R E P O R T R E S U M E S

ED 015 505 CG 001 124

ADEQUACY OF TEST VALIDITIES FOR INDIVIDUAL PREDICTION.

BY- WEITZ, HENRY

PUB DATE 2 SEP 67

EDRS PRICE MF-$0.25 HC-$0.56 12P.

DESCRIPTORS- *COUNSELING, *COUNSELING GOALS, *TEST VALIDITY, *PREDICTIVE VALIDITY, *TEST RELIABILITY, VOCATIONAL COUNSELING, TEST INTERPRETATION, TEST RESULTS, TESTING PROBLEMS, APTITUDE TESTS, PREDICTIVE MEASUREMENT, STATISTICAL ANALYSIS, PROBLEM SOLVING, PSYCHOLOGICAL TESTING, PERSONNEL SELECTION, DATA ANALYSIS,

COUNSELORS OFTEN ADMINISTER TESTS OF QUESTIONABLE VALIDITY. IN RELIABILITY STUDIES, EVERY PRECAUTION IS TAKEN TO STABILIZE THE STIMULUS SITUATION. IN ASSESSING VALIDITY, CONCERN CENTERS ON BEHAVIOR UNDER DIFFERENT STIMULUS CONDITIONS. CRONBACH'S THEORETICAL LIMIT FOR A VALIDITY COEFFICIENT OF A TEST IS THE SQUARE ROOT OF THE RELIABILITY COEFFICIENT. GHISELLI, IN REVIEWING HUNDREDS OF VALIDITY STUDIES COMPLETED BETWEEN 1919 AND 1964, FOUND THAT NONE OF HIS FOUR MAJOR CLASSES OF APTITUDE TESTS FORECAST PROFICIENCY ON ANY JOB WITH A HIGH DEGREE OF ACCURACY. ALTHOUGH TESTS CAN HAVE A SUFFICIENTLY HIGH DEGREE OF PREDICTIVE POWER TO BE OF PRACTICAL VALUE IN PERSONNEL SELECTION, THEY MAY ALSO FRUSTRATE THE COUNSELOR'S CHANCES OF HELPING THE CLIENT SOLVE IMMEDIATE PROBLEMS. IN ACTUARIAL INTERPRETATION AND BEHAVIOR PREDICTION BASED ON TEST DATA, GREATER RELIANCE SHOULD BE PLACED ON MULTIVARIATE ANALYSIS. RIGOROUS COLLECTION, ANALYSIS, AND REPORTING OF PREDICTION AND CRITERION DATA ARE NECESSARY. USE OF GOLDMAN'S MULTIDIMENSIONAL APPROACH CAN HELP THE CLIENT BY DEFINING THE ESSENTIAL ELEMENTS OF THE PROBLEM. UNTIL MORE VALID TESTS ARE DEVELOPED, WE MUST CHOOSE BETWEEN REFUSAL TO USE TESTS, AND USING THEM AS PART OF AN EXTENSIVE DESCRIPTION. THIS SPEECH WAS PRESENTED AT THE AMERICAN PSYCHOLOGICAL ASSOCIATION CONVENTION, WASHINGTON, D.C., SEPTEMBER 2, 1967. (FR)

SYMPOSIUM ON

CONVEYING TEST INFORMATION IN THE COUNSELING PROCESS:

Adequacy of Test Validities for Individual Prediction

Henry Weitz

APA Division 17
2 September 1967

Counselors collaborate with clients to assist them in resolving their immediate problems. These problems may involve vocational choice, educational decisions, social pressures, emotional stress, or, more likely, some mixture of stimuli for which no appropriate response pattern is spontaneously available to the client. The effective counselor, however, goes beyond this immediate objective and seeks to achieve the ultimate goal of counseling, aiding the client to acquire generalized problem-solving behavior.

It makes little difference how we characterize the ultimate purpose of counseling. We may label this adjustment, enhancement of the phenomenal self, development of ego strength, acceptance of the existential condition, serenity, or what have you. What the conscientious and responsible counselor seeks to do is to aid his client in acquiring a highly organized and highly energized system of psychological mechanisms that will permit the client to check conflict as or before it arises and to move forward toward meaningful, satisfying, distant goals. We have chosen to call these psychological mechanisms generalized problem-solving behavior and the state in

which they function spontaneously and efficiently, serenity (Weitz, 1964, p. 143-144).

