

PRACTICE EFFECTS IN INTELLIGENCE TESTS¹

BY KNIGHT DUNLAP AND AGNES SNYDER

The Johns Hopkins University

The possible practice effects in intelligence tests are of importance where groups of reactors are to be given relative gradings on tests. If some of the reactors have had experience in the taking of tests; especially tests of the same general sort as that on which the grading is to be done; while others have had less experience, or none at all, with tests of the type used; it is obvious that the ratings will be unfair, unless no considerable practice effects can be demonstrated. The matter is of especial importance in connection with tests for admission to college or school, since the possibility of 'coaching' for these tests must be definitely known, or else the tests are misleading.

For the purpose of a preliminary investigation into practice effects, a college class of 44 men, mostly seniors, was tested four times, with intervals of approximately three weeks between the tests, using four forms of the Army "Alpha" composite test. The Alpha test was selected not because of any assumed superiority of that test over others, but because a number of nearly equivalent forms of that test were available, whereas there were at most only two forms of any other composite test available, with the exception of one standardized test for which the price asked was nearly a dollar each, rendering it prohibitive for experimental work.

The 'practice' in this investigation consisted solely in the taking of the tests: no attempt was made to give further practice between the tests, although this might have been done by using various single tests corresponding in type to the parts of the Alpha composite. No attempt has been made so far to find what effect practice on tests of one type

¹ A paper presented before the Southern Society for Philosophy and Psychology, New Orleans, April 23, 1920.

