INCREASED RELIABLE GRADES AND VALIDATING TESTS

HENRY D. RINSLAND, University of Oklahoma, Norman

Generally speaking the correlation between test scores and grades are low. Undoubtedly one of the major uses of tests—most kinds of tests—is prediction. The actual magnitudes of correlation coefficients which usually appear in the literature upon which predictions are made vary from .25 to .70, with the vast majority running between .30 and .60. For most counselors and predictors these correlations are usually considered very low. If the coefficient of correlation is interpreted in terms of the coefficient of alienation, it is necessary to have an r of .866 to have a k of .50 which, subtracted from 1, gives an improvement of prediction and proficiency which is only 50% better than pure guess; really not a satisfactory prediction because the coefficient of alienation actually measures the residual deviations about the regression lines, but according to this an r of .866 is not high—just "half high."

Recently, Jackson and Phillips (2) have pointed out that these predictions can be improved if we will establish definite lines of passing and failing and predict in terms of deciles. Those of us who were in the designing and test construction business in the Army are acquainted with the predictions or chances in a hundred that a man receiving a certain score will achieve average or better in training, as described in *Personnel Classification Tests*, 1942 and 1946, War Department. Those are about the same techniques as in the Russell-Taylor tables (5). Although this situation has been pointed out by a Committee of the American Council on Education (1), it appears from their report that little has been done to use these methods, or what has been of greater concern to this writer since 1937, no improvement in methods of grading has been adopted by schools.

In 1937 the author (3) pointed out that the cumulation of points on highly reliable subjective tests, highly reliable objective tests, readings, reports, problems and the like, can be converted into essentially reliable letter grades of *A, B, C, D* and *E*. He issued the first military bulletin on this in December, 1941, as an instruction manual for the Adjutant General's School under the title, *Test Construction and Grading in Military Schools*. Each letter grade is defined as approximately one standard deviation wide; specifically one standard deviation for *B, C,* and *D*, but for *A* from $+1\frac{1}{2}$ S. D. to infinity, and for *E* from $-1\frac{1}{2}$ S. D. to infinity—a score scale which is called a "quintine", similar to the Army Air Forces' stanine, although his R-score of 0 to 100 running from $-2\frac{1}{2}$ S. D. to $+2\frac{1}{2}$ S. D. is much preferred in accurate scaling. He reported that median coefficient of reliability of semester grades range from .88 to .91. Since then he reports (4), by use of the Kuder-Richardson formula, reliabilities ranging from .86 to .94. These reliabilities are within the respectability of standardized tests used in high school and college predictions.

For some time the writer has held that if the criteria, which are usually high school or college grades, can be raised to the respectability of the reliability of our standardized tests, then the prediction of grades from tests would be greatly improved and the validity of tests whose elements depend upon correlations with grades (Toops-L technique, and others) will thereby be greatly improved. This is undoubtedly one of the most important next steps in the validating of almost all types of intelligence and aptitude tests used to predict achievement.

If the improved statistical predictions as mentioned in the above introductory paragraph can be applied, predictions for guidance and for validating tests will be greatly raised as a consequence. To overcome the present low reliability of grades in the correlation of tests with grades, the usual formula was altered for the correction for attenuation by dividing the correlation by the square root of the unreliability of the criterion. So, if an obtained corre-

lation between tests and grades is, say .77 and the reliability of the criterion is .60 the estimated true correlation is 1.00. For practical prediction purposes this is useless. Nothing except improved reliability of grades will ever correct the gross error in the measurement of school and college work. The present traditional system is, to use students' language, "grossly unfair."

However, nowhere in the literature are tables set up for these predictions for varying coefficients of correlations based upon normal distribution. To accomplish this, the statistical definitions of letter grades A, B, C, D and E must be the same for both tests and grades, and, therefore, the definition of these letter grades as recommended by the writer must be applied to tests. Of course, for most tests this is very simple as means and standard deviations are published. However, the determination of theoretical tables of prediction is not established.

Such tables have been calculated* and will be published in the forthcoming revision of *Test Construction and Grading* under the title *Evaluation, Testing and Marking*, (Prentice-Hall). A comment or two is in order to make clear the construction of these prediction tables since all data are based upon the assumption of theoretical frequencies of a normal bivariate surface. It is necessary first to reduce both the regression equation and the standard error of estimate to a standard score form where $\overline{Y} = rX$ and *S. E. (est.)* is equal to *S. D.* $\sqrt{1 - r^2}$.