One of the principal considerations in the solution of human problems is a recognition of and a willingness to accept reality as it is. Unfortunately, in the case of human interactions, reality is the product of behavior performed on the basis of perceptions and misperceptions of objects and events. Thus, reality is seldom an unambiguous event. Counseling, however, is designed to aid in the achievement of generalized problem-solving behavior by reducing the client's misperceptions and clarifying some of the ambiguities in his segment of reality.

It is in this context that I should like to consider with you the use of tests in counseling. With few exceptions, counseling psychologists use tests in helping their clients identify their problems and find solutions to them, although as we all know, counselors of certain persuasions do so reluctantly and some even truculently with the consequent ill-effects equalled only by the "test and tell" school of counselor. If tests, then, are widely used in counseling---and the economic euphoria enjoyed by many test publishers is evidence of this wide use---we are confronted with the question of whether or not test scores can provide us a basis for the correction of client's misperceptions and a means of clarifying some of the ambiguities of perceived reality. A lifetime of using tests in counseling has led some of us---in our few moments of clearer vision---to suspect that testing in guidance may be a piece of superstitious tribal ritual that permits us to accept our error-laden

predictions with less pain and anxiety.

Answers to the question of whether or not testing is likely to contribute to our client's acquisition of generalized problem-solving behavior may be found in an examination of the validities of some of the kinds of tests we use.

The validity of a test, as we all know, is the degree to which it measures what it purports to measure. There are, again as we all know, many different kinds of test validity all more or less suited to the different purposes of different test users. In counseling, however, we are most concerned with predictive validity. When a counselor orders the administration of a test in the course of counseling, he wants to be able to say, "If the client exhibits a given level of a definable kind of behavior---as measured by the test---today, he is likely to exhibit a similar level of the same kind of behavior in the future under different, but similar, circumstances."

Let us look at it this way: A test is, in essence, a sample of behavior. If we wish to make some estimate of a client's mathematical behavior, say, we confront him with a sample of mathematical stimuli and observe the ways in which he makes his responses. If our sample of mathematical stimuli is well chosen, it will represent the total set of stimuli that have confronted or are likely to confront the client and hence will evoke a variety of mathematical responses available to the client. When his responses are compared with the responses of other examinees who have had a

similar opportunity to acquire mathematical responses we can assign a number to his behavior that not only tells us something about his present total mathematical behavior but also tells us something about his probable future behavior in this domain. This prediction is possible because of our experience with large numbers of observations of mathematical behavior.

When we are assessing the reliability of a test, we try, insofar as possible, to maintain fairly constant stimulus situations. Consider, for example the test-retest method of estimating reliability. Here we try to reproduce the same test conditions and in many instances, as when the retesting is done with the same form of the test, the specific mathematical stimuli remain the same. In the case of alternate form retesting, the specific stimuli are changed, but they are assumed to activate the same behavioral responses. Thus in reliability studies, we take every precaution to stablize the stimulus element of the behavior product in order to estimate the effects of chance factors irrelevant to the restricted range of behavior being sampled.

When we are assessing validity, however, we are concerned with degree to which the behavior under observation functions under different sets of stimulus conditions. Thus we may sample mathematical behavior with a test of arithmetic speed and accuracy in which the stimulus items are problems of addition, subtraction, multiplication, and division of three digit numbers. In order to measure the validity of this test we may sample the behavior of a group of bookkeepers and rate them with respect to speed and

accuracy on their jobs over a period of time and then determine the degree of correlation between their test scores and their rating. The significant factor here is that a major element in the behavior product, the stimulus element, is markedly different in the test and criterion samples although both have the generally common factor of arithmetical stimuli.

When we take two samples of behavior in which every effort is made to insure similarity of both stimuli and responses as in the case of reliability studies, we never get a perfect correspondence between the two samples. Reliability coefficients in excess of .90 are considered exceptional within restricted range samples of subjects. Consequently, one would expect that samples of behavior in which a major element, the stimulus, is varied, would produce even lower coefficients of validity. And, in fact, as we all know, they do. But how low? Cronbach (1960, p. 132) tells us that a validity coefficient can never exceed the square root of the reliability coefficient of a test. Thus a test with a reliability of .90 can be expected to have a validity coefficient of .95 or less with some external criterion. This is the <u>theoretical limit</u>, but how much less do we find in actual practice in the case of the tests used in counseling?

