VOLUME III JUNE, 1912 NUMBER 6

THE JOURNAL OF
EDUCATIONAL PSYCHOLOGY

METHODS OF RESEARCH IN EDUCATION.[1]

EDMUND C. SANFORD,
President Clark College, Worcester, Mass.

The aspect of educational research, which the Executive Committee was so kind as to offer to me, is that of methods; specifically, that of methods which are promising for immediate application by instructors in departments of education not yet abundantly supplied with material facilities. The topic is a congenial one, and I have taken it up with interest, though I have, I think, no illusions as to the value of the suggestions I shall be able to make, except perhaps as incitements to discussion.

A fact that meets us on the threshold of a practical consideration of methods and which must determine our subsequent course is that in actual research it is the problem that dictates the method, and not the method the problem; or, more exactly, it is the problem, the state of knowledge of the field in which the problem lies, the interest of the investigator, and the facilities at his disposal. If our question, for example, is one of lighting, heating and ventilation, we find ready at hand the precise quantitative methods of physics and chemistry, and where they can be used there is usually no excuse for others. If it is a matter of grade of intelligence and advancement relative to age, we have the Binet-Simon test-scale. If it is the rôle of inheritance, we must gather statistics and calculate the correlations between parent and child. If it is the general character of children's impulses and interests, we are com-

[1]Read at the Wellesley meeting of the New England Association of College Teachers of Education, December 2, 1911.

pelled to rely on the systematic observation of single children, on the questionnaire, and upon collation of biographies. In the case of problems of practical administration we must frequently make use of comparative studies of the experience of others. In still other cases—since pedagogy is a practical science and cannot always wait for the methodical solving of its problems—we are obliged to fall back on crude empiricism, the method which the comparative psychologists have called that of "trial and error," the method of guessing at what ought to be done and trying it on.

If the problem dictates the method, it is with a handful of problems that I ought to begin what I have to say about methods; and as I am to deal with methods readily accessible to college teachers of education, I shall confine myself to college problems, for, while many of us lack special laboratories and experimental schools, we have each of us *ex officio* a college under our hands, and are in some measure responsible for the solution of its problems. I do this with the more confidence also, because I believe that these problems are for the most part typical, both in nature and in the methods of investigation by which they must be approached, of educational problems all along the line.

Here, then, is a group of problems, already somewhat worked upon, which can be treated by purely statistical methods on the basis of data to be found in the college registrar's office or readily to be obtained from other sources.

1. What are the causes of failure in college work? Studied by means of the office records and information as to the college life and activities of the men who are dropped from college for poor standing. This question has been studied by President Schurman, Professor Miner and others.
2. The correlation between high standing in college and later success in the professional schools, and the correlation between success in certain special studies, as in history or in science, and later high standing in the law or medical school.

Both of these have been studied by President Lowell.

3. The correlation between high standing in mathematics or the classics and success in other directions. This has been studied by many different champions of these studies as means of mental discipline.
4. The extent to which a student's academic fate is sealed before his entrance to college, as shown by the correlation between scholastic records in the high school and the college. Studied by Professor Dearborn.
5. The inverse correlation between the habit of smoking and success in college work. Investigated some time ago by one of the Clark students.

Such problems as these are the easiest upon which to begin educational research, because the questions can be stated clearly and the material has already been gathered. The following are of the same sort, and, so far as I know, have not yet been studied to the point of publication:

6. Is there any relation between age of entrance to college and success in college work?
7. How far, if at all, can maturity and other similar advantages compensate for imperfect preparation for admission?
8. What, in general, is the effect on the work of a student in the graduate or professional schools of having taken the college course in three years? (President Lowell has gathered data on this point.)
9. A study, similar to that on college laggards, on men who do exceptionally well in college, and the causes of their success—their age, home conditions, participation in college activities, social standing among their mates, physical vitality, moral and religious character, etc.
10. Relative standing of men who concentrate on a few subjects and those who scatter over a good many.