The second point of observation is to read from a table of area of normal curve the percentage of cases that will fall within the following respectively named class intervals of approximately 1 S. D. width: *A*, +1.5 to infinity; *B*, + .5 to + 1.5; *C*, (— .5) to .5; *D*, (— 1.5) to (— .5); *E* or *F*, (— 1.5) to minus infinity where we find for 1,000 cases respectively 27, 242, 383, 242, and 67 (strictly speaking, 1001 cases).

For a given correlation then the number of cases is calculated in the class interval *A* of the test for 67 cases who will probably achieve the grade of *A*, *B*, *C*, *D* or *E*. Assuming a correlation of .80, knowing that the mid-point of the *A*, the interval from +1.5 to infinity or +2.5, is 2, the predicted *Y* is 1.6. The standard error of estimating is $\sqrt{1 - (.80)^2}$ or .6.

In order to use a normal curve table of areas, the *Y* variate must be changed to standard values. The $z = \dfrac{Y - 1.6}{.6} = .166$. From this interval to infinity a normal table of areas lists 56.4% of our 67 cases or approximately 38 cases; meaning that 38 out of 67 cases who made *A* on the test would be likely to make a letter grade of *A*. To continue the same operation for the other letter grades, it is found that the number of cases are 27 for *B*, 2 for *C*, and none for either *D* or *E*.

Taking another illustration with an *r* of .60, the investigator would have predicted letter grades for the 67 people making *A* on the test as follows: 24 *A*'s 31 *B*'s, 12 *C*'s, 1 *D* and no *E* score.

These illustrations show very definitely, in terms of percentage of cases for correlation .80 which is now extremely high, and the correlation .60 which is now high under the traditional grading systems, that these predictions are necessarily better than one could infer from any generalization of the coefficient of alienation. But this is exactly what is needed in all predictions from a given score on a test or battery of tests whose scores might be summated by the multiple regression equation. The researcher wants to know, what are the chances of making the respective letter grades *A*, *B*, *C*, *D*, and

*The writer is indebted for suggestions in making these calculations to Robert W. B. Jackson, Toronto University, 1950.

E; or if it is desirable to define failures as being all cases below —1.5 standard deviation (there is here a cut-off point which is a fair assumption where the concept of failure is actually employed). However, with a calculable letter grade, a *z* score, or an *R*-score as proposed by the writer, a given faculty can vote on a failure line and thus make failure statistically calculable and fairly uniform from instructor to instructor and even from school to school. Toops (6) proposed a method by which a university could convert grades from all high schools into comparable standardized grades by knowing the correlation between high school grades and some standardized test such as an intelligence test or a general college ability test, or as he illustrated between such grades and the Ohio State Physchological Examination.

By way of summary, it is proposed that first of all an accurate grading system be set up which will approximate in accuracy the reliability of standardized tests used to predict college grades, say .90 or above; and secondly that tables of estimate be calculated for predicting these grades from tests for a large number of coefficients of correlation fom say .30 to .95. Two distinctive improvements would result: one proven by the writer which is the increased reliability of grades approximating the reliability of standardized tests; the other improved validity of tests used to predict these more reliable grades.

BIBLIOGRAPHY

1. American Council on Education. 1949. Predicting success in professional schools. Washington: The Council.
2. Jackson, Robert W. B. and A. J. Phillips. 1945. Prediction Efficiencies by deciles for various degrees of relationship. Department of Educational Research, Bulletin No. *11*, University of Ontario, Ontario.
3. Rinsland, Henry D. 1937. Test construction and grading. New York: Prentice-Hall, Inc.
4. ————. 1949. A clinical method of grading. Norman, Oklahoma: University of Oklahoma Book Exchange.
5. Taylor, H. C. and J. T. Russell. 1939. The relationship of validity coefficients to the practical effectiveness of tests in selection. J. of Applied Psychol. *23*: 565-578.
6. Toops, Herbert A. Ohio College Bulletin No. *88*, Ohio Psychological Association. Columbus: Ohio State University, undated.