DOCUMENT RESUME

ED 101 011 TM 004 244

AUTHOR Strobel, Michael G.; Dudek, Stephanie Z.
TITLE On the Testing of Tests: A Longitudinal Study.
PUB DATE [74]
NOTE 9p.; Paper presented at the Annual Convention of the American Psychological Association (New Orleans, Louisiana, 1974)

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS Achievement Tests; Elementary Education; Elementary School Students; Intelligence Tests; *Longitudinal Studies; Middle Class; Personality Tests; Predictive Ability (Testing); *Testing; *Test Reliability; *Test Reviews; *Test Validity

ABSTRACT

Fifty-eight middle class children were tested over 6 years with 25 achievement, I.Q., and personality tests. Consistency of test results were evaluated by a variance comparison method and a simple signal detection model. Both methods lead to the conclusion that achievement tests are far better predictors than personality tests with I.Q. scales placing in between. (Author)

ON THE TESTING OF TESTS : A LONGITUDINAL STUDY

M.G. Strobel & S.Z. Dudek

Université de Montréal

The purpose of the present study was twofold : (1) to evaluate the consistency of a number of standard tests used to measure cognitive, perceptual-motor, achievement and personality variables in children and (2) to assess the degree to which inconsistency of a test can be attributed to uneven developmental growth in children. We aimed at conceptually simple constructs to derive the predictive value of tests during the development of normal school children, with particular emphasis on long-term achievement.

METHOD

SUBJECTS: Ss were the 103 children of the Kindergarten grade from two Montreal suburban schools; 52 were boys and 51 were girls. The schools were selected to represent a typical middle class background: The median interval of annual family income in the sample was $7,500 to $10,000 (1964). 40% of the fathers had attented at least one year of university, 32% had finished High School and 28% had not completed their High School education. Socio-economic distribution was as follows: 19% were professionals, 36% held sales and clerical jobs, 17% were self-employed, 11% occupied administrative posts. The remaining 17% were placed in the miscellaneous category.

Testing Procedure

The children were followed over 6 years, from Kindergarten through Grade 5 (1964-1970). Depending on the test, the children were either individually tested or seen as a group at yearly intervals by two psychotechnicians. Achievement Tests for Reading, Arithmetic and Language, were administered in the 10th month of each Grade. Scoring of tests was supervised by a psychologist.

The following is the list of tests used in the study:

1. Lincoln-Oseretzky Motor Development Scale.
2. Goodenough-Harris Draw-a-man-test.
3. Goodenough-Harris Draw-a-woman-test.
4. WISC Performance Scale.
5. WISC Verbal Scale.
6. WISC Full Scale.
7. Lorge-Thorndike Group Intelligence Scale.
8. Piaget test of causal and operational thinking (Total score)
9. - 21. Cattell's Children's Personality Questionnaire, Scales A to Q4.
22. California Achievement Test : Reading.
23. California Achievement Test : Arithmetic.
24. California Achievement Test : Language.
25. California Achievement Test : Total score.

Attrition rate for the group between Kindergarten and Grade 5 was 35%, with 67 children remaining by Grade 5 in the sample. For an analysis of test stability from year to year only those children for whom data were complete for all years were included. With this constraint the final sample consisted of only 58 children but the intra-group variability was not distorted by extraneous subjects.

INTRA-TEST STABILITY ANALYSIS

Previous longitudinal research has shown that test measures, including I.Q. tests, tend to increase over the years. The present study confirmed this trend: groups showed an average gain of 10 to 20 percent depending on the test. Thus, a child who maintained the same score over the years was in fact losing points if the group mean had increased. Rate of developmental change was evaluated in relation to the childs group, and therefore standard scores were used as the ba-

sis for a quatitative measure of rate of change.

Change in a standard score reflects a change in the child's relative standing in the group. Differences in standard scores for each child, from year to year, provide a measure of his mobility within that group. Summing these differences in Z scores over the years results in a value which is numerically identical to the difference between the first and last measures. That is, the difference between the first and last measures represents the total amount of change for a particular child (relative to his group). In order to asses the amount of movement within the group as a whole, one is tempted to take the mean of those differences. However, due to the fact that the measurements are in Z scores, the mean of these differences will be zero. The sum of squared differences, divided by N, will give the desired quantity. It is easily shown that this is a between subject variance and represents average intra-group mobility. This variance was used to discriminate between tests. High variance within a test over the years signifies a great deal of instability and the test will be a poor predictor.

There remains the problem of how to differentiate test instability due to poor test construction from instability due to the idiosyncratic variability of individual children. Since the samples for different tests were not always made up of the same Ss it was possible that some tests fared badly because they were plagued with highly unstable children. We expressed the inconsistency of the individual child as the difference of the standard scores from year to year and calculated the variance of these difference scores for each of the 27 tests. These variances are directly comparable and measure the extent to which Subject variability contributes to the uncertainty in longterm predictions from these tests. Thus comparatively high variability in some tests cannot be attributed to the Ss if the same children show erratic scores in just those tests while remaining quite consistent in others. By the same token, tests which include sizeable proportions

of children with erratic scores throughout may be excused for not predicting better than they do.

