

# SUGGESTIONS LOOKING TOWARD A FUNDAMENTAL REVISION OF CURRENT STATISTICAL PROCEDURE, AS APPLIED TO TESTS

BY SIDNEY L. PRESSEY

*University of Indiana*

The past three or four years have been notable in psychological history for the remarkable development of statistical methods as applied to the problems of mental measurement. This advance is undoubtedly of the very greatest importance. The writer has come to feel, however, that, with the first enthusiasm in such work, there has been a tendency toward over-elaborateness and diffuseness of treatment, and a lack of directness and incisiveness in the statistical procedure. And he wishes to point out certain limitations to the present concepts of "reliability" and "validity" as applied to tests, and certain objections to the customary use of the theory of the normal curve in test building, which he feels to be of distinct importance.

The situation can most readily be made clear by a very concrete example. Suppose, then, that a high-school principal desires to give a group test for measuring general intelligence to his entering class, in order to pick out in advance those who are likely to fail in their freshman work. He has a number of scales under consideration. And he wishes evidence as to the relative merits of these scales for this purpose,—for the selection of potential failures. He will very likely be given data with regard to the comparative 'reliability' and 'validity' of these scales; information may also be produced with regard to the organization of the tests, especially in respect to the normality of the distribution of scores yielded. The present paper aims to show that no one of these three sets of facts gives that close contact, which is desirable, with the practical problem.

## I. INADEQUACY OF THE PRESENT CONCEPT OF 'RELIABILITY'

The principal may be urged to use a particular scale because the scale has a high 'reliability.' The exact meaning of 'reliability' must, however, first be carefully looked into. The meaning of the concept can best be understood by considering the way in which 'reliability' is usually measured. The most common method is simply to give two duplicate forms of the same test, one after the other, to the same subjects. The ratings obtained by the subjects on 'Form A' and 'Form B' are then correlated. And the closeness of the correlation indicates the reliability of the test.

The significance, and the limitations, of a measure thus obtained are fairly obvious. Two such limitations are especially important. (a) The measure is evidently a measure of the reliability of the sampling,—of the particular type of performance involved in the test. When one speaks of the reliability of an instrument, one naturally thinks of its reliability *for some purpose*. Such a connotation must be guarded against here. One must not come insensibly to think of the reliability coefficient of a test of intelligence, for instance, as indicating the value of the test, as a measure of intelligence. Such a conclusion is sound only if a test is a simple sampling of the ability which it is sought to measure; and this happens much more rarely than might be supposed.<sup>1</sup> The term 'consistency' would, therefore, seem a more accurate term; the 'reliability' coefficient indicates only the extent to which a test is consistent with itself. And it is entirely possible that a test should yield highly consistent results which were, nevertheless, not at all measures of the function which it was desired to measure.<sup>2</sup>

<sup>1</sup> It might seem, for instance, that the Courtis Scale B was a simple sampling of ability in the fundamentals. But recent research has shown the situation to be by no means so simple. (See Thorndike, E. L., and Courtis, S. A., 'Correction Formulas for Addition Tests,' *Teachers' College Record*, 1920, 21, 1-24.)

<sup>2</sup> Thus, not so many years ago, cancellation tests were frequently included in 'batteries' of tests intended for the measurement of mental endowment. (See, for instance, Pyle, 'The Examination of School Children,' Macmillan, 1913.) It now seems quite clear that cancellation tests are not good tests of intelligence. (See McCall, 'Some Correlations between Mental Traits,' *Teachers' College*, 1916.) But cancellation tests appear to be quite 'reliable' measures,—they are simply not good tests of general intelligence. They are, therefore, not 'reliable' for the purpose for which Pyle used them.

It must also be kept in mind (*b*) that such a measure of the reliability of the sampling may be considered an adequate measure for this purpose only if the scores obtained on 'Form A' and 'Form B' may be considered entirely random samplings of performance on such a test. Usually they cannot be so considered. There may be an initial difficulty with directions at the beginning of 'Form A' and a slight fatigue toward the last of 'Form B.' What is, with many of the tests, more important—the method as described above tells us nothing whatever about the 'consistency' of the results from one examiner to another, one scorer to another, from one day to another, or one time of the day to another.

To come back to the original problem, then: such a measure of the consistency of the test with itself, under certain circumstances, tells the high school principal surprisingly little as to the value which that test may have in distinguishing his potential failures from the rest of their class.<sup>1</sup> And information with regard to the 'validity' of the scale is naturally turned to, to settle this practical question.

