






LIBRARY
OF THE
UNIVERSITY
OF ILLINOIS

370

IPG

no. 43-53



BULLETIN No. 48

BUREAU OF EDUCATIONAL RESEARCH
COLLEGE OF EDUCATION

EXPERIMENTAL RESEARCH
IN EDUCATION

By

WALTER S. MONROE

Director, Bureau of Educational Research

and

MAX D. ENGELHART

Assistant, Bureau of Educational Research

PUBLISHED BY THE UNIVERSITY OF ILLINOIS
URBANA

PREFACE

There is urgent need for a comprehensive description of the techniques employed in educational research. There are a large number of texts dealing with statistical methods, especially the more elementary ones, but statistical procedures represent only one group of the techniques of educational research. Among the techniques for which we have no adequate treatment, the need is probably most urgent for those relating to setting up and conducting experiments. Experimental research is a means for evaluating educational procedures and, hence, occupies a position of importance. In general outline, the procedure is simple, but an analysis reveals its complexity. The idea of "controlled experimentation" is easy to comprehend, but it is not easy to specify precisely what is involved in maintaining a control group.


In this bulletin an attempt is made to describe in some detail the procedure of controlled experimentation, and on the basis of the requirements revealed, a small group of experiments is evaluated. The analysis of the factors affecting pupil achievement and the evaluation of the factors considered are largely subjective. An attempt was made to utilize the best data obtainable, but the supply is inadequate and in some cases the information is not highly dependable. Consequently, both the analysis and the evaluation must be considered tentative and subject to revision in the light of future investigations. The writers, however, believe that they have succeeded in showing controlled experimentation to be a highly complex and an intricate type of research, rather than one which can be carried out successfully by any novice who is sufficiently interested.

The bulletin should be of interest to teachers, supervisors, and administrators, as well as to research workers. The latter will find it helpful as a guide in planning and conducting an experiment and in interpreting the results. To the others, it should give a set of criteria that may be used in evaluating the experimental investigations reported in our educational literature.

The writers are glad to take this opportunity to express their indebtedness to Dr. C. W. Odell for a careful reading of Chapter III and to Mr. T. T. Hamilton, Jr., for the editing of the entire manuscript.

January, 1930.

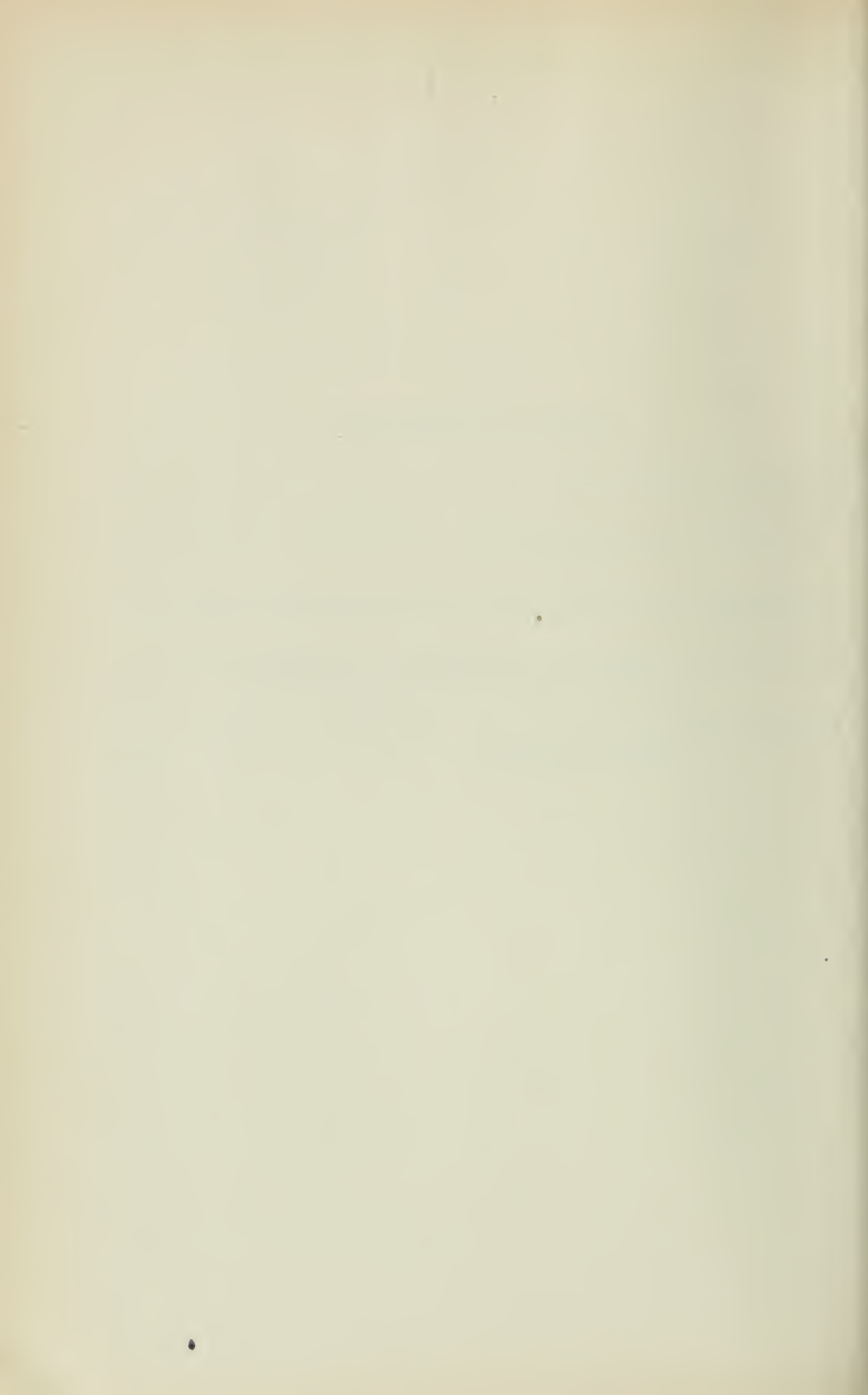
WALTER S. MONROE
MAX D. ENGELHART



Digitized by the Internet Archive
in 2012 with funding from
University of Illinois Urbana-Champaign