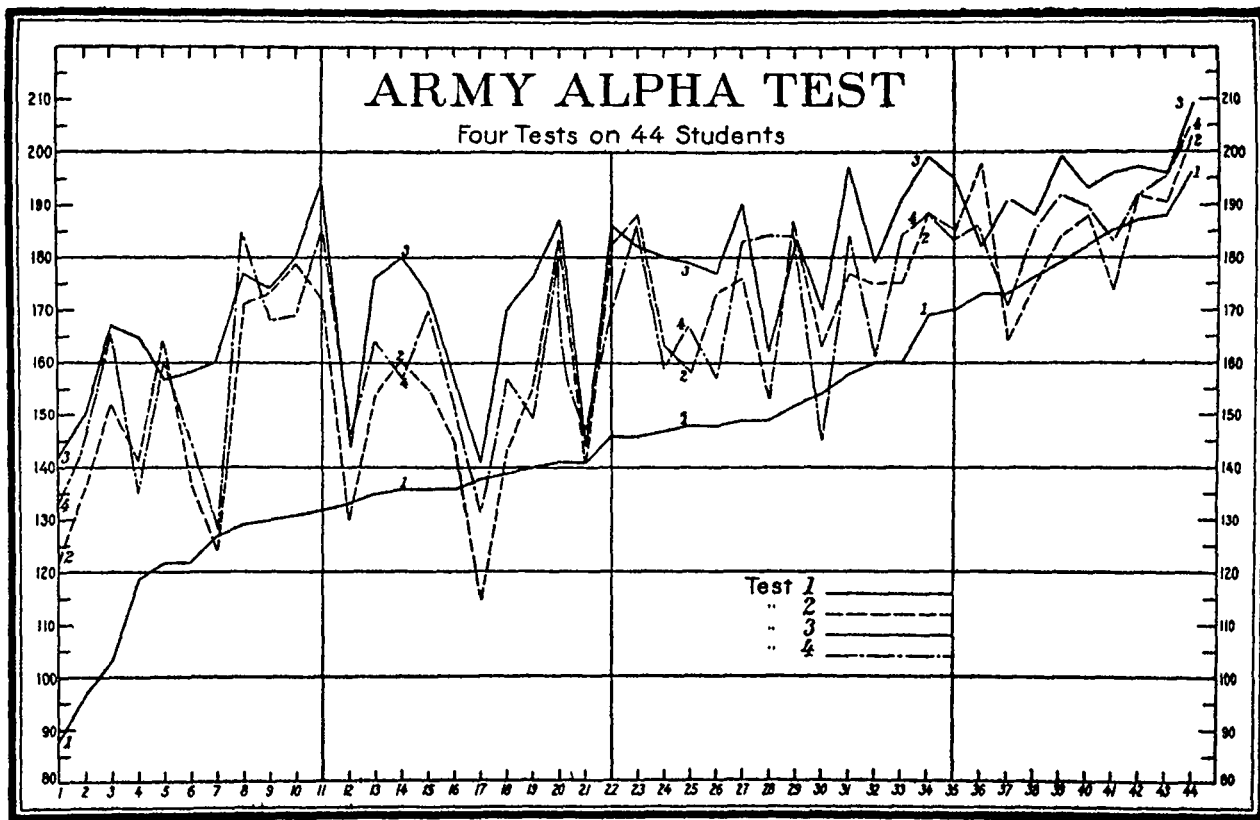


DIAGRAM 1. The Abscissæ numbers, from 1 to 44, represent the men, arranged in order of grades attained on the first test, the Graph for which is marked 1. The Ordinate numbers, at the left, indicate the scale of markings in absolute scores on the test, not percentages. Graphs 2, 3 and 4 represent the grades received by the men on the second, third and fourth tests respectively.

exercises on tests of a different type. The purpose has been merely to obtain data under conditions which might most easily occur; namely, where persons expecting to submit to a certain sort of group-test might obtain previous forms of that test and study them or work through them for practice.

The results of the four tests are presented in Diagram 1. For this diagram, the members of the class were arranged in the order of their grades on the first test, from low (left) to high (right). The grades for the second, third and fourth tests were then plotted for the same arrangement. The reason for the specific arrangement is that thereby one fairly simple graph is obtained, making comparison with the graphs of the succeeding tests easy.

It will be noticed that there is a general improvement on the second test, and again on the third test, but a general falling back on the fourth. While five men on the second test fell below their marks on the first test, and one other made exactly the same mark on both tests; and five men on the third test fell below their marks on the second; every one made better marks on the third test than on the first. That is: not only was there a striking general improvement from the first to the second, and from the second to the third, but *every man* showed improvement on the third test, over the first.

The improvement on the second and third tests, and the slip-back on the fourth, is shown in another way in Diagram 2, where the ordinates represent the number of men making improvements, or losses, of magnitudes represented by the abscissae.

The main reason for the deterioration on the fourth test was made clear by the reports of the reactors. There was a practically unanimous report that the taking of the first, second and third tests was interesting; but that the fourth was a bore. The fact of lessened interest I inferred also from details of the behavior of the class during the test, before I obtained any reports. Of course, the fourth test should have shown at least as high an average grading as did the third if anything had depended on the test beyond the mere

personal ambition to make a good showing, which had been largely satisfied by the preceding tests.

It is probable that the four forms used were not exactly equivalent for this particular group of men. The most likely part in which differences would be found might be assumed to be the part on General Information; and the details of the average gains and losses on the different parts of the test (Diagram 3) indicate that the General Information

