ing the departments from being wiped out inch by inch than in keeping them going. But what of the other six—a three-fifths majority of the Cabinet?

Mellon handles the finances of the Treasury Department admirably, but when he reaches out to perform the other duties of his office, as in legislation dealing with taxes and debts, he always is beaten to death. Fall probably is below the average of his predecessors in the Interior Department, considering him from the test of fitness for that particular office. No arresting evidence of great power has appeared in the Post Office Department, either under Hays or Work. Wallace is no better and no worse than the Agricultural Department has been accustomed to—a good, sufficient man. Davis is only now showing signs of an elementary understanding of the real business of the Labor Department.

And there is Daugherty.

JOHN W. OWENS.

The Reliability of Intelligence Tests

III.

SUPPOSE, for example, that our aim was to test athletic rather than intellectual ability. We appoint a committee consisting of Walter Camp, Percy Haughton, Tex Rickard and Bernard Darwin, and we tell them to work out tests which will take no longer than an hour and can be given to large numbers of men at once. These tests are to measure the true athletic capacity of all men anywhere for the whole of their athletic careers. The order would be a large one, but it would certainly be no larger than the pretensions of many well known intelligence testers.

Our committee of athletic testers scratch their heads. What shall be the hour's test, they wonder, which will "measure" the athletic "capacity" of Dempsey, Tilden, Sweetser, Siki, Suzanne Lenglen and Babe Ruth, of all sprinters, Marathon runners, broad jumpers, high divers, wrestlers, billiard players, marksmen, cricketers and pogo bouncers? The committee has courage. After much guessing and some experimenting the committee works out a sort of condensed Olympic games which can be held in any empty lot. These games consist of a short sprint, one or two jumps, throwing a ball at a bull's eye, hitting a punching machine, tackling a dummy and a short game of clock golf. They try out these tests on a mixed assortment of champions and duffers and find that on the whole the champions do all the tests better than the duffers. They score the result and compute statistically what is the average score for all the tests. This average score then constitutes normal athletic ability.

Now it is clear that such tests might really give some clue to athletic ability. But the fact that in any large group of people sixty percent made an average score would be no proof that you had actually tested their athletic ability. To prove that, you would have to show that success in the athletic tests correlated closely with success in athletics. The same conclusion applies to the intelligence tests. Their statistical uniformity is one thing; their reliability another. The tests might be a fair guess at intelligence, but the statistical result does not show whether they are nor not. You could get a statistical curve very much like the curve of "intelligence" distribution if instead of giving each child from ten to thirty problems to do you had flipped a coin the same number of times for each child and had credited him with the heads. I do not mean, of course, that the results are as chancy as all that. They are not, as we shall soon see. But I do mean that there is no evidence for the reliability of the tests as tests of intelligence in the claim, made by Terman,* that the distribution of intelligence quotients corresponds closely to "the theoretical normal curve of distribution (the Gaussian curve)." He would in a large enough number of cases get an even more perfect curve if these tests were tests not of intelligence but of the flip of a coin.

Such a statistical check has its uses of course. It tends to show, for example, that in a large group the bias and errors of the tester have been cancelled out. It tends to show that the gross result is reached in the mass by statistically impartial methods, however wrong the judgment about any particular child may be. But the fairness in giving the tests and the reliability of the tests themselves must not be confused. The tests may be quite fair applied in the mass, and yet be poor tests of individual intelligence.

We come then to the question of the reliability of the tests. There are many different systems of intelligence testing and, therefore, it is important to find out how the results agree if the same group

*Stanford Revision Binet-Simon Scale, p. 42.

of people take a number of different tests. The figures given by Yoakum and Yerkes† indicate that people who do well or badly in one are likely to do more or less equally well or badly in the other tests. Thus the army test for English-speaking literates, known as Alpha, correlates with Beta, the test for non-English speakers or illiterates at .80. Alpha with a composite test of Alpha, Beta and Stanford-Binet gives .94. Alpha with Trabue B and C completion-tests combined gives .72. On the other hand, as we noted in the first article of this series, the Stanford-Binet system of calculating "mental ages" is in violent disagreement with the results obtained by the army tests.