TABLE OF CONTENTS

CHAPTER I. INTRODUCTION.	7
CHAPTER II. THE REQUIREMENTS FOR CONTROLLED GROUP EXPERIMENTATION	18
CHAPTER III. THE INTERPRETATION OF DIFFERENCES IN GAINS	59
CHAPTER IV. A CRITICAL EVALUATION OF EXPERIMENTAL STUDIES RELATING TO SUPERVISED STUDY .	77
CHAPTER V. EXPERIMENTATION AS A PROCEDURE IN EDUCA- TIONAL RESEARCH	99



EXPERIMENTAL RESEARCH IN EDUCATION

CHAPTER I

INTRODUCTION

The passing of speculation and authority. Until recently the typical method of answering questions relative to education has been that of speculation, and the pronouncements of those recognized as authorities have been accepted generally as final; but history records a number of attempts to solve thought questions in education by means of trial and observation of results. For example, Vittorino da Feltré (1378-1446) followed this procedure in devising methods of teaching that attracted much attention to his school, the *Casa Giocosa* at Mantua.¹ Wolfgang von Ratke, or Ratich, (1571-1635) also attempted to prove the value of his method by actual trial in practice.² The theories of Comenius and Rousseau found expression in practice through the founding of the *Philanthropinum* at Dessau by Johann Bernhard Basedow (1723-1790).³ Johann Heinrich Pestalozzi (1746-1827) put his educational theories into practice in his schools at Stanz, Burgdorf, and Yverdun.⁴ Johann Friederich Herbart (1776-1841) was a firm believer in the value of experimental procedure and inaugurated a practice school along with his pedagogical seminar at the University of Königsberg.⁵

The evaluation of pedagogical theory by trial in practice was the aim of several pioneer experimental schools in the United States. Among the most notable of these were the Oswego Primary Teachers

¹Woodward, W. H. *Vittorino da Feltré and Other Humanist Educators*. Cambridge, England: Cambridge University Press, 1905. 261 p.

²Raumer, Karl von. *Geschichte der Pädagogik*. Gütersloh: Druck und Verlag von C. Bertelsmann, 1902, p. 27-29.

A briefer description of this "experiment" is given in:

Graves, F. P. *Great Educators of Three Centuries*. New York: The Macmillan Company, 1912, p. 20-26.

³Raumer, *op. cit.*, p. 212-52.

Brief descriptions are given in: Graves, *op. cit.*, p. 112-21.

Monroe, Paul. *A Textbook in the History of Education*. New York: The Macmillan Company, 1929, p. 580-83.

⁴An account of his visit to Pestalozzi's institution at Yverdun is to be found in:

Raumer, *op. cit.*, p. 340-59.

Other descriptions of Pestalozzi's work are to be found in:

Barnard, Henry. *Pestalozzi and his Educational System*. Syracuse, New York: C. W. Bardeen Company, 1906. 751 p.

Graves, *op. cit.*, p. 122-66.

Monroe, *op. cit.*, p. 601-22.

Parker, S. C. *A Textbook in the History of Modern Elementary Education*. Boston: Ginn and Company, 1912, p. 273-74.

⁵For discussions of the work of Herbart see:

Compayre, Gabriel. *Herbart and Education by Instruction*. New York: Thomas Y. Crowell and Company, 1907. 142 p.

De Garmo, Charles. *Herbart and the Herbartians*. New York: Charles Scribner's Sons, 1896. 268 p.

Graves, *op. cit.*, p. 167-93.

Monroe, *op. cit.*, p. 622-39.

Parker, *op. cit.*, p. 375-430.

Training School, with its model school for observation, established by Edward A. Sheldon in 1861;⁶ the experimental school inaugurated by Francis W. Parker when he assumed the principalship of the Cook County Normal School in 1883;⁷ and the Laboratory School at the University of Chicago, established by John Dewey in 1896.⁸

Early experimentation handicapped by inadequate conception of control of educative factors and by lack of instruments for measuring pupil material and pupil achievement. The pioneer experimentation in education failed to yield dependable results because of an inadequate conception of control of educative factors. A single group of pupils was subjected to a complex of educative influences, including the novel procedure that was being tried, and after the close of the experiment, the results were ascribed, in many cases erroneously, to the novel procedure alone. A repetition with another group of pupils secured contrary results. This is well illustrated by the success of enthusiastic reformers who, in their own schools, showed an apparent superiority of their methods. Repetition by less enthusiastic schoolmen often failed to substantiate the contentions of the reformers.

A second handicap in these early experiments was the lack of instruments for measuring pupil material and pupil achievement. Measurement is fundamental to experimentation. The investigator must measure the original status of the pupils participating in the experiment, submit them to the experimental procedure, and measure them again. The pioneer experimenters were handicapped by their inability to secure quantitative measurements of the initial status of their pupils and of their final status after they had been subjected to the experimental procedure.

The development of the concept of control of experimental conditions. The investigations of Rice, which were made between 1894 and 1897, were transitional in the techniques used. The results obtained with one group of pupils were compared with the results secured from other groups of pupils. Comparison of results obtained by one procedure with results obtained by other procedures is a means of

⁶For a description of this school see:

Autobiography of Edward Austin Sheldon. New York: Ives-Butler Company, 1911, p. 133-80.
Dearborn, N. H. "The Oswego Movement in American Education," *Teachers College, Columbia University Contributions to Education*, No. 183. New York: Bureau of Publications, Teachers College, Columbia University, 1925. 191 p.