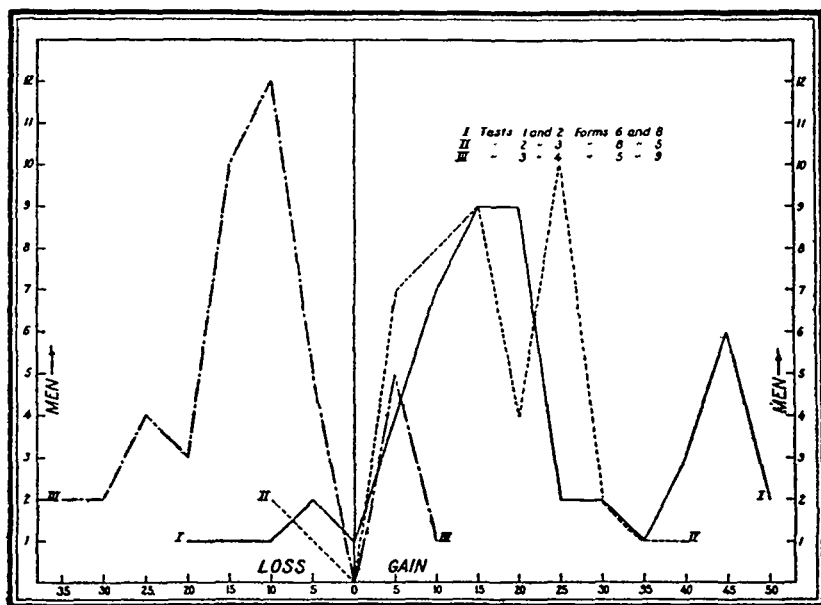


DIAGRAM 2. The abscissæ here represent the number of points of gain or loss between successive tests, the ordinates representing the number of men making a gain or loss not greater than the abscissa at the point of erection of the ordinate, but greater than the next smaller abscissal value marked. Graph I., for example, indicates that two men lost not more than five points on the second test (as compared with the first); one man lost more than five but not over ten; one lost not over fifteen, and one not over twenty.

part of the fourth test was indeed more difficult than the corresponding parts of the other forms. The probability of consequential differences in the other parts of the test is less definitely determinable; but since the excessive loss on the general information part of the fourth form accounts for only

a part of the general loss, the apathy or lessened interest of the class may be tentatively assumed as the most plausible major cause for the general decline on the fourth test.

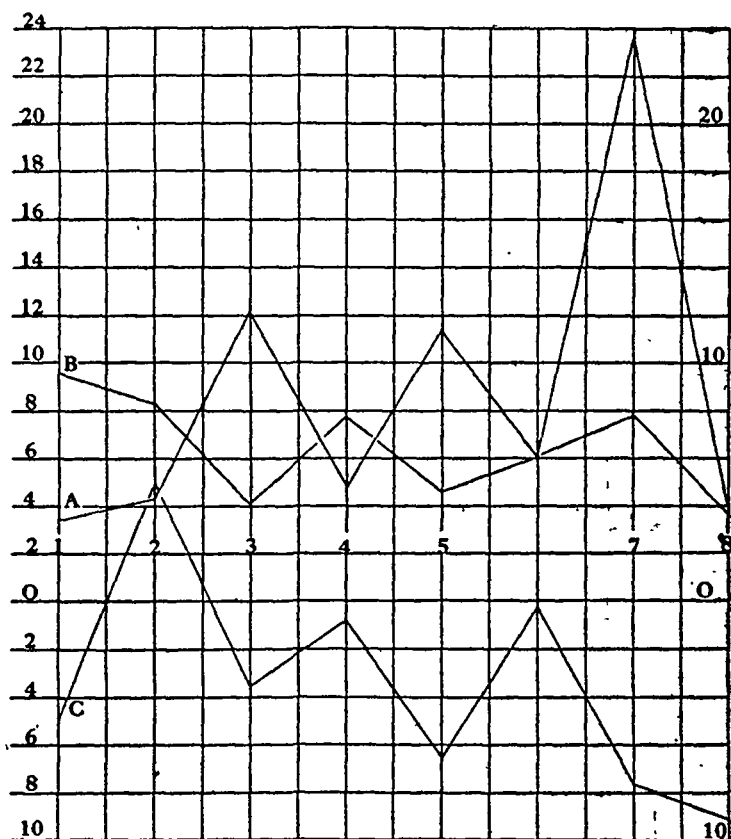


DIAGRAM 3. This represents the gains and losses, in percentages, on the eight parts of the composite test, the percentages being represented by the ordinates, and scaled at the left. The eight points marked on the horizontal axis represent the eight parts of the test. Graph *A* shows the average percentage of gains of the class on the second test as compared with the first; graph *B* represents the average percentage of gains on the third test as compared with the second; and graph *C* represents the gain and losses on the fourth test as compared with the third.

