

THE JOURNAL OF EDUCATIONAL PSYCHOLOGY

Mental Tests as a Means of Selecting and Classifying College Students

AGNES L. ROGERS

Goucher College.

IN March and April, 1919, a series of mental tests was applied to a group of 98 Seniors and 182 Freshmen at Goucher College. The purposes in view were threefold:

(1) to determine their reliability as measures of mental capacity for college women.

(2) to weigh their worth as indices to future academic success.

(3) to establish in the event of their proving satisfactory in the foregoing respects adequate standards both for the selection of candidates for admission and for the classification of entrants in the various divisions of the larger courses, required and elective, in accordance with capacity.

The tests were applied by the writer to both Seniors and Freshmen simultaneously in one large group of three hundred students approximately. The tests were distributed and collected and general supervision was exercised by ten other college instructors.

The Thorndike test of Mental Alertness¹ was used along with the Roger's Interpolation test² and the Rogers' Reasoning³ test. Series

¹The Thorndike Mental Alertness Test with instructions for giving and scoring can be obtained from the Bureau of Publication, Teachers' College, Columbia University.

A of the former was applied in March and Series K in April. Alternative forms of the other tests were likewise given on the second occasion. The Thorndike Tests are comparable in general character to the Army Scale Alpha. An enumeration of the names of

¹A description of these tests and their method of application and scoring is given in *Experimental Tests of Mathematical Ability and Their Prognostic Value*, Teachers College, New York City, 1918.

the tests included will indicate the extent of similarity between them. They are as follows: the first is a directions' test; the second, a series of disarranged sentences; the third, an arithmetic test consisting of a series of reasoning problems; the fourth, an arithmetic test in the four fundamentals; the fifth, an information test; the sixth, a synonym-antonym test; the seventh, a selective judgment test in which the task consists in indicating which of four things should be done in a given situation or which of four alternatives is the cause of a given course of action; the eighth, an interpolation test; the ninth, an analogies test; the tenth, a test in which the task consists in perceiving and indicating the largest and smallest numbers in a series of columns of numbers; the eleventh, a test in the detection of absurd statements; the twelfth, a test in logical inferences; the thirteenth, a test in the recognition of spatial forms.

These had been applied to candidates for Officers' Training Schools for the Air Service. They also were used as one feature of the Emergency Admission plan of the S. A. T. C. and seemed better adapted a discrimination between higher degrees of ability than the Army Scale Alpha. Their power to determine the mental level of individuals of superior mentality had already been partially demonstrated, and tentative standards for the particular groups examined had been determined. They were, therefore, applied now with the expectation of their proving a reliable means of measuring general mental alertness and aptitude for college work.

Another conspicuous advantage of these tests was that there were available fifteen duplicate copies, genuine alternative forms, equal in difficulty, but differing in content, thus making it possible to repeat the application on generations of Freshmen over a period of years and facilitating the standardization of norms for college women. The tests will be re-applied with such improvements as this application suggests at Goucher College, in October, 1919. This article therefore describes only one stage of the investigation, on which further reports will be published later.

Each application of the entire group of tests occupied 45 minutes. The conditions under which the tests were administered, though comparable to those used in applying the tests to candidates for army schools, seemed to the writer to fall short of what the object in view demands. In order to secure reliable measures of individual

ability the groups tested should be smaller. Where three hundred persons are examined simultaneously, the discipline must inevitably become more rigorous and consequently more emotionally disturbing. While it is true that more than merely intellectual functions are represented in the scores derived from tests of intelligence, it is nevertheless desirable that we should approximate as closely as possible to customary examination conditions and accordingly groups not exceeding 50 should be examined together. The reliability coefficients obtained would probably have been greater, if such conditions had prevailed. When calculated by the Pearson method, they were, however, as indicated in Table I.

TABLE I.