⁷Rugg states, ". . . he built up in the Cook County Normal School a faculty of experimentalists, of fearless innovators, real students of childhood, and a practice school which proved an influential object lesson for both teachers and the general public." See:

Rugg, H. O. "Curriculum-Making in Laboratory Schools," *Twenty-Sixth Yearbook of the National Society for the Study of Education*, Part I. Bloomington, Illinois: Public School Publishing Company, 1926, p. 87-91.

⁸Dewey, John. *The School and Society.* Chicago: University of Chicago Press, 1900. 129 p.

A revised edition of 164 pages was published by the University of Chicago Press in 1915.

securing a measure of control of experimental conditions whose importance was recognized by Rice in the following statement:

By a comparative study of results, even on a much narrower basis than I have indicated, a great deal might be accomplished in a very brief period toward the solution of the problem of methods.⁹

The influence of Rice is evident in the report of an experimental investigation of spelling by Cornman. This research had as its object the determination of the relative effectiveness of formal instruction and incidental teaching in spelling. The results obtained in the two experimental schools were compared with those obtained in schools retaining the formal instruction.¹⁰

Prior to 1910 the use of control groups was most prevalent in learning experiments conducted by psychologists under laboratory conditions, but several notable experiments were carried out with the use of control groups under school conditions. Three may be mentioned from the field of transfer of training.

BAGLEY, W. C. and SQUIRE, C. R. "Experiment on Transfer of Ideals of Neatness," performed in 1905 and reported in Bagley, W. C. *Educational Values*. New York: The Macmillan Company, 1911, p. 188-89.

RUEDIGER, W. C. "The Indirect Improvement of Mental Functions Thru Ideals," *Educational Review*, 36:364-71, November, 1908.

WINCH, W. H. "The Transference of Improvement in Memory in School Children," *British Journal of Psychology*, 2:284-93, January, 1908; 3:386-405, December, 1910.

The extent to which the use of control groups has been recognized by experimenters in education is indicated by the fact that control groups were employed in thirty-five out of seventy-two experimental investigations reported in the *Journal of Educational Research* from January, 1920 to June, 1927,¹¹ and in seventeen out of twenty-six experiments reported as *Teachers College Contributions to Education* from 1918 to 1926.¹² It is evident that this technique is almost universally recognized as essential, even though a large proportion of contemporary experimenters fail to employ it.

The development of instruments for measuring pupil material. The use of control groups, as an experimental technique, rests on the assumption that equivalent groups can be secured. In order that equivalence may be secured, it is essential to measure the pupils with respect to characteristics which influence learning in the experiment.

⁹Rice, J. M. *Scientific Management in Education*. New York: Hinds, Noble and Eldredge, 1914, p. 51.

The chapter from which this quotation was taken was first published in *The Forum* for January, 1897.

¹⁰Cornman, O. P. *Spelling in the Elementary School: An Experimental and Statistical Investigation*. Boston: Ginn and Company, 1902, p. 59.

¹¹Monroe, W. S., et al. "Ten Years of Educational Research, 1918-1927," *University of Illinois Bulletin*, Vol. 25, No. 51. Bureau of Educational Research Bulletin No. 42. Urbana: University of Illinois, 1928, p. 79-80.

¹²*Ibid.*, p. 82.

Educational experimentation has acquired one of its most important tools in the development of tests to measure the chief of these characteristics—intelligence. The following paragraph briefly traces their development.

The work of Galton (1869-) and Cattell (1890-) and other American psychologists on the differences in mental abilities of individuals has been said to mark the beginning of modern intelligence testing.¹³ In 1905, Binet, in collaboration with Simon, published the first individual intelligence scale.¹⁴ Intelligence testing became fairly common when Terman's revision of the Binet-Simon Scale became generally available in 1916. In 1918 appeared the first group intelligence scale designed for school use, that of Otis,¹⁵ and since 1918, group intelligence tests have been widely used in elementary and secondary schools, and to some extent in colleges and universities. It is estimated that seven to ten million are used annually at present.¹⁶ In 43 per cent of the learning experiments reported in the *Journal of Educational Research* from January, 1920 to December, 1928, intelligence tests were used to measure pupil material for the purpose of securing equivalent groups.

The development of instruments for measuring pupil achievement.

For securing control groups that are equivalent to experimental groups in such an important characteristic as previous school achievement and for measuring the experimental achievement, valid and reliable instruments are essential. In 1908, Stone, under the direction of Thorndike, devised the first standardized achievement test.¹⁷ This was followed in the next few years by Curtis' Arithmetic Tests, Series A (1909), Thorndike's Handwriting Scale (1909), Hillegas' Composition Scale (1912), Buckingham's Spelling Scale (1913), and Ayres' Handwriting and Spelling Scales (1912-15).¹⁸ In more recent years there have been developed a multitude of achievement tests in almost all of the school subjects, both elementary and secondary, and to some extent in subjects of higher education. Some progress is being made at present in the development of measurements of character and personality. It has been estimated that thirty to forty million standardized tests and scales are used annually, of which, three-fourths are tests of achievement.¹⁹ In 58.3 per cent of the learning experiments

¹³Monroe, *et al.*, *op. cit.*, p. 89.

¹⁴*Ibid.*, p. 90.

¹⁵*Ibid.*, p. 98.

¹⁶*Ibid.*, p. 114.

¹⁷*Ibid.*, p. 90.

¹⁸*Ibid.*, p. 91.

¹⁹*Ibid.*, p. 114.

reported in the *Journal of Educational Research* from January, 1920 to December, 1928, standardized achievement tests were used to measure pupil achievement for the purpose of evaluating the effect of the experimental procedure.²⁰

Development of statistical techniques used in securing equivalent groups and in interpreting differences in gains in achievement. The theory of correlation, discovered by Galton about 1875²¹ and extended by Pearson, Yule, Spearman, and others,²² has enabled experimenters to evaluate the validity and reliability of intelligence tests used to secure equivalence and of educational tests used to measure gains in achievement. Gauss, Encke, Quetelet, Galton, Pearson, Sheppard, Yule, Spearman, Filon, and Kelley should be mentioned for their work in the development of the statistics of errors. The error of a difference formula, particularly useful in the interpretation of differences in gains in experimentation, has evolved as a result of the work of Encke, Airy, Sheppard, and Yule.²³ The suggestion of the "experimental coefficient" by McCall in 1923 has provided experimental workers in education with a criterion for testing the statistical significance of a difference.²⁴ It would be possible to mention many other statistical devices that have been developed in recent years and that are of service in educational experimentation.