Nevertheless, in a rough way the evidence shows that the various tests in the mass are testing the same capacities. Whether these capacities can fairly be called intelligence, however, is not yet proved. The tests are all a good deal alike. They all derive from a common stock, and it is entirely possible that they measure only a certain kind of ability. The type of mind which is very apt in solving Sunday newspaper puzzles, or even in playing chess, may be specially favored by these tests. The fact that the same people always do well with puzzles would in itself be no evidence that the solving of puzzles was a general test of intelligence. We must remember, too, that the emotional setting plays a large rôle in any examination. To some temperaments the atmosphere of the examination room is highly stimulating. Such people "outdo themselves" when they feel they are being tested; other people "cannot do themselves justice" under the same conditions. Now in a large group these differences of temperament may neutralize each other in the statistical result. But they do not neutralize each other in the individual case.

The correlation between the various systems enables us to say only that the tests are not mere chance, and that they do seem to seize upon a certain kind of ability. But whether this ability is a sign of general intelligence or not, we have no means of knowing from such evidence alone. The same conclusion holds true of the fact that when the tests are repeated at intervals on the same group of people they give much the same results. Data of this sort are as yet meager, for intelligence testing has not been practised long enough to give results over long periods of time. Yet the fact that the same child makes much the same score year after year is significant. It permits us to believe that some genuine capacity is being tested. But whether this is the capacity to pass tests or the capacity to deal with life, which we call intelligence, we do not know.

This is the crucial question, and in the nature of things there can as yet be little evidence one way or another. The Stanford-Binet tests were set in order about the year 1914. The oldest children of the group tested at that time were 142 children ranging from fourteen to sixteen years of age. Those children are now between twenty-two and twenty-four. The returns are not in. The main question of whether the children who ranked high in the Stanford-Binet tests will rank high in real life is now unanswerable, and will remain unanswered for a generation. We are thrown back, therefore, for a test of the tests on the success of these children in school. We ask whether the results of the intelligence test correspond with the quality of school work, with school grades and with school progress.

The crude figures at first glance show a poor correspondence. In Terman's studies* the intelligence quotient correlated with school work, as judged by teachers, only .45 and with intelligence as judged by teachers, only .48. But that in itself proves nothing against the reliability of the intelligent tests. For after all the test of school marks, of promotion or the teacher's judgments, is not necessarily more reliable. There is no reason certainly for thinking that the way public school teachers classify children is any final criterion of intelligence. The teachers may be mistaken. In a definite number of cases Terman has shown that they are mistaken, especially when they judge a child's intelligence by his grade in school and not by his age. A retarded child may be doing excellent work, an advanced child poorer work. Terman has shown also that teachers make their largest mistakes in judging children who are above or below the average. The teachers become confused by the fact that the school system is graded according to age.

A fair reading of the evidence will, I think, convince anyone that as a *system of grading* the intelligence tests may prove superior in the end to the system now prevailing in the public schools. The intelligence test, as we noted in an earlier article, is an instrument of classification. When it comes into competition with the method of classifying that prevails in school it exhibits many signs of superiority. If you have to classify children for the convenience of school administration, you are likely to get a more coherent classification with the tests than without them. I should like to emphasize this point especially, because it is im-

†Army Mental Tests, p. 20.

*Stanford Revision of Binet-Simon Scale, Chapter VI.

portant that in denying the larger pretensions and misunderstandings we should not lose sight of the positive value of the tests. We say, then, that none of the evidence thus far considered shows whether they are reliable tests of the capacity to deal intelligently with the problems of real life. But as gauges of the capacity to deal intelligently with the problems of the classroom, the evidence justifies us in thinking that the tests will grade the pupils more accurately than do the traditional school examinations.