	Seniors		Freshmen	
	r1	r2	r1	r2
Thorndike Test of Mental Alertness,76	.86	.66	.80
Rogers Interpolation Test,76	.86	.75	.86
Rogers' Reasoning Test,54	.70	.52	.68

r1 is the reliability coefficient or coefficient of correlation between the two applications of the tests.

r2 is the reliability coefficient for the two applications combined.

$$r2 \text{ equals } \frac{2r1}{1+r1}$$

These coefficients cannot be considered satisfactory, particularly when it is remembered that our ultimate aim is individual measurement. Even if we admit that greater reliability coefficients would result from improved conditions of administration, still a more trustworthy gauge of the capacities tested seems necessary. The very brief period of testing is a partial explanation of the imperfect character of the tests as measuring rods. Their reliability would be increased by merely extending the duration of the time of testing. To determine to what extent these same tests would need to be lengthened to produce reliable results, use was made of the formula sug-

$$\text{gested by Brown}^8, \text{ namely } r_n = \frac{nr_1}{1 + (n - 1)r_1}.$$

In order to obtain a reliability coefficient of .95, the Thorndike test, if applied under similar conditions, would have to be extended to six times its present duration, that is six of the forms would have to

⁸BROWN, WILLIAM. *The Essentials of Mental Measurement*. Cambridge University Press, Cambridge, 1911. 101-102.

be given and the results pooled to secure such a measure of the student's ability as would yield a reliability coefficient of .95. Similarly for the Rogers' Interpolation, a test six times as long would be required for college women, whereas in the case of High School Girls, examined in smaller groups, a reliability coefficient of .94 was obtained from this test.⁴ For the Reasoning test apparently at least sixteen applications would be necessary, although when applied to High School Girls, this test yielded a reliability coefficient of .73.⁵ Thorndike's claim that "about 30 minutes of fore-exercise and about 200 minutes of test" is necessary to obtain an individual's true status in comparison with other individuals in a "standard test of intelligence" is supported by our results.⁶

As a measure of academic accomplishment the grades obtained by the students in all courses were averaged. The usual practice is to record these in literal form, the scheme in use at Goucher College being to assign 3 per cent. of the students in both required and elective courses to grade A, 22 per cent. to grade B, 50 per cent. to grade C, while to grades E and F respectively in elective courses 22 per cent. and 3 per cent., and in required courses 15 per cent. and 10 per cent. are assigned. In the Goucher Bulletin it is stated that these letters, A to F correspond with the work frequently described as excellent, good, average, poor, conditioned and failed. In the case of an F grade, the course in question must be repeated. An E grade indicates that certain work remains to be completed. If a student fails to remove a condition on a course by the time set, she is regarded as having failed in that course. In converting these literal measures into numerical terms it was decided to have one system of grading for both required and elective courses, to count all conditions that were later removed as D and all unremoved conditions as F. In this way were obtained for Senior and Freshman measures of their accomplishment in academic work. For the Seniors, the academic mark represents the composite result of approximately 30 separate grades and for the Freshmen 4-6 grades.

⁴ROGERS, AGNES L. *Experimental Tests of Mathematical Ability and Their Prognostic Value*. Teachers' College, Columbia University, N. Y. City, 1918. p. 45.

⁵ROGERS, AGNES L. *Experimental Tests of Mathematical Ability and Their Prognostic Value*. Teachers' College, Columbia University, 1918. p. 45.

⁶THORNDIKE, E. L. *Tests of Intelligence, Reliability, Significance, Susceptibility to Special Training and Adaptation to the General Nature of the Task*. School and Society IX, 216. February 15, 1919.

To determine the extent to which the test results are symptomatic of academic attainment, two methods were used. First, the degree of correspondence between the two sets of measures obtained from the tests and from averaging academic grades was calculated. Secondly, the extent to which students maintain their place in certain divisions of the group in the two sets was computed. Pearson coefficients of correlation are presented in Table II, which reveal the amount of interrelation between the tests and college marks.

TABLE II.