Development of educational tests accompanied by interest in experimentation under school conditions. The development of educational tests was accompanied by increasing interest in experimentation under school conditions. Leaders in the field of education stimulated this interest by speeches at educational meetings and by editorials in educational journals. The following quotation from an editorial in the first number of the *Journal of Educational Psychology* is characteristic of these utterances.

Educational practice is still very largely based on opinion and hypothesis, and thus will it continue until competent workers in large numbers are enlisted in the application of the experimental method to educational problems. Little more than a beginning has been made in this important movement.²⁵

²⁰It is, of course, not essential that an achievement test be standardized for it to be suitable for use in an experiment. Standardized tests are usually better constructed than tests made informally, and as such, are better measures of achievement. See:

Odell, C. W. *Traditional Examinations and New-Type Tests*. New York: The Century Company, 1928, p. 21.

²¹Adrain, Laplace, Plana, Gauss, and Bravais had developed some ideas of correlation before Galton, but the first clear statement of the theory and the first use of the term "correlation" in 1888 must be credited to him. See:

Walker, H. M. *Studies in the History of Statistical Method*. Baltimore: The Williams and Wilkins Company, 1929, p. 92-106.

²²*Ibid.*, p. 107-41.

²³*Ibid.*, p. 114-15.

²⁴*Ibid.*, p. 180.

The idea of such a ratio was first developed by De Moivre, Kramp, and McGaughy.

²⁵*Journal of Educational Psychology*, 1:2, January, 1910. (An editorial.)

In an article published in this same issue of the journal, which has since published more learning experiments than any other periodical, Thorndike states:

Schoolroom life itself is a vast laboratory in which are made thousands of experiments of the utmost interest to "pure" psychology. . . . Experts in education studying the responses to school situations for the sake of practical control will advance knowledge not only of the mind as a learner under school conditions but also of the mind for every point of view.²⁶

Dearborn urged the repetition of laboratory experiments under school conditions. His emphasis on the use of appropriate techniques and his plea for careful work, coming as it did in 1911 long before "mass production" in educational research had been reached, should not fail to be noted. The following quotation illustrates the character of Dearborn's pleading:

If this is to be a serious school experiment, practice should be carried out for months at a time, and longer. The entire subject may be dropped for a year or more from the work of one class or group and carried on with regular and persistent practice in a comparable group. Such an arrangement of the work in the early years of the elementary school in view of the importance of the experiment and in view of the possible flexibility of the elementary-school course would not be an undue interference with the work of the school.²⁷

In 1914, Whipple urged that learning experiments under laboratory conditions and using adults as subjects be repeated under school conditions with children as the subjects. Such a statement as the following could not help but stimulate research workers in the field of education to undertake investigations of the type advocated.

. . . . I believe that in one important phase of experimental work—that dealing with the effects of practice and its spread or transfer—experimentation with children has been somewhat neglected, and that most of the conclusions now current upon the nature of formal discipline have been based upon observations carried on with adults. . . . The whole problem of practice might be recanvassed to advantage with children working under classroom conditions.²⁸

A sublime faith in the value of experimentation in the solution of educational problems is expressed in the following quotations:

Now comes the experimentalist, and with clear, unfaltering eye and steady, relentless tone, he demands of each subject the justification for its existence.²⁹

Everywhere there are evidences of an increasing tendency to evaluate educational procedures experimentally. . . . Scientific organizations, research committees, an institute of educational research, and large educational foundations are lending such impetus as make experimental education the most important current movement in education.³⁰

²⁶Thorndike, E. L. "The Contribution of Psychology to Education," *Journal of Educational Psychology*, 1:12, January, 1910.

²⁷Dearborn, W. F. "Experimental Education," *School Review Monograph* No. 1. Chicago: University of Chicago Press, 1911, p. 10.

²⁸Whipple, G. M. "Applicability to Children Secured with Adults," *Journal of Educational Psychology*, 5:362, June, 1914. (An editorial.)

²⁹Bell, J. C. "A New Humanism Needed," *Journal of Educational Psychology*, 9:165, March, 1918. (An editorial.)

³⁰McCall, W. A. *How to Experiment in Education*. New York: The Macmillan Company, 1923, p. 2.

It is to the experimental method that education must look for the solution of many of its most vexing problems. It is upon this basis that the ultimate establishment of education as a science must rest.³¹

Thus within a relatively short period, controlled experimentation reached the highest vogue in the repertoire of research workers in education. Through the stimulation of leaders in this field, the multitude engaged in educational experimentation. The following paragraphs portray the awakening of a few to the limitations of present experimental method even with perfected techniques.

Recent criticism of the experimental method in education. In recent years the feeling has arisen on the part of some leaders in the field of education that educational experimentation, as it is carried on at present, is largely futile.

The need for a program of research in teaching becomes more apparent when the nature of the so-called "scientific investigations" in that field is considered. In general, many of the investigations are too limited in duration, involve too few subjects, and are too crudely done to warrant satisfactory conclusions. The topics investigated are unrelated, and many of those attempting research have not been properly trained for such work.³²

A survey of the learning experiments reported in the *Journal of Educational Research*, *Journal of Educational Psychology*, and the *Teachers College, Columbia University Contributions to Education* during the period 1918-27 provided the data on which the following conclusion is based.