## II. THE ARTIFICIAL NATURE OF CURRENT CONCEPTS REGARDING VALIDITY

The principal is, then, urged to use a particular scale because the scale has a high 'validity' as a measure of general ability. That is, data are presented showing that the scale gives results having a high correlation with independent criteria as to general intelligence, and congruence with current theories regarding the nature of general intelligence,—there

<sup>1</sup> The writer is inclined to feel that most problems of consistency can best be dealt with in general terms. That is, what difference, in general, may one expect in test results if one tests Monday instead of Friday, at 9 o'clock instead of 3 o'clock? What difference, in general, may be expected, with a given type of directions, from one examiner to another? What differences, with various scoring methods, may be expected from one scorer to another? What differences may result in the score of an individual as the result of fluctuations from one time to another, in general feeling tone, energy, vigor, health? The writer believes that, until evidence to the contrary appears, it may be taken for granted that such factors affect all tests in more or less the same way; certain general theorems with regard to their operation should, then, be possible—or general precautions taken. The only problem of consistency that needs specific determination for each test would then be consistency as it relates to the subject matter of the test.

is a regular rise in score from year to year until maturity, a relative freedom from the influence of specific training, and so on. This concept of 'validity' is also, the writer feels, beside the point, if not misleading, so far as the practical problem of the high-school principal is concerned. And again there are two difficulties.

In the first place (*a*), since the extent to which general intelligence is the fundamental factor, in conditioning success and failure in the Freshman year of high school, is not known, the usefulness of the scale (even if proven a satisfactory measure of general intelligence) is still an unknown quantity. Stability of character, willingness to apply oneself even though the restraints of grammar school supervision are now removed, interest in the more mature subjects of the high-school curriculum—such elements are probably more important than is often supposed, in the total situation.<sup>1</sup> Differences in the adequacy of previous preparation may also be of importance. So proof of the 'validity' of a scale as a measure of 'general intelligence' is by no means proof of the value of the scale in sorting out potential failures among these high-school freshmen. In fact, it might almost be said that in proportion as the scale measured one element only, in a complex situation, to just that extent was the scale inadequate for dealing with that total situation!<sup>2</sup>

It remains to be pointed out, however, that even though

<sup>1</sup> For a discussion of this tendency to overestimate the comparative importance of intelligence see Rosenow, Curt, 'Is Lack of Intelligence the Chief Cause of Delinquency?' *PSYCHOL. REV.*, March, 1920.

<sup>2</sup> The more extreme theories in regard to general intelligence surely make up, in the aggregate, an extraordinary concept. It should surely be kept in mind that it is, in the first place, an analytical concept and so dependent for its character upon the methods of analysis employed. It should also be pointed out that such a concept naturally receives successive accretions in the way of theory and may, by a mental synthesis largely adventitious to the facts, acquire a reality which is very largely an artifact. Scores on various tests are lumped and the aggregate used as a measure of general ability. A more or less close relationship is naturally found between such an aggregate and the average marks of the children in school. Teachers, especially in the grades, naturally think of the child's work as a whole, and give marks showing high correlations between abilities in different subjects. And the children come to this attitude and react to their school work as a whole. And—the whole situation is cumulative. One might, in fact, imagine the concept of general ability thus developing even though abilities were, as a matter of fact, diverse and uncorrelated.

the demonstration of a close correlation between the scale in question and a general intelligence *were* supplemented by evidence that general intelligence was the fundamental factor in the situation, still the suitability of the scale for the particular problem would remain to be shown. That is, (b) the usual method for stating relationships between two variables—the correlation coefficient—does not express satisfactorily the nature of that relationship, for diagnostic purposes, at a particular point in the distribution. The problem is: How unmistakably will the scale set off the lower 15 per cent. or so in scholastic ability? A correlation coefficient is only very general evidence in regard to this particular matter.<sup>1</sup> And it is evidence with regard to such diagnostic efficiency that the school principal should require.

### III. THE IRRELEVANCY OF THE THEORY OF THE NORMAL CURVE IN PRACTICAL PROBLEMS IN CLASSIFICATION

Proof of the validity of a scale as a measure of general intelligence is, then, not proof of the value of that scale for sorting out potential high-school failures, since failure is not conditioned by general ability alone, and since the diagnostic efficiency of a measuring instrument is not the same thing as the general relationship of that instrument to the factor concerned. A third set of facts may, nevertheless be introduced in evidence of the value of the scale in question. It may be pointed out that the tests of the scale are very carefully constructed so that equal increments on the scale represent equal increments in ability, and so that the total distribution of abilities yielded is closely similar to the distribution of abilities that would be expected according to

<sup>1</sup> See, for instance, Thurstone, L. L., 'Mental Tests for College Entrance,' *J. of Educ. Psychol.*, March, 1919 and Pressey, S. L., 'Suggestions with Regard to Prof. Thurstone's "Method of Critical Scores,"' *J. of Educ. Psychol.*, December, 1919.