	Academic Seniors		Success Freshmen	
	r	P. E.	r	P. E.
Thorndike Test of Mental Alertness Series A,	.42	.056	.37	.043
Thorndike Test of Mental Alertness Series K,	.39	.058	.37	.043
Thorndike Test of Mental Alertness Series A and K, combined,43	.054	.40	.040
Rogers' Interpolation 1,36	.059	.41	.040
Rogers' Interpolation 1a,31	.061	.29	.050
Rogers' Reasoning 1,21	.065	.23	.048
Rogers' Reasoning 1a,20	.065	.12	.050

It is clear from these coefficients that the independence of the functions measured by the Mental Alertness and the Interpolation tests and by college marks is substantial. Between the Thorndike tests and academic grades the coefficient obtained is at least eight times the amount of the probable error and for the Interpolation tests it is not less than five times. As the shortness of the period of testing for the individual tests was believed to lower their reliability and this would tend to decrease the coefficients of correlation between tests and academic grades, various combinations of the tests were made after the manner suggested by Woodworth⁷ and the amount of correspondence between these composites and college marks was computed for the Seniors. These are presented in Table III.

TABLE III.

	Academic		Success	
	r	P. E.	r	P. E.
Thorndike Test of Mental Alertness Series A and K combined,	.43	.05		
Thorndike Test of Mental Alertness A and Rogers Interpolation 1 combined,43	.05		
Thorndike Test of Mental Alertness K and Rogers' Interpolation 1a combined,90	.01		
Thorndike Test of Mental Alertness A and K and Rogers' Interpolation 1 and 1a combined,43	.05		

⁷WOODWORTH, R. S. *Combining the Results of Several Tests*. Psychological Review, XIX: 97.

With the exception of the composite obtained from the Mental Alertness Test K and Interpolation Test 1a these new composites yield coefficients but little larger than the Thorndike tests alone. Where the correlation obtained is so small, it would be unpermissible to use the method of the regression equation to predict even *on the average* future performance at college from achievement with the tests, much less the *individual's* probable status. The standard error made in using the regression equation to estimate academic success from mental performance, when Series A and K of the Thorndike Test are combined is in the case of the Seniors .45. The total range in academic grades for the Seniors is 2.13. This means that the probable error of our estimate is approximately one-fifth of the difference between the best and the poorest Senior. Since practical certainty is attained only between limits of 4 or 5 P. E., which in this instance covers the entire range almost, we cannot rely upon the tests as measuring instruments of individual promise of college students.

When we represent graphically the correlation between the Mental Alertness Test, Series A and K combined and College Marks as in Figures 1 and 2, this becomes very apparent. The wide variability in academic success corresponding to any particular degree of achievement in the tests is obvious.

The limitations of the tests as indicators of academic success are even more striking when we use the second method and consider the median, tertile, and quartile retention. The percentage of students retaining their place in the same halves of the distribution is shown in Tables IV and V, which indicate roughly the amount of displacement existing. Evidently we would have been wrong four times out of ten, if we had used the tests to predict academic success even so roughly as this classification implies.

TABLE IV
Median Retention of Seniors.
Academic Success.

Mental Alertness Test, A and K		1	2
	1	29	20
	2	18	31

Median Retention=60 or 61.2 per cent.

TABLE V
Median Retention of Freshman.
Academic Success.

Mental Alertness Test, A and K		1	2
	1	58	34
	2	35	57

Median Retention=113 or 62 per cent.

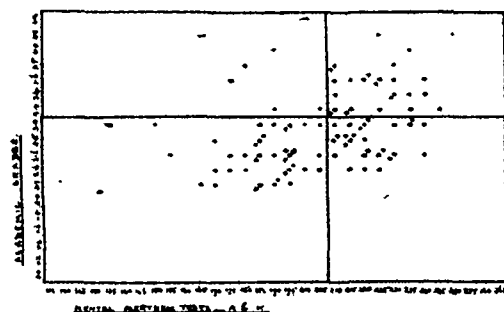


FIGURE 1

Scatter diagram showing the relation between the scores obtained by the Seniors in the Mental Alertness tests, Series A and K combined and their academic grades.