Although no systematic survey has been made, it appears that the permanent accomplishments of educational research during this period are much less than the quantity of production would lead one to expect. This is especially true of experimental studies.³³

Henmon, after three years work with the Modern Foreign Language Study in the production of tests and in the setting up of controlled experiments, reflects as follows on the possibilities and difficulties of experimentation:

We teach our students to be scornful of tradition and mere observation and insist that all things must be subjected to the test of controlled experimentation.

This is undoubtedly a healthy attitude to take if education is to become a science but the constant reader of the present day educational literature cannot in his critical moments help but be troubled by the imperfections and ambiguities of our measurements and the inconclusiveness of our sporadic experiments. When, for example, on such an important problem for educational theory and practice as the effect of equal practice on individual differences, whether equal practice increases or decreases them, we find out of twenty-four experimental studies twelve of them leading at least tentatively to the conclusion that differences are increased and twelve to the conclusion that differences are decreased, we cannot

³¹Good, C. V. *How to do Research in Education*. Baltimore: Warwick and York, 1928, p. 146.

³²Woody, Clifford. "The Values of Educational Research to the Classroom Teacher," *Journal of Educational Research*, 16:175, October, 1927.

³³Monroe, W. S., et al. "Ten Years of Educational Research, 1918-1927." *University of Illinois Bulletin*, Vol. 25, No. 51, Bureau of Educational Research Bulletin No. 42. Urbana: University of Illinois, 1928, p. 84.

help wondering about our experiments and about the conclusions derived from them.³⁴

The following quotations show that the feeling of distrust for the results of educational experiments reported in the literature is not restricted to the men quoted above.

Or the investigator gives a few standard tests; he finds the pupils very deficient. He calls the teachers together; he arouses great enthusiasm, doubles the time to be given to the subject, introduces an entirely different method, works up a high degree of skill in the use of it, and after a few months "concludes that the new *method* was alone responsible for the improvement observed. Everybody should at once follow suit."³⁵

Perhaps the extreme case is that of the examination and treatment of a fourth-grade pupil, found to be deficient in reading. After a brief diagnosis and application of "remedial measures," the announcement is gravely made that in the light of this experience we may safely assume that the proper method of dealing with all fourth-grade pupils having similar disabilities is that used in this case. Making a sweeping generalization on the basis of a single instance would seem to exhaust the possibilities of the scientific method in education and leave nothing to be desired in the way of economy, efficiency, and dispatch. Many of the "conclusions" appended to recent "scientific" investigations have little more to support them. We are in a fair way to be able to prove anything. A few figures and a graph will turn the trick.³⁶

We have observed in many of the practices of educational research workers a tendency to shallowness. We have taken occasion to point out more than once a lack of sustained effort, a willingness to flit from one thing to another, and an unwillingness to stay with a problem until fundamental—the word seems to haunt us—until fundamental results are secured . . . We are threatened with becoming mere dabblers in research, foolishly confident of the virtues of a fresh start.³⁷

Another line of inquiry has to do with the operations of the classroom. Some of the most influential investigations made in recent years have had to do with the problems of classroom procedure, and yet anyone who contrasts the facts which appear during observation of a good teacher and the recommendations made in even our best textbooks on methods knows that the scientific description of teaching is in its infancy.³⁸

We must use greater care to make certain that the conclusions we state in our reports follow logically from the data presented. Too many reports state conclusions that are not fully supported by the research data included in them. This association should interest itself in the quality as well as in the quantity of educational research.³⁹

Nevertheless, I can not evade the conviction that, relatively speaking, the published research in education is, on the whole, inferior in quality, and more especially inferior in ultimate significance, to the published research in other branches of scientific endeavor. Too many contributions seem essentially futile.

³⁴Henmon, V.A.C. "Measurement and Experimentation in Educational Methods," *Journal of Educational Research*, 18:185-186, October, 1928.

³⁵"Assuming the Major Premise," *Journal of Educational Method*, 2:229, February, 1923. (An editorial.)

³⁶*Loc. cit.*

³⁷"Fundamentalism in Research," *Journal of Educational Research*, 9:331, April, 1924. (An editorial.)

³⁸Judd, C. H. "Research in Elementary Education," *Journal of Educational Psychology*, 17:224-225, April, 1926.

³⁹Trabue, M. R. "Educational Research in 1925," *Journal of Educational Research*, 13:344, May, 1926.

After you read them, you feel like saying: "Well, suppose it is true; what of it?"⁴⁰

It is easily charged, and must be admitted, that initial effort to apply experimental techniques to the intricate problems of human affairs is often a lame and halting procedure, and far too much may easily be claimed by way of fact and inference as forthcoming from first efforts in this direction.⁴¹

Educational experimentation in a plateau period. The previous discussion has traced the past of educational experimentation. It has been shown that this method of answering thought questions in education has undergone an evolutionary development over a period of some centuries. The contributions of laboratory experimentation in the field of psychology and the aid rendered by the production of more suitable measuring instruments have been mentioned. Finally, some indication was given of the effect of the writings of prominent leaders in the field. These writings, stimulating and optimistic for the most part a few years ago, have been replaced by others reflecting disillusionment and an attitude of distrust for this method of research in education. However, the feeling seems to be that the fault lies not with the fundamental theory of experimentation nor with the difficulties involved when human beings are the subjects of experiment, but with our present experimental techniques. That is to say, there is a feeling that the mediocre quality of experimental results has been due to the lack of adequate techniques and to the belief that conclusive results will be secured when techniques are perfected. If this is true, possibly an analogy may be drawn with plateaus in learning. The lower order "habits" have been formed; improvement is at a standstill until higher order "habits" have been perfected.

Definition of experiment. The foregoing discussion has been given before defining the term "experiment," since the concept represented by this term has undergone an evolution analogous to the historical development of the method itself. The concept of "educational experimentation" that is expressed in the following paragraphs is the culmination of this development; it is, therefore, appropriately given at this time.