A detailed study of the gains and losses on the several parts of the test has given perplexing results. The close similarity of the corresponding parts of the four forms; especially the second part (problems in arithmetic), in which

the same wordings are used in some of the corresponding items of the several forms; and in the fourth part (number series), in which some identical number-series occur; may account for the significantly higher gain in these parts as compared with other parts of the tests. The problems of transfer of training are sharply presented here. The important question is whether repetition-forms of tests can be constructed in such a way as to test the same 'capacities,' and offer the same degrees of difficulty, without permitting a large amount of transfer. It is obvious that a considerable amount of experimental work must be done before tests can be used with assurance as to the absence of important practice effects.

Subject to the foregoing considerations, the following interesting points may be indicated:

In the first place, the relative rating on the third test is somewhat different from that on the first. If the tests had been used to select the best fourth of the group there would have been but little difference in the individuals selected, since 9 of the 11 who were in the first (high) quartile in the first test remain in the first quartile in the third, while the two who go back move back but to the next quartile. The tendency of movement backward or forward to adjacent quartiles is marked in the three lower quartiles. While this has no direct bearing on the present problem, it is a matter of passing interest to note the comparative uniformity in the results of the first and third tests in the first (high) quartile and the contrasting fluctuation of results in the fourth (low) quartile. Were the test given with a view to eliminating the lowest quartile the first and third tests would thus yield quite different results.

In the second place: if the men in the lower half of the class (as indicated by the first test) had had practice in the test equivalent to that actually obtained by the taking of the test twice, and the men who are in the upper half had not had practice, a single test on the class would have given a relative grading which would have differed markedly from either the first or the third of the actual ratings, and which (in all probability) would have been maximally unfair.

The practical conclusions to be drawn from such data are extremely important. The question as to which would be the fairest method of grading a group (of applicants for admission to college, let us suppose): grading them on a test with no preceding practice, or on a test preceded by uniform practice (the same amount for each individual), is of great theoretical interest, but of none at all practically, since groups will not be obtainable in either condition, if the system of testing becomes established and individuals know for months in advance that they are to be tested.

Obviously, if tests susceptible to practice effects are used, the only system which will be at all fair involves the condition that all candidates shall have old forms of the tests in advance, and shall have an opportunity to practice on them; or that in other ways all candidates shall receive an amount of practice which will put them in an equivalent "practiced" condition. The exact amount of practice required for that purpose must be experimentally determined, in order that the necessary minimum may be assured to all the candidates. Since coaching on such intelligence tests is distinctly possible and since the maximal unfairness is obtained where some are coached and others are not, pains must be taken to have all coached effectively.

Whether it is possible to develop types of test which are not susceptible to practice effects, and which yet test the same 'capacities,' and offer equal difficulties, is a question which cannot be answered definitely without prolonged research on that specific problem. The best conjectures we may make on this point are not encouraging. However, if the plan indicated above is followed systematically, we may be able to equalize practice effects and thus render them harmless.

TABLE I
CHANGES IN QUANTILES FROM TEST 1 TO TEST 3

	Fourth Quartile	Third Quartile	Second Quartile	First Quartile
To fourth.....	6	4	1	0
To third.....	3	4	4	0
To second.....	1	3	5	2
To first.....	1	0	1	9

TABLE II

AVERAGE PERCENTAGE OF IMPROVEMENT ON THIRD TEST OF MEN IN QUARTILES OF FIRST TEST

Fourth Quartile	Third Quartile	Second Quartile	First Quartile
54	11	40	18
53	41	33	12
64	44	29	11
46	37	31	25
36	21	41	11
35	3	13	8
33	31	31	10
46	36	16	19
44	3	39	19
62	36	19	13
General Average: 47.4	28.1	29.4	16

TABLE III

AVERAGE PERCENTAGE OF IMPROVEMENT ON THIRD TEST OF MEN IN QUARTILES OF THIRD TEST

Fourth Quartile	Third Quartile	Second Quartile	First Quartile
3	64	49	11
54	31	44	62
3	9	33	25
11	37	36	11
53	44	31	8
35	41	40	10
41	36	46	39
36	46	13	25
33	29	41	19
13	31	31	30
46	19	18	13
General Average: 28	35.2	34.7	23