Statistical Methods. In the use of all statistical methods there are, of course, certain essential precautions to be taken in gathering the original data and in its subsequent manipulation. And there is always necessary a certain circumspection in the interpretation of the final results. The most essential of these points of care are: First, to be sure that the data represent facts; then not to draw inferences from too few cases; not to mix figures which do not really stand for facts of the same kind; and never to forget that numerical results may frequently be explained with equal ease in several different ways. The first of these points is too obvious to need illustration. The second was sometimes illustrated by President Wright by a case in which, in a study of the ratio of criminality to population in a certain city, the Turks appeared to be the most criminal, having a record of 300 per cent., which, however, was in reality due to the fact that there was but one Turk in the city, and he had been arrested three times. An instance of the third, coming under my own notice, was that of lumping the statistics of conversion of young men and young women, which gave the age of most frequent conversion at a point above what it was in reality for the one group and below what it was for the other. In illustration of the last, I may say that it appears that students in small classes in college get higher marks on the average than those in large classes, but it will not do to conclude that this is due to the better pedagogical conditions in small classes, because smaller classes are generally in advanced subjects and the poorer students may have dropped by the way; or, since advanced subjects are usually elective, the small classes may be attended by those only who have a spontaneous interest in the subject. For the same reason the common inference, from the fact that high scholarship in all subjects is more frequent among those that have taken Latin, Greek and mathematics in the preparatory school and the lower college classes, that these subjects furnish a general training of a superior kind is not justified, for the weaker students may steer clear of them or fail if they take them, *i. e.*, these studies may act as a sieve instead of a gymnastic; or they may be taken chiefly by students from homes where the cultural standards are higher, *i. e.*, students from more successful

families, students endowed by inheritance with physical and mental powers above the average. For the detailed precautions to be observed in statistical matters, one would go, of course, to works on statistics; and naturally, in educational research, to Thorndike's "Mental and Social Measurements," or his "Educational Psychology."

I may mention, however, one other matter of importance in this connection, and that is the dependence to be placed upon the marks for college standing on which we rely for our evidence of success. There is, unfortunately, a widespread suspicion as to the value of college marks, not only on the part of students who succeed usually in getting a "gentleman's grade," but on the part of those who have to do with such marks for other purposes. This is to be regretted, for it is by no means unavoidable.

As now managed, college marks in different subjects probably do give some index of the relative success of students in those subjects as judged by their several instructors. A paper graded 95 by one instructor is probably distinguishably better than one graded 85 by the same instructor, but it is by no means certain that it is better than a paper graded 75 by another instructor whose subjective grade-scale is different. That such differences are common has been shown by Cattell, Meyer, Dearborn and others, and can easily be verified by anyone who cares to take the trouble to compare the marks assigned by his colleagues. This makes combination of the marks unsafe and their comparison often misleading.[2]

The variation comes largely from the fact that college instructors have usually, when they begin, no special training in assigning marks, and no effort is made to urge or assist them toward uniformity of standards. Improvement naturally results through social pressure when each instructor knows how his own marking habits compare with those of his colleagues and what in the long run the proportionate frequency of the marks of various grades ought theoretically to be, which can be shown

[2] Difference of grades from instructor to instructor in any single year is, however, not by any means a sure sign of difference in subjective standards, even when the instructors deal with parallel sections in the same subject. It may be due to an unfortunate chance in the grouping of students, as was brought out by Dr. Norton of Harvard in the discussion following this paper.

ly the publication of the proportion of marks of each grade assigned by each instructor at the last previous examination period, or his average for the past year or two, together with the theoretical proportions, and, if necessary, with a simple explanation of what the figures mean. It is even possible, as Dean Ferry of Williams has done, to calculate, from a sufficiently large collection of marks by different instructors, a numerical coefficient by which the marks of the aberrant can be reduced to the proper general standard.