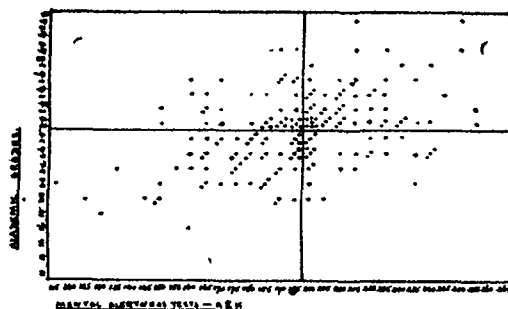


FIGURE 2

Scatter diagram showing the relation between the scores obtained by the Freshmen in the Mental Alertness tests, Series A and K combined and their academic grades.

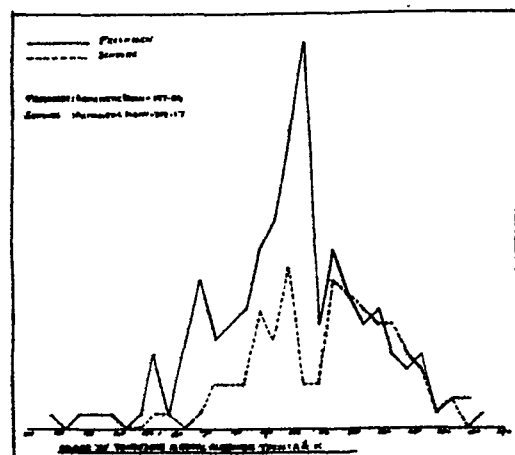


FIGURE 3

This result is emphasized by the more exact and detailed analysis of the same data obtained by calculating the actual number and percentage of students occupying the same quartile in both sets of measures. This is done in Tables VI and VII.

TABLE VI.

Quartile Retention of Seniors.
Academic Success.

Mental Alertness Test, A and K

	1	2	3	4
1	12	7	8	3
2	5	7	9	4
3	4	7	5	7
4	3	8	8	11

Quartile Retention=35 or 35.7 per cent.

TABLE VII.

Quartile Retention of Freshman.
Academic Success.

Mental Alertness Test, A and K

	1	2	3	4
1	16	13	9	7
2	13	14	12	7
3	11	13	10	11
4	5	6	11	24

Quartile Retention=64 or 35.1 per cent.

To predict an individual's probable status in academic work from his performance in the tests would obviously be rash. In this instance we would have erred in such a rough approximation to measurement as assigning a student to a definite quartile, sixty-five times in every hundred in the case of both Seniors and Freshmen. It is not even true that students in the highest quartile in college scholarship records never get into the lowest quartile in the test results nor that students in the lowest quartile in college records never get into the highest quartile in the test results.

A division of larger classes into three sections is not only more customary, but better justified on grounds of the distribution of mental capacities in the form of the normal frequency curve, in which there is observable a tendency to cluster around one type, small deviations from that central tendency being frequent and large deviations rare. The amount of tertile retention has therefore a special interest for us from a practical standpoint and is presented in Tables VIII and IX for Seniors and Freshmen respectively.

TABLE VIII
Tertile Retention of Seniors.
Academic Success.

Mental Alertness Test, A and K		1	2	3
	1	18	8	6
	2	10	12	11
	3	4	12	17

Tertile Retention=47 or 47.9 per cent.

TABLE IX.
Tertile Retention of Freshman.
Academic Success.

Mental Alertness Test, A and K		1	2	3
	1	34	13	13
	2	25	19	18
	3	13	17	30

Tertile Retention=33 or 45.6 per cent.

In this classification likewise we would be in error in 52 per cent. of the cases as far as the data derived from the Seniors are concerned and in 54 per cent. of the cases as shown by the results obtained from the Freshman. Speaking roughly we can state that every other assignment would have been an unwarranted assignment.

That these Mental Alertness tests requiring only an hour of time have some value, however, for purposes of classifying students in accordance with capacity as seen from a comparison of these results with those that are to be expected by mere random assignment of students to sections. By the latter method we have one chance out of two of assigning a student to the correct half of the group, one out of three of assigning her to the correct tertile and one out of four of assigning her to the correct quartile. The corresponding chances, where we make use of the Mental Alertness tests, Series A and K combined, employed in this investigation are approximately three out of five, two out of five and one out of two. To appraise adequately the practical value of the tests we would have, however, to compare their power to classify with that of the previous school record.