A child's achievement from a period of learning is the resultant of several educative factors. "Experimentation" is the name given to the type of educational research in which the investigator controls the educative factors to which a child or group of children is subjected during the period of inquiry and observes the resulting achieve-

⁴⁰Whipple, G. M. "The Improvement of Educational Research," *School and Society*, 26:251, August 27, 1927.

⁴¹Haggerty, M. E. "The Scholarly Study of College Education," *Journal of Educational Research*, 19:140, February, 1929. (An editorial.)

ment. The meaning of this definition will be clearer if consideration is given to some of the procedures employed. In the simplest type of educational experiment the investigator seeks to evaluate the influence of some one educative or "experimental" factor on a single group of children. He must start the experiment with some measurement of the initial attainment of the children in the trait or ability to be influenced. He then subjects the group to the experimental factor, such as a particular type of drill material in arithmetic, for the duration of the experiment. At the end, the investigator applies a final test for the purpose of determining the gain in achievement that has resulted from the application of the experimental factor. This simple type of experiment may be illustrated by describing briefly one reported by Glick.

This experiment had as its problem the determination of the effect of practice on intelligence test scores.⁴² Students were tested at the start of the experiment with one of the forms of the Army Alpha Intelligence Examination. The experimental factor consisted of practice exercises similar to, but not identical with, the exercises in the sub-tests of the intelligence examination. After certain intervals other forms of the Army Alpha were administered to these same students. The increase in scores from one application of the intelligence examination to another is a measure of the effect of the experimental factor operating over the interval of practice.

A single group experiment, such as the one just described, is appropriate when it is evident that the effect is to be ascribed to the operation of only one educative factor. In many cases, however, such a situation does not exist. Instead of being able to have the subjects influenced by one factor, they are influenced by many. If one were to use a single group of individuals, it would be impossible to say how much of the effect was due to any particular cause. When two or more groups are used, it is possible to subject them to identical conditions with the exception of the experimental factor. The difference in the effect when one group is compared with the other may be ascribed to the operation of this single factor. This method may be illustrated by describing an experiment by Anibel.

The problem of this investigation was the determination of the comparative effectiveness of the lecture-demonstration and individual-laboratory methods in chemistry.⁴³ Anibel set up two groups pre-

⁴²Glick, H. N. "Effect of Practice on Intelligence Tests," *University of Illinois Bulletin*, Vol. 23, No. 3, Bureau of Educational Research Bulletin No. 27. Urbana: University of Illinois, 1925, p. 6.

⁴³Anibel, F. G. "Comparative Effectiveness of the Lecture-Demonstration and Individual Laboratory Method," *Journal of Educational Research*, 13:355-65, May, 1926.

sumably equivalent in intelligence as follows: "A student from the lecture-demonstration or test group was paired, for the purpose of comparing achievement records, with a student from the individual-laboratory class or control group."⁴⁴ The investigator sought to keep certain educative factors identical in both of these groups. He states, "all classes in chemistry met five times per week for forty-five minute periods The classroom instruction was identical for the two groups, thus equalizing any factors that might be present in classroom instruction."⁴⁵ After getting these educative factors under control, it was possible for him to use different instructional procedures in the laboratory instruction of the groups. These instructional procedures were demonstrations of chemical experiments before one group of the pupils, while the pupils of the other group were required to perform the experiments for themselves. The difference between the two groups in gains in achievement is to be ascribed, with certain limitations, to the superiority of one of these instructional procedures over the other.

The problem of this investigation. The previous discussion has indicated to the reader that educational experimentation has undergone a long period of development. In the last twenty years, hundreds of learning experiments have been performed under school conditions. Early enthusiasm has been replaced to some extent by expressions of distrust. Hence, there appears to be a need for a critical analysis of experimentation as a procedure in educational research. In the chapters that follow, the present writers attempt:

1. To describe in detail the procedure that should be followed in educational experimentation to arrive at dependable conclusions.
2. To apply the procedure outlined as a means of evaluating a group of experiments.
3. To formulate an appraisal of the present status of experimentation as a procedure in educational research.

⁴⁴Anibel, *op. cit.*, p. 356.

⁴⁵*Ibid.*, p. 356.

CHAPTER II

THE REQUIREMENTS FOR CONTROLLED GROUP EXPERIMENTATION

The general plan of controlled group experimentation. In a controlled experiment there are two groups of pupils which are equivalent in all respects that affect learning in the field of experimentation. The instruction and other educative influences to which the two groups are subjected are the same except for one factor. This *experimental factor* may be an instructional technique, the size of the class, the textbook, or any other educative influence that may be studied experimentally. The difference in the gains in achievement made by the two groups during the period of experimentation is an index of the relative merits of the two forms of the experimental factor.¹ This plan may be described more formally as follows:

Let E_1 = mean initial status of experimental group in the abilities that the application of the experimental factor is expected to affect.

C_1 = mean initial status of control group in the same abilities.

E_2 = mean final status of experimental group in the abilities that the application of the experimental factor is expected to have affected.

C_2 = mean final status of the control group in the same abilities.

$E_2 - E_1$ = Gain E

$C_2 - C_1$ = Gain C

The "difference in gain," D, equals the result found when Gain C is subtracted from Gain E. If this difference is positive, the status of the experimental factor prevailing in the experimental group represents the more effective instructional conditions. If the difference is negative, the opposite conclusion is indicated. The validity of this interpretation depends upon the satisfaction of three requirements: (1) The two groups of pupils are equivalent at the start of the experiment. (2) All educative factors except the experimental one are the same for both groups. (3) The measures of achievement from which the gains are computed are both valid and accurate. When any one

¹An educational experiment may involve more than two groups of pupils and may be more complex in other respects, but the following discussion assumes the simple plan described here. Later, attention will be given to the procedure of the more complex experiments.

of these requirements is not fully realized, it becomes necessary to discount the difference in the gains made by the two groups. If the difference is small, and if the departure from the requirement is large, the relative merits of the two procedures compared will not have been determined.