Besides this general lack of uniformity in the meaning of the grades assigned, there is also another habit of instructors which is useful perhaps pedagogically, but injurious to the marks as data for pedagogical research. I mean the habit of using the marks as rewards and punishments, giving the student who deserves encouragement a higher grade than the bare quality of his work would justify, and shading downward the mark of a student who has not worked so faithfully as his instructor believes that he ought. Of the same sort is the custom of certain instructors to cut down the scholarship mark as a penalty for absence. These customs introduce into the records factors which tend to irregularity, and which it is impossible to remove. It would be much better if the students could be graded separately for each particular aspect of their work—scholarship, diligence, attendance and other matters if necessary—and the marks combined after filing in the registrar's office according to an established system. We should then have data on scholarship which would be purer than under the present system.[3]

After making sure of the reliability and adequacy of the original data, the next precaution is, as I have said, to avoid misleading mixtures of the data having reference to different classes of cases. In studying the correlation of the age at entrance to college with success in college work, it would, of course, be necessary to discriminate those who came late to col-

[3]It would perhaps be possible to test the efficiency of college marks as general indexes of ability by testing their correlation with the grading of the students in (1) the skilful use of English, (2) their skill in discovering and exposing the error in fallacious arguments, (3) with their grades when graded by their fellow-students after the manner used by Cattell in his study of American Men of Science.

lege through ill-health or through the necessity of earning money to pay their way from those who came late from sheer stupidity; and in interpreting the result it would be necessary to consider the various distractions and other causes of poor success in college work and to determine whether these were operative equally at all ages, before one could arrive at a statement with reference to age alone. Similarly, if it should appear that three-year men were less (or more) successful than four-year men in their later professional studies, it would be necessary, before concluding that the three-year course was in general a bad (or a good) thing, to consider the type of men who now take it. The value of statistical results lies in their just and skilful interpretation, and their interpretation is often an especially serious matter in the complicated cases with which educational researches have to deal.

The Questionnaire. In the case of many other problems of interest to educators the data are not already in existence, but must be secured by some sort of a census-taking or questioning. If, for example, one should undertake a study of the heredity or of the present home conditions of the bright or dull men in college, or, for light on vocational training in college, should try to find what proportion of students enter with a definite vocational bent, how many change their life-plan during college, and how many follow, after graduation, the plan which they formed in college, it would evidently be necessary to depend chiefly on data furnished by the students themselves in interviews or in response to a schedule of questions.

The method of the questionnaire is one that can be applied with all degrees of rigidity and laxness, and can furnish accordingly data of all degrees of value from those which are susceptible of rigid mathematical treatment and yield results of practical certainty to those which cannot fairly be made to yield any quantitative statement at all, and are perhaps overpraised if regarded as worthily suggestive. The ideal questionnaire is one which asks for information which the questionee is sure to possess and willing to give, and which he can state briefly and unequivocally—which asks questions, for example, that can be answered by "yes" or "no," or by definite statements of facts or figures.

When it goes beyond questions of this sort and asks for matters of personal history or of opinion and belief, it soon reaches a point where a quantitative treatment of the data is impossible, or, if possible, is often misleading. The best that can then be done is to gather the material into characteristic groups or types and to attempt little beyond a liberally illustrated description of these.

A second important point is to see that the questions are answered by all of the group which you have under consideration, or, if it is too large for that, by a perfectly fair and adequate sample of the whole group. A set of questions with reference to conversion answered by a class of clergymen would, of course, be hardly a fair basis for inference as to the religious experiences of the man in the street.