There are of course many reasons why perfect correspondence between ability as measured by the Mental Alertness tests and ability as measured by college grades should not be found. One obvious cause of the low correlation is that we are here dealing with a selected group, from which presumably those unfit to profit by college study have largely been eliminated. The Seniors have certainly

passed through a prolonged winnowing process. A comparison of the distributions for Seniors and Freshmen presented in Figure 3 gives evidence in support of this. It will be noted that the group average for Freshmen is less than that for Seniors and that the range towards the low end of the scale is greater for the former than the latter. Only 26.3 per cent. of the Freshmen reach or exceed the median for the Seniors. This may be partly due to chance errors of measurement, but it also indicates a constant cause at work, namely the elimination of the weaker students during the four years of the college course. The fact that both groups represent a selection of college candidates justifies the statement that the actual correspondence between success with the tests and academic promise is higher than the coefficients of correlation presented in Tables II and III indicate.

Furthermore it is obvious that the mental abilities underlying the two sets of measures—those derived from the tests and those based on academic records are not identical. The tests gauge predominantly innate intellectual dexterity, whereas college marks rather furnish measures of progress in learning, into which there enter to a very great extent emotional and moral elements. The conscious possession of purposes and the degree to which they dominate the mind are powerful factors in determining academic success. The extent to which the student identifies herself with the college class-work is as essential an element in the grades she receives as the intellectual acuity or skill she possesses. It has to be admitted that the mental tasks demanded by these tests are humble. Given sufficient time college women could master all of them readily. The emphasis is on speed rather than on difficulty and in so far as this is true they fall short of providing an ideal gauge of intellectual power. While well adapted to younger or less carefully selected individuals, they are, it would seem, too simple for use in classifying students on the basis of mental capacity. There is a general law that the harder the test, the greater is the scatter. The weak fall more conspicuously and the able succeed more markedly. To distinguish between the higher levels of intelligence we need therefore harder tests, involving more searching intellectual tasks.

Moreover, as has already been pointed out, the tests used in this investigation do not measure with sufficient accuracy the traits

they gauge. Their reliability coefficients show to what extent they fail in this respect. It is equally true that academic grades are imperfect, by no means determining exactly the abilities they purport to measure. It has to be borne in mind that college marks are not assigned to all students by the same instructors even in the same subjects of study and despite the standardization of grades secured by the use of a defined system, such as the Missouri system of grading, still misplacements are likely to occur. Students are graded by different instructors in different years and this tends to lower the accuracy of the ratings made and so far these compare unfavorably with the tests.

Certain capacities essential in academic work are undoubtedly measured by the tests and in spite of their defects they can be of service in two respects. In the first place they are superior, even if only slightly so, to haphazard guessing as a basis for allocating students to sections on grounds of mental capacity. Using Series A and K combined of the Mental Alertness test we have roughly one chance in two of making a proper assignment, whereas by haphazard selection we have but one chance in three. In the second place, they are of value in determining a lower limit, which when coupled with all the other information about an applicant to which the college has access, can reinforce a judgment as to fitness to undertake a college course. Thus it is practically certain that a student failing to attain a score of 125 in Series A and K combined lacks the ability that college work demands and should be discouraging from entering. Her efforts would be applied to better effect in other pursuits.

This investigation points to the directions in which the tests are susceptible of improvement. The period of testing must be lengthened. More capacities and more fundamental capacities should be measured. The student must be confronted with more subtle problems and more intricate tasks, so that intellectual power rather than speed is emphasized in the scores obtained. Smaller groups should be measured simultaneously so that the conditions of application are as normal as possible. Provided such improvements are made, the tests give promise of being a helpful supplementation of existing methods of selecting college entrants.

The need for a more satisfactory method of determining fitness to pursue a college course is great. Changes in College Entrance Ex-

aminations, notably by four of the colleges for women, have led to renewed interest in the problem. Columbia and other colleges for men have recently adopted psychological tests. It would seem incumbent that this new instrument should itself be carefully tested before it is accepted as the nonpareil and the satisfactory solution of a difficult situation.