I should like to consider with you what can happen in the case of using tests in the area of vocational counseling. To be sure, this is not the only kind of problem confronting the counselor, but we seem to know more about tests in this area than we do in some others so that what we have to say about the findings here can

be applied in other areas provided due allowance is made for our more limited knowledge there.

Ghiselli reviewed hundreds of validity studies, both published and unpublished, that were completed between 1919 and 1964, and summarized his findings in a fascinating if frightening little book entitled The Validity of Occupational Aptitude Tests (Ghiselli, 1966). After finding that interest and personality tests contributed little to the prediction of occupational performance, he grouped the remaining tests into four major categories: (1) intellectual abilities, (2) spatial and mechanical abilities, (3) perceptual accuracy, and (4) motor abilities. He summarized the validity coefficients for many different tests in each of these "aptitude" areas against two principal criterion measures for a wide variety of occupational groups. The criterion measures were training and job proficiency.

Ghiselli reports that:

> . . . none of the major classes of tests forecasts proficiency on any job with a high degree of accuracy. Although in a number of instances tests have moderate validity, their power to predict success on the actual job is substantially less than their power to predict trainability. (Ghiselli, 1966, p. 64.)

He points out that, "Taking all tests as a whole for training criteria, it will be observed . . . that nearly half the average validity coefficients, 47.3 per cent, are at least moderately high, being .30 or greater. Indeed," he continues, "nearly three quarters of them, 72.5 per cent, are above .20, which is perhaps, the lower limit of usefulness." (Ghiselli, 1966, p. 123.)

When we move to proficiency criteria, however, the picture darkens. Here Ghiselli (1966, p. 124) reports that for all tests only 14.7 per cent had validity coefficients of .30 or better while somewhat less than half, 44.5 per cent, had coefficients in excess of .20.

Ghiselli draws the following conclusions about average validity coefficients:

> . . . while the general predictive power of aptitude tests in forecasting occupational success is by no means zero, it is far from impressive. For all tests and jobs as a whole, a coefficient of the order of .30 describes the general validity of tests for training criteria, and one of the order of .20 gives the value for proficiency criteria. (Ghiselli, 1966, p. 125.)

It might be worth remembering, at this point, that if the validity coefficient is based on 100 cases, it needs to exceed .195 to be significantly different from zero at the .05 level of significance and to exceed .230 at the .01 level. Or to put it another way: A client whose test score places him in the upper quarter of his group has about one chance in four of being in the upper quarter on the criterion measure and an equal chance of being in the lowest quarter when the validity coefficient is about .0. Even when the validity coefficient is as high as .40---and this is somewhat higher than the average for Ghiselli's findings--- the client's chances of being in the upper quarter on the criterion measure when he scores in the upper quarter on the predictive test is considerably less than 50-50; it is 428 chances out of 1000. One is inclined to paraphrase the old saying as "with validity

coefficients like this who needs enemies?"

But, of course, these are average validity coefficients reported by Ghiselli, and we all know that counselors use only the best tests. What does Ghiselli have to say on this score? Just this:

> The highest validities found in any of the single studies reviewed for this summary, studies in which more than one hundred workers were used, were .77 for training criteria and .66 for proficiency criteria. In both of these investigations the coefficients were for intelligence tests applied in the trades and crafts. (Ghiselli, 1966, p. 126.)

With a correlation coefficient of .77 the coefficient of determination is .59 and with a correlation of .66 the coefficient of determination is .44. (See Croxton and Crowden, 1943, pp. 663-664.) This suggests that in the former case slightly over half the variability in the criterion is accounted for by variability in the predictor and in the case of the validity coefficient of .66 only 44 per cent of the variability in job proficiency is accounted for in the predictor measure.

Ghiselli's parting comment needs to be considered by counselors who plan to use tests:

> It is apparent that even the most optimistic supporter of tests cannot claim that they predict occupational success with what might be termed a high degree of accuracy. Nevertheless, in most situations tests can have a sufficiently high degree of predictive power to be of considerable practical value in the selection of personnel. (Ghiselli, 1966, p. 127.)