A good many years ago the English Society for Psychical Research gathered many replies to a questionary, of which the chief question was as follows: "Have you ever, when believing yourself to be completely awake, had a vivid impression of seeing or being touched by a living being or inanimate object, or of hearing a voice; which impression, so far as you could discover, was not due to any external physical cause?" In case of an affirmative answer, a description of the experience was asked, with particulars as to the time and circumstance. The question was carefully worded, and seems tolerably plain; the answers ran into the thousands, and the returns were conscientiously and even critically worked over, and yet they left the real facts of the matter in many respects uncertain.[4] The figures seemed to show that women are more subject to such experiences than men, but it is not sure that the experiences are not as frequent with men, for they very likely forget them more frequently in their busier and more varied lives, and so report them less often. Even the frequency of the experience among both men and women is in doubt. As the number of returns rose, the proportion of affirmative answers fell off, showing probably that the collectors had quite unconsciously not taken people at random, but asked those first whom they judged would be interested to answer the questions, and

[4] See criticism by Parish, "Hallucinations and Illusions," New York, 1902, pp. 83 ff.

this often meant people who had had such experiences. Later, when these interested people were exhausted, the proportion of those asked who had not had such experiences increased. Even the great preponderance of visual hallucinations over those of the other senses, which the figures seem to show, has been called in question on the ground that those of vision are better remembered than those of the other senses, and evidence to support the contention has been drawn from other portions of the data themselves.

If complex matters, where misunderstandings are easy and where forgetfulness or inattention is liable to make the replies unreliable, are to be studied, it is well that the questioner make his inquisition in person and explain and cross-question as occasion demands. A few cases thus thoroughly worked out at first hand are, of course, of more value than a much greater number collected carelessly and embodying only such things as the questionee chanced at the moment to think of and set down.

Two small questions which might possibly repay investigation in this way I may mention in passing:

1. The moral code of students with reference to questions involving the college or behavior toward other students and its sanctions. This has been touched upon experimentally by Sharp in a paper some years ago in the *American Journal of Psychology*.

2. Students' methods and habits of study—not so much as to time spent as to the manner in which the work is done.

Experimental Methods. Much more difficult in execution, but also more certain and definite in results, are the methods of direct experimentation. Of these there are, roughly, three types:

1. The painstaking introspective study of various forms of the learning process under laboratory conditions. This is hardly to be distinguished from experimental psychology, and most of it has so far been done in the psychological laboratories. I refer to such work as that of Meumann and other German investigators on economic methods of memorizing, and to such work as that of Bryan and Harter, Swift, Book

and others in this country on the acquirement of skill in teleggraphy, ball-tossing, shorthand, typewriting, and the like. Such experiments as these require the resources and technique of the laboratory, and are not likely to be undertaken by teachers of education unless they are both fortunately situated and have strong psychological leanings.

2. Then there are the studies of individual differences and the testing of capacities of various kinds, which promise in the end to give us a science of psychography—in other words, a schedule and a procedure by which we can take the psychic measure of a man in all important particulars and inventory all his characteristic qualities. This work, in anything like completeness, is yet in its beginning, and is connected in its recent developments especially with the name of Stern, in Germany. In a simpler way it has given us the Binet-Simon test-scale for determining the psychological age of backward children. To this type also would belong the tests and measurements worked out so fully in Whipple's "Manual of Mental and Physical Tests." And closely connected with this type of experimentation, though important enough to deserve the name of a method by itself, is Thorndike's work in the establishment of standard scales of excellence for the quantitative measurement of excellence in penmanship and elementary English, and that of Courtis and others for work in elementary arithmetic. Many of the experiments of this type also call for equipment and technique, though the practical value of the studies of the individual pupil and the simplicity of the test material or its use of schoolroom products bring it nearer to the daily business of the student of education.

3. The experiments of the third type are strictly pedagogical—experiments made under school conditions, with the ordinary materials of the classroom, in order to answer strictly pedagogical questions. Such, for example, is the work of Lay in Germany on the first steps in number, of Winch in England on the possible transfer of the educative effect of work on the fundamental operations of arithmetic to improvement of mathematical reasoning, of Gilbert in our own country on the teaching of zoology in such a way as to keep in view its practical bearings as compared with the teaching of the same sub-

ject as pure science, and of Pearson on the teaching of spelling —the last three reported in the JOURNAL OF EDUCATIONAL PSYCHOLOGY.