Questions considered in this chapter. In this chapter the following questions are considered:

1. What is required to secure equivalent groups of pupils?
2. What are the important educative factors that affect the achievements of pupils?
3. What is involved in controlling the important educative factors that affect pupil achievements?

The analysis of the causes that affect achievement and the determination of the important educative factors is, of course, only tentative. Although there are a number of causal investigations in which an attempt has been made to determine the contributions of certain factors to achievement and their relative potency, the available evidence is fragmentary, and there is reason to doubt the validity of the findings, at least in a number of cases.² Consideration of experimentation as a research procedure, however, requires that an attempt be made to identify the more important educative factors. In doing this, the present writers have endeavored to make use of the best data obtainable, but they are not unmindful of the fact that in this case the best data may not be sufficiently valid to accomplish the desired result. Consequently, the conclusions presented in the following pages relating to the factors to be considered in educational experimentation should be thought of as tentative and subject to modification when more dependable data are available.

The significant characteristics of pupil material. In listing the significant characteristics of pupil material, we are concerned only with those that affect achievement in the field of experimentation. Obviously, such characteristics as color of hair, degree of beauty, and height, do not belong in the list. On the other hand, general intelligence and previous achievement in the field of experimentation must be included. The following characteristics appear to deserve consideration:

1. General intelligence in terms of point scores, or of mental age
2. Chronological age
3. Previous achievement in the field of experimentation

²Burks, B. S. "On the Inadequacy of the Partial and Multiple Correlation Technique," *Journal of Educational Psychology*, 17:532-40, 625-30, November, December, 1926.

4. Study habits
5. Personality traits (attitudes, ideals, and interests)
6. Physical condition (health)
7. Sex
8. Race

1. There is abundant evidence that *general intelligence*, as measured by typical intelligence tests, influences the achievement of children. Many investigators have concluded that it is the most important factor. The following conclusion from the report of a recent investigation by Heilman is indicative of this belief:

Our results also appear to show that under the prevailing conditions of the home and school organization, intellectual endowment, or whatever is measured by the Stanford-Binet test, has by far the most powerful influence in determining differences in achievement in the traditional curriculum. It is not unlikely that a similar statement could be made for achievement in general.³

This may not be true in the case of some pupils, but the general statement appears to be justified. Hence, general intelligence (mental age, or test scores) may be placed at the head of the list of significant characteristics of pupil material.

2. The significance of *chronological age*⁴ becomes apparent when a child having a mental age of twelve years and a chronological age of ten years is compared with one whose corresponding ages are twelve and fifteen. The first child has an I. Q. of 120 and the second one, an I. Q. of 80. Although the two children have equivalent mental ages, the first one is "bright" and the second is "dull." The significance of chronological age is further shown by a comparison of two children of the same I. Q. but of different chronological ages. Although the children are equally "bright," the difference in mental ages, as well as the differences in physiological and social maturity, emphasizes the importance of chronological age as a factor in school achievement. The importance of chronological age as a factor in school achievement is recognized by those who recommend homogeneous grouping on the two bases, mental age and chronological age.⁵ An excellent discussion of the influence of chronological age, or the maturity of which it is an index, is to be found in a recent monograph by Commins.⁶

³Heilman, J. D. "Factors Determining Achievement and Grade Location," *The Pedagogical Seminary and Journal of Genetic Psychology*, 36:454, September, 1929.

For a comprehensive account of the influence of general intelligence upon school achievement, see:

Terman, L. M., et al. "Nature and Nurture, Their Influence Upon Achievement," *Twenty-Seventh Yearbook of the National Society for the Study of Education, Part II*. Bloomington, Illinois: Public School Publishing Company, 1928. 397 p.

⁴The I. Q. might have been listed as a pupil characteristic instead of chronological age. The measurement of the latter is objective; hence it is to be preferred. When chronological age is included, the I. Q. is superfluous.

⁵Freeman, F. N. *Mental Tests*. Boston: Houghton Mifflin Company, 1926, p. 23.

⁶Commins, W. D. "Maturity and Education," *Educational Research Bulletin*, Vol. 3, No. 7, Catholic University of America. Washington: Catholic University Press, 1928, p. 36.

3. *Previous achievement*⁷ is a significant characteristic of the pupil material when it functions as a prerequisite for the learning involved in the experiment. For example, ability to read functions as a tool in learning arithmetic, geography, history, literature, and the like.⁸ Certain abilities in arithmetic and algebra function as tools in the study of chemistry, and achievement in chemistry contributes to achievement in physics. Achievement in the first year of a foreign language functions as a tool in the more advanced study of that language. It would be easy to enumerate a large number of cases in which abilities engendered in a school subject function later in the learning of that subject or related subjects.

Abilities that function as a prerequisite for learning in one school subject may, or may not, be significant for learning in another school subject. For example, achievement in the first year of a given foreign language would be of more significance in an experiment in the second year of that language than it would be in an experiment in a different language. Achievement in the first year of a foreign language would probably be of least significance in an experiment that involved type-writing. The previous achievement of children becomes of increasing importance as a factor in the achievement of the experiment in proportion to the extent to which the children have experienced subject-matter similar in content to that of the experiment.

4. The term *study habits* is used to designate a somewhat indefinite group of procedures employed in doing assignments. Their general nature is indicated by samples of the rules proposed by Whipple.⁹

- a. When possible, prepare the advance assignment in a given subject directly after the day's recitation in it.
- b. Form a time-study habit.
- c. Form a place-study habit.
- d. Don't stop work when you have just barely learned the material, but keep on until you have over-learned it.
- e. Begin work promptly.
- f. Train yourself to ignore distractions from without.
- g. Do your work with the intent to learn and to remember.
- h. Mentally review every paragraph as soon as you have read it.

⁷The total outcome of learning includes general patterns of conduct as well as specific habits and knowledge. Among the possible outcomes are study habits which are not included here under the head of "previous achievement."

