# Validity, Reliability, and Baloney

Edward E. Cureton

*Educational and Psychological Measurement* 1950; 10; 94
DOI: 10.1177/001316445001000107

# VALIDITY, RELIABILITY, AND BALONEY[1]

EDWARD E. CURETON
University of Tennessee

IT is a generally accepted principle that if a test has demonstrated validity for some given purpose, considerations of reliability are secondary. The statistical literature also informs us that a validity coefficient cannot exceed the square root of the reliability coefficient of either the predictor or the criterion. This paper describes the construction and validation of a new test which seems to call in question these accepted principles. Since the technique of validation is the crucial point, I shall discuss the validation procedures before describing the test in detail.

Briefly, the test uses a new type of projective technique which appears to reveal controllable variations in psychokinetic force as applied in certain particular situations. In the present study the criterion is college scholarship, as given by the usual grade-point average. The subjects were 29 senior and graduate students in a course in Psychological Measurements. These students took Forms Q and R of the *Cooperative Vocabulary Test*, Form R being administered about two weeks after Form Q. The correlation between grade-point average and the combined score on both forms of this test was .23. The reliability of the test, estimated by the Spearman-Brown formula from the correlation between the two forms, was .90.

The experimental form of the new test, which I have termed the "B—Projective Psychokinesis Test," or Test B, was also applied to the group. This experimental form contained 85 items, and there was a reaction to every item for every student. The items called for unequivocal "plus" or "minus" reactions, but in advance of data there is no way to tell which reaction to a given item may be valid for any particular purpose. In this

[1] This paper was presented in Denver, Colorado, September 7, 1949, at a meeting sponsored jointly by the Division on Evaluation and Measurement of the American Psychological Association and the Psychometric Society.

respect Test B is much like many well-known interest and personality inventories. Since there were no intermediate reactions, all scoring was based on the "plus" reactions alone.

I first obtained the mean grade-point average of all the students whose reaction to each item was "plus." Instead of using the usual technique of biserial correlation, however, I used an item-validity index based on the significance of the difference between the mean grade-point average of the whole group, and the mean grade-point average of those who gave the "plus" reaction to any particular item. This is a straightforward case of sampling from a finite universe. The mean and standard deviation of the grade-point averages of the entire group of 29 are the known parameters. The null hypothesis to be tested is the hypothesis that the subgroup giving the "plus" reaction to any item is a random sample from this population. The mean number giving the "plus" reaction to any item was 14.6. I therefore computed the standard error of the mean for independent samples of 14.6 drawn from a universe of 29, with replacement. If the mean grade-point average of those giving the "plus" reaction to any particular item was more than one standard error *above* the mean of the whole 69, the item was retained with a scoring weight of *plus one*. If it was more than one standard error *below* this general mean, the item was retained with a scoring weight of *minus one*.

By this procedure, 9 positively weighted items and 15 negatively weighted items were obtained. A scoring key for all 24 selected items was prepared, and the "plus" reactions for the 29 students were scored with this key. The correlations between the 29 scores on the revised Test B and the grade-point averages was found to be .82. In comparison with the Vocabulary Test, which correlated only .23 with the same criterion, Test B appears to possess considerable promise as a predictor of college scholarship. However, the authors of many interest and personality tests, who have used essentially similar validation techniques, have warned us to interpret high validity coefficients with caution when they are derived from the same data used in making the item analysis.

The correlation between Test B and the Vocabulary Test was .31, which is .08 higher than the correlation between the

Vocabulary Test and the grade-point averages. On the other hand, the reliability of Test B, by the Kuder-Richardson Formula 20, was −.06. Hence it would appear that the accepted principles previously mentioned are called in question rather severely by the findings of this study. The difficulty may be explained, however, by a consideration of the structure of the B−Projective Psychokinesis Test.

The items of Test B consisted of 85 metal-rimmed labelling tags. Each tag bore an item number, from 1 to 85, on one side only. To derive a score for any given student, I first put the 85 tags in a cocktail shaker and shook them up thoroughly. Then I looked at the student's grade-point average. If it was B or above, I projected into the cocktail shaker a wish that the student should receive a high "plus" reaction score. If his grade-point average was below B, I projected a wish that he should receive a low score. Then I threw the tags on the table. To obtain the student's score, I counted as "plus" reactions all the tags which lit with the numbered side up. The derivation of the term "B−Projective Psychokinesis Test" should now be obvious.

The moral of this story, I think, is clear. When a validity coefficient is computed from the same data used in making an item analysis, this coefficient cannot be interpreted uncritically. And, contrary to many statements in the literature, it cannot be interpreted "with caution" either. There is one clear interpretation for all such validity coefficients. This interpretation is—

"Baloney!"