RESULTS

Figure 1 describes the relative stability of tests as the between Ss variance calculated from the difference of standard scores over the span of 5 or 6 years. The lower the variance the greater the stability of the test. It appears that the achievement tests show the strongest predictive power. The total achievement score on the California Achievement Test (#25) has a variance of .23 which means the average displacement to be expected of Ss taking this test is less than 1/2 SD (.23 =.48). The three achievement scales of the CAT, reading, arithmetic and language (#22-24), as well as the intelligence tests (#4-8) and the motor tests (#1, 2) with the exception of the Goodenough-Harris (woman) test (#2), have a comparable stability index of about 1., meaning that the average displacement of Ss within the group did not exceed one SD. The predictive power of personality tests (#9-21), where subject's standing in the group over successive years changes a great deal, is weak.

This conclusion is supported by an analysis of the degree to which individual children show test variability over classes of tests. Taking again the difference in Z values between the first and the last year as a score and calculating the variability of these scores for each S over the four groups of tests, only one child in 29 (3%) exceeded a variance of 1 in the achievement tests, 15 out of 58 children (26%) had variances greater than 1 in the motor and intelligence tests, but 34 out of 43 children (79%) exceeded this value in the personality tests.

Therefore the instability encountered in these latter tests seems to be due largely to the poor caracteristics of the tests rather than the ideosyncratic

variability in the children. The fact that some scales are better than others e.g. #20, versus #14, (two factors on the Cattell Scale), cannot obscure the finding that the personality tests as a whole cannot be interpreted with the same degree of confidence as the other tests.

TEST CONSISTENCY OF BROAD CLASSIFICATIONS

So far the evaluation of test stability was based on comparisons of variances derived from differences of Z scores. For the practitioner it might be of greater value to know whether the groupings and distinctions he makes on the basis of his test results are reasonably consistent over the long run.

We therefore applied a second method, non-parametric in form, derived from a signal-detection model, which has the appeal of using empirical concepts likely to be encountered in practice.

Suppose test results at school entrance examination were used to form classes of children with special programs. Let the arbitrary class boundary be one SD. Thus children scoring 1 SD or more above the mean would go to an accelerated program, those scoring 1 SD or worse below the mean would receive auxiliary training and the bulk of 68% would be devided by the mean into an above average and a below average group. The question asked by the practitioner is how many of the children thus classified would still turn up in the original group 5 or 6 years later. In terms of the signal detection model: How many "hits" did the test score? If a child turned up two or more categories removed from its original classification it surely was a "bad misplacement". Using a rather strict criterion for hits but a larger one for misplacements takes into account the graded severity of consequences for misses. Presumeably less harm is done if misclassification is by only one category. The proportion of correct predictions and bad misplacements were calculated for each test. Again only Ss participating in all test

scores over 5 or 6 years, depending on the test, were included in the sample.

Figure 2 shows the outcome of the analysis with the signal detection model. The tests are plotted in terms of correct predictions and bad misplacements (i.e. more than two categories removed from the initial placement). The graph can be read by dividing the plot into 4 quadrants. Let one third be the minimum acceptable correct prediction and one in 5 be the maximum tolerable rate of bad misplacements (.33 and .20 on the Y and X axes respectively); then quadrant II contains the tests with the absolute best performance both in terms of high number of hits and small number of bad mistakes. Quadrant IV points out the tests with all around wrong predictions, while quadrants I and III contain those ambigous test performances with either not enough usable correct predictions or too many misplacements. Achievement tests, motor - and intelligence tests accumulate in quadrant II while the personality tests fall into quadrant IV, or into the ambiguous and unsatisfactory categories. The close correspondence of the two modes of analysis can be taken as an indication that the rather complex numerical analysis involving transformations and laborious searches and matchings can be bypassed by the rather simple counting procedure necessary to build the signal detection model.

In conclusion, it might be said that while intelligence, achievement and perceptual motor tests appear to measure relatively stable dimensions of functioning, personality tests, at least of the inventory type, present serious problems for predictive purposes. Reasons for this are undoubtedly complex. It is possible to say that personality, particulary in normal children, is not yet crystallized and is constantly emerging. Therefore stable measures should not be expected. This conclusion contradicts psychoanalytic theory which may be referring to more basic personality structures. Tests of the CPQ type probably do not reach this level. Whatever these tests are measuring would seem to be closer to the more fluctuating traits which vary from situation to situation and from year to year.

ABSTRACT

58 middle class children were tested 6 years with 25 achievment, I.Q., and personality tests, Consistency of test results were evaluated by a variance comparison method and a simple signal detection model. Both methods lead to the conclusion that achievement tests are far better predictiors than personality tests with I.Q. scales placing in between.