If school success were a reliable index of human capacity, we should be able to go a step further and say that the intelligence test is a general measure of human capacity. But of course no such claim can be made for school success, for that would be to say that the purpose of the schools is to measure capacity. It is impossible to admit this. The child's success with school work cannot be a measure of the child's success in life. On the contrary, his success in life must be a significant measure of the school's success in developing the capacities of the child. If a child fails in school and then fails in life, the school cannot sit back and say: you see how accurately I predicted this. Unless we are to admit that education is essentially impotent, we have to throw back the child's failure at the school, and describe it as a failure not by the child but by the school.

For this reason, the fact that the intelligence test may turn out to be an excellent administrative device for grading children in school cannot be accepted as evidence that it is a reliable test of intelligence. We shall see in the succeeding articles that the whole claim of the intelligence testers to have found a reliable measure of human capacity rests on an assumption, imported into the argument, that education is essentially impotent because intelligence is hereditary and unchangeable. This belief is the ultimate foundation of the claim that the tests are not merely an instrument of classification but a true measure of intelligence. It is this belief which has been seized upon eagerly by writers like Stoddard and McDougall. It is a belief which is, I am convinced, wholly unproved, and it is this belief which is obstructing and perverting the practical development of the tests.

WALTER LIPPMANN.

(To be continued.)

[*A number of letters have been received, commenting on the two articles of Mr. Lippmann's series already printed. We have thought it best not to print any of these letters until the completion of the series, when it will be possible to classify and present the points brought up by our correspondents more intelligently.*—THE EDITORS.]

Loyalties

Loyalties, by John Galsworthy. Gaiety Theatre. October 23, 1922.

MR. GALSWORTHY'S play holds many people spellbound by three kinds of logic. There is the logic of melodrama, the logic of the social idea, the logic of the loyalty theme. The melodrama holds the interest by its suspense and plot devices; the sociological meaning sanctifies the curiosity with which one follows the characters and their fortunes; the loyalty theme gives everything a background and assurance of philosophical thinking. No wonder the audience sits there piqued, thrilled and flattered all at the same time. You take first the melodrama of catch the thief and weave that into a skilful plot structure; on top of that you put the Shylock motive and make the play's struggle look more racial than theatric: then you tie all this up with the idea of loyalties, loyalties working so that everyone in the theatre can see them enter and exit, social class loyalties, racial, the policemen, the Italian's daughter, the lawyer, the butler, and finally the wife's loyalty to the husband, with those last touches of his, for her sake, loyally shooting himself and her loyally fainting on the sofa.

If you are Puritan enough still to mistrust what is more or less mere pleasure, you can fall back on the profundities of meaning in the whole play; and after you see what happens about the robbery and catching the thief and what the various people do about it, you can go home and think deeply about the problems of race and class. But Loyalties depends on the story, on the thrill and suspense of tangled incidents. And that is nothing against it. Melodrama is good enough in itself. It can be a sound dramatic pattern. The test of a play may, in certain types at least, be the pantomime of it, as everyone has heard. The important point is that we should keep this clear, and not go fuddling up an entertaining melodramatic framework with abysms of philosophy and subtle creation.

In Loyalties a Jew, who is received on account of his money, comes late at night into his host's room to report that he has lost a thousand pounds. He insists that his host ask everyone in the house to appear. A hero of the war is the man suspected by the Jew. The English gentlemen resent this aspersion by an outsider on one of their set. There is reason to suspect the hero, for one of the gentlemen finds accidentally that his sleeve is wet, which might be proof that he has entered by way of the balcony. But they bully the Jew with lofty conduct and social blackmail. Later the accusations are renewed and taken to court. The stolen notes are traced. The Englishman shoots himself to escape shame.

This is tragic material. There is one scene in Loyalties, the interval of the inspector's examination, into which we must read supernaturally significant comments on police inefficiency to keep it from being very dull and obvious; otherwise the play moves like a clock from start to finish. It is smooth, adroit, quiet and, so far as any unseemly jars in its unfolding go, invariably true to the kindred points of heaven and home. The motives of class prejudice are deftly introduced into the situations. The loyalty theme is superbly completed at the very last by one of the characters remarking that they had kept the faith but it was not enough.

But there is one scene that gives the game away. The moment comes when the climax of the struggle between