The writer has often wondered whether the early introduction of the Pearson products-moments formula for calculating the correlation coefficient has not hindered rather than helped the study of relationships, in psychology. There are, of course, no right and wrong methods; methods are simply more or less adequate to the data and the problem in hand. One could almost say, dogmatically, that the particular type of data and problem to which the Pearson method is applicable were relatively rare. Most practical problems require a two or threefold division.

the theory of the 'normal curve.' Once more the writer would object to the relevancy of the information to the practical problem, and on two counts.

(a) Construction of the scale so that equal increments of ability are related to equal increments in score means, probably, transmutation of values in terms of the per cent. passing different items into positions on the normal curve or some such procedure.<sup>1</sup> It need only be said here that items which give a satisfactory scaling on such a curve need by no means be the most diagnostic items. An item may appear in a test because it is the only item appearing at 1.52 P.E. (when scaled as mentioned above) or it may appear in a test because most of the potential failures cannot pass it and most potential successes can. The last criterion is obviously the fundamental one if the problem in hand is to obtain a test that shall most completely differentiate the potential failures.

(b) It may also be pointed out shortly that for the particular practical problem under consideration a normal distribution of scores is hardly to be desired. If a scale sets off the potential failures very completely, it will lump the assured failures at the bottom and the assured successes at the top, and spread out the questionable cases in between. In short, equal increments of ability and a normal distribution of scores are *not* to be desired if the greatest efficiency, for the practical problem postulated, is sought.

#### IV. DISCUSSION

Well—most of these points seem obvious enough, perhaps. But the concept back of them indicates a fundamentally different statistical attack, in the development and use of tests. If differentiation of the potential failures in high school is an important problem, why not build a scale specifically for that purpose? Select items simply according to their ability to make the desired division. Combine those items so that such a lumping of cases at the two extremes is obtained; the reverse of the normal distribution is the distribu-

<sup>1</sup>Of which procedures, transmutation of percents passing at different chronological ages into supposed units of mental growth is surely more questionable still.

tion to be desired.<sup>1</sup> Then measure the value of the test by measuring its 'efficiency' in dealing with the practical problem for which the scale has been designed. Deal with each important problem in some such empirical and concrete fashion. And, if, out of a large number of such attempts, there emerge certain unitary factors,—a general ability, a series of character types, or what not,—well and good. But the postulation of such elements in advance, with verification primarily by reference back to these postulates, is both an unscientific and a practically dangerous proceeding.

First a very specific problem; then, after that—*everything subservient to the solution of it!* Every item chosen with reference to that one problem, every method aiming only at the most direct and empirical solution of that problem—no hypotheses, as thoroughly empirical treatment as may be! The result will be, the writer believes, an essentially new statistical approach (methods now in use suggest something of this sort, particularly the methods used in the development of the army trade tests). Such a revision of methods is, the writer has come to feel, necessary, for a clarifying of the total situation.<sup>2</sup>

<sup>1</sup> Is this not really the solution of the problem of the normal curve in mental measurement? (See, for instance, Boring, *Amer. J. of Psychol.*, January, 1920). The actual distribution of various traits is a matter of academic interest only. But meantime, the distribution to be sought in test work will be determined by the problem.

<sup>2</sup> And now the apology! There is little essentially new in the paper, of course. (In fact, it should be said that a detailed discussion, with full use of the literature, was first attempted, but was found to extend beyond reasonable limits.) The important thing, however, is the total implication of the various points presented. Our statistical methods as applied to tests have been largely borrowed methods,—and methods borrowed from the descriptive sciences. So the question has been: What is the test measuring, and how accurately is this thing being measured? But mental testing is not a descriptive, but a technical science. And the question should be, instead: What are we trying to do, and how well are we doing it? The distinction is, the writer believes, of the very most fundamental importance, involving fundamental differences in statistical approach.

It remains to be mentioned that the points made apply equally to measures of achievement in the school subjects or other like tests. Instead of measuring "ability in arithmetic" in the eighth grade—and then commenting mildly on the extent to which arithmetical ability in the eighth grade overlaps on the seventh, why not tackle a definite practical problem,—attempt to define the passing point in arithmetic for the eighth grade? The distribution, again, should be bi-modal, not normal,—and the other points mentioned follow.