In this sort of experimentation pedagogy comes to her own. The essence of the method is to submit to educative processes which are different in one clearly defined particular two groups of pupils which are, as regards native ability, previous schooling and all other essential items, as much alike as they can be made, and to compare results, stated in some definitely quantitative way. If, for example, a teacher of the mother tongue wished to determine whether facility in composition is to be more readily acquired by much practice in writing or by memorizing extracts from the best models (as a French experimenter has seemed to find it), he would proceed as follows: He would first divide his class into two sections, having regard to marks previously obtained in composition, in such a way that the average ability was as nearly as possible the same in both sections. He would then for a semester, say, demand a daily theme of one section and of the other an equal time spent in committing to memory and reciting aloud an assignment of good prose, while care was taken to secure as great uniformity as possible in all other work done by the two sections. At the end of the semester test themes would be required of both sections—or perhaps several themes—and the productions would be marked and compared, and the value of the methods judged according to the results. It would be desirable, probably, that marks be assigned for more than one quality; for example, for correctness as well as for facility of diction, and that the papers be graded by someone ignorant of the purposes of the experiment and of the membership of the sections. For greater certainty the treatment of the sections should be reversed during the second semester and the results again compared.

In all such experiments the attitude of the instructor is, of course, a very important factor, but, if he is worthy to be an investigator at all, he will know enough not to spoil his own experiment by carelessness in the management of it or by bias in estimating its results.

The comparison of the products in such an experiment would

be possible on the basis of gradings according to the usual school methods, but it would be improved in certainty and precision if it could be carried out by several readers, and especially if it were checked by the use of a standard scale of attainment in English composition against which each individual's production could be measured off and assigned a definite value.

The construction of such scales is worth a little consideration, for they promise to bring within range of quantitative measurement many phenomena the proposal to measure which has until recently seemed visionary, if not absolutely chimerical. To Dr. Thorndike belongs the credit of having pushed forward as a pioneer in their construction. See his papers in the *Teachers College Record,* XI, No. 2, March, 1910, and the JOURNAL OF EDUCATIONAL PSYCHOLOGY, II, No. 7, September, 1911, 361-368.

The production of a reliable scale is no easy matter, and requires the co-operation of a number of conscientious judges, but it is worth the labor. In substance the method is this: Given a large number of samples of the product in question, say 1000 specimens of handwriting, of all grades of quality from the very poorest to the very best. The whole set is classified individually by a number of competent judges (perhaps 30 or 40) into 10 groups as nearly equally different in quality as each judge can make them. The record of the classes in which each specimen is placed by each judge is kept, and by calculation gives an average rating for each specimen depending on the judgments of all the judges—a rating which it is fair to assume is very much more reliable, that is, is very much more nearly right than the rating of any single judge.

From the whole 1000 specimens thus rated it will be easy to find one or more examples which correspond almost exactly with the ideal grades 1, 2, 3, 4, etc., which measure off by equal distances the range of difference from the worst to the best specimens. These taken by themselves can be preserved as permanent samples, and together form the standard scale, which can be photographed or otherwise reproduced, and can be had by anyone who needs to grade handwriting. Any new piece of handwriting to be graded can be compared with these

and be assigned its grade as it falls above or below or exactly coincides with one of the samples. Such a scale does not, of course, enable us to perceive differences which we cannot perceive without it. It is not a microscope, but a footrule. It should enable us, however, to give our impressions of difference an objective and quantitative statement which can be apprehended by any other worker who has a standard series like ours.

The invention of such scales is, in my opinion, full of promise for the scientific study of educational problems. I believe with Thorndike that "if a number of facts are known to vary in the amount of anything which can be thought of, they can be measured in respect to it," and I am hopeful that the time may yet come when it will be possible to meet critics with figures and to demonstrate in a quantitative fashion that one sort of college training is better than another.