If, as we indicated earlier, counseling is intended to help the client correct his misperceptions and reduce ambiguity in reality, and if, as Ghiselli suggests, tests are likely to provide us with only a limited and often distorted view of the events we wish to observe, then we may be inclined to conclude that testing in counseling may not only frustrate our chances of assisting the client in solving his immediate problem, but may in fact, engender in the client a self-defeating approach to generalized problem-solving, for the inclusion of highly ambiguous data in the problem-solving process may, and I fear, often does, acclimatize the client to a toleration of superstition and ritual in the solution of problems and teach him irrationality in problem-solving which, to succeed, needs to become a highly logical and rational process.

All is not lost, however, despite the obvious limitations placed on tests by their apparently low validities. Goldman has suggested a model for the dimensions of interpretation in counseling (Goldman, 1961, p. 143ff.) Three dimensions are suggested (1) type of data (including test and non-test data), (2) type of treatment of data (including actuarial and clinical), and (3) type of interpretation (including descriptive, genetic, predictive, and evaluative). It is possible in this three-dimensional format to generate sixteen kinds of interpretation of events which play a part in the client's problem and its solution. About half of these relate to test data. About half of the test interpretations---according to the model---are actuarial (see Meehl, 1954) in which test validity plays an important role. Thus according to the

Goldman model only about one-quarter of the kinds of interpretations generated in the course of problem-solving in counseling are dependent upon test validity. (The validity of other types of interpretation need also to be questioned, but this is not our present concern.)

This state of affairs may help to lead us out of our morass of ambiguity. In the case of actuarial interpretation and prediction of behavior based on test data we need to place greater reliance on multivariate analysis than we have done heretofore. (See, for example: Rulon, et al., 1967.) Devices and programs for their use are now available. What is required now is the rigorous collection of prediction and criterion data, all sorts of data, on a wide scale, and the analysis and reporting of it in a form that is not only useful to the counselor, but also communicable to the client. Such an approach appears to give promise of moving test data away from their present status as a piece of superstitious tribal ritual and nearer to the role of the unambiguous picture of reality so desperately needed in the solution of human problems.

Another approach is the application of what I have called description by extension (Weitz, 1964, pp. 84-85). Here the other three-quarters of the Goldman model come into play. If an event is described in sufficient detail, the inconsistent elements will begin to emerge. Thus when we find a student whose level of general test measured ability is high, whose test measured abilities and interests would seem to make him admirably suited to a career in medicine, but who appears to be failing the first chemistry course in a premedical curriculum, we do not throw out the test scores on

the basis of their low validity in this instance nor do we urge the client to "keep at it, do your best, because the tests show that you can become a successful and happy surgeon, if you'll only try." The inconsistency of the data here suggests that we may be trying to solve the wrong problem or, at least that we are attacking problems in the wrong order. Further extensive description of the client's behavior including additional, but different, test data, may help us redefine our problem in a way that permits a rational and realistic solution.

With such an approach we are more likely to help our client resolve his immediate problem and at the same time assist him to acquire one of the first essentials in generalized problem-solving behavior, defining the problem in <u>all</u> of its essential elements and avoiding quick solutions to poorly defined problems that may, in fact, not exist. Until such time as tests of ability and inventories of personality traits and interests are developed that have validity coefficients sufficiently high (.90 or better) for the prediction of individual performance against clearly specified and relevant criteria, we are faced with the alternative of refusing to use tests in counseling at all or using them as part of an extensive description including multivariate analysis. Of course, we can always go on doing what we have been doing for many years---using tests as a ritual for placating the gods of chance.

REFERENCES

Cronbach, L. J. Essentials of psychological testing. New York: Harper and Brothers, 1960.

Croxton, F. E. and Crowder, D. J. Applied general statistics. New York: Prentice Hall, 1943.

Ghiselli, E. E. The validity of occupational aptitude tests. New York: John Wiley and Sons, Inc., 1966.

Goldman, Leo. Using tests in counseling. New York: Appleton-Century-Crofts, Inc., 1961.

Meehl, P. E. Clinical vs statistical prediction. Minneapolis: University of Minnesota Press, 1954.

Nalen, P. J. Multivariate statistics for personnel classification. New York: John Wiley and Sons, Inc., 1967.

Waitz, Henry. Behavior change through guidance. New York: John Wiley and Sons, Inc., 1964.