⁸For a discussion of the contribution of achievement in reading to achievement in arithmetic, see:

Lessenger, W. E. "Reading Difficulties in Arithmetical Computation," *Journal of Educational Research*, 11:287-91, April, 1925.

⁹Whipple, G. M. *How to Study Effectively*. Bloomington, Illinois: Public School Publishing Company, 1927. 96 p. (Revised edition).

It is evident from an examination of this list that study habits vary widely in specificity and in value. Habits with respect to the time and place in which study is carried on are far more specific in nature than those of mentally reviewing a paragraph or studying with the intent to remember. Mentally reviewing a paragraph implies organization of knowledge and as such may be classed as a method of thinking. Studying with the intent to remember is an attitude toward study that may function in all study situations. While the precise effect of conforming to recommended study habits is not known, it is probable that their value is a function of the degree of their generalization. The more specific study habits may, or may not, be useful since the brighter students can get along without them. The more general study habits are indispensable since they are the methods and the driving forces of reflective thinking. The following quotations tend to substantiate the above contentions with respect to the importance of this factor in learning activity:

When the pupil had acquired effective methods of study and observed that he really could learn, a new and happy interest was the common result. Some of the pupils, for example, began to read books, a thing that had never previously been done because reading was difficult work.¹⁰

During the latter half of the period, the methods used in the preparation of actual assignments were given special attention. Certain of the students made noticeable progress; one sophomore made his first "A" since entering high school while receiving training in ancient history, a freshman showed a decided gain in algebra.¹¹

The work of the class was greatly improved through the use of better methods of study. The pupils became more independent and more alert to the importance of the history topics and their relation to our lives today.¹²

. . . that when reasonably effective methods are used to control admission to college, the failure of students subsequently is not commonly due to inadequate intelligence; that, on the contrary, the failures are mainly due to several factors, among which, according to the reports gained from the students themselves, a prominent place is to be assigned to neglect of proper instructions in the art of study.¹³

What is still more significant, perhaps, is the fact that 87 per cent of the freshman students enrolled in these two "How to Study" sections completed their total enrolment in the university in a way that was satisfactory to *all* their instructors, while only about half of the freshman students entering the university last year, (46 per cent of the boys and 63 per cent of the girls) completed their total enrolment in a satisfactory manner.¹⁴

¹⁰Gates, A. I. "A Study of Reading and Spelling With Special Reference to Disability," *Journal of Educational Research*, 6:20, June, 1922.

¹¹Monroe, W. S. and Mohlman, D. K. "Training in the Technique of Study," *University of Illinois Bulletin*, Vol. 22, No. 2, Bureau of Educational Research Bulletin No. 20. Urbana: University of Illinois, 1924, p. 20.

¹²Fisk, E. M. "An Experiment in How to Study," *Elementary School Journal*, 27:138, October, 1926.

¹³Whipple, G. M. "Experiments in Teaching Students How to Study," *Journal of Educational Research*, 19:1, January, 1929.

¹⁴Book, W. F. "Results Obtained in a Special 'How to Study' Course Given to College Students," *School and Society*, 26:534, October 22, 1927.

On the same level of intelligence the methods of study are of great importance. As a rule the students of low intelligence who were successful in college were employing good study technique.¹⁵

Symonds, in concluding an excellent review of the research on "How to Study," is probably correct in stating ". . . that the commonly accepted rules of study are often non-consequential,"¹⁶ but his later statement, "While one would not deny the fact that all these rules are factors in efficient study one may question their relative importance,"¹⁷ would cause one to believe that some study habits are important factors in learning activity. While further experimentation is necessary before it may be known definitely which of the more specific study habits are most effective and, therefore, most important to the experimenter, it may be safely stated that the status of the pupils with respect to study habits, particularly the more general ones, should be considered by the experimenter in forming equivalent groups.

5. The term *personality traits* is used to designate a group of attitudes, interests, ideals, and other reaction tendencies. This group has not been fully analyzed, but several traits have been identified that appear to influence pupil achievement. After canvassing the available literature, Herriott listed five attitudes:

- a. Ambitious — Indifferent
- b. Cheerful — Despondent
- c. Evaluative — Non-evaluative
- d. Persevering — Vascillating
- e. Self-confident — Dependent¹⁸

His investigation to determine the importance of these attitudes as factors of scholastic success indicated that the last three are major factors. The third and fourth are related to scholastic success in a positive way. That is to say, the student who has the attitudes of evaluating and persevering is more successful, in general, than the student whose behavior is characterized by their opposites, non-evaluative and vacillating. The fifth is related to success in a negative way. Self-confidence is apt to be dangerous as an attitude of coxsureness,

¹⁵Ross, C. C. and Klise, N. M. "Study Methods of College Students in Relation to Intelligence and Achievement," *Educational Administration and Supervision*, 13:562, November, 1927.

¹⁶Symonds, P. M. "Methods of Investigation of Study Habits," *School and Society*, 24:151, July 31, 1926.

¹⁷*Ibid.*, p. 152.

¹⁸Herriott, M. E. "Attitudes as Factors of Scholastic Success," *University of Illinois Bulletin*, Vol. 27, No. 2, Bureau of Educational Research Bulletin No. 47. Urbana: University of Illinois, 1929, p. 31.

The two words used to designate an attitude represent opposite extremes of what Herriott calls a "single attitude." Thus the pupil who is "ambitious" and the one who is "indifferent" are merely exhibiting extreme differences in the same attitude rather than two different attitudes.

or the instructor is likely to favor an attitude of dependence on himself and the text.¹⁹ Herriott concludes with the following statement relative to the significance of these traits: "These data support the belief expressed by many authorities that traits such as study habits and attitudes are factors of success comparable to the seemingly more tangible and more usually measured factors such as intelligence and previous preparation."²⁰

Statements by several other investigators are reproduced as indicative of the recognition of the importance of personality traits as factors in learning:

"Without doubt, some of the backwardness was due to a lack of interest and effort,"²¹

It is possible to obtain statements of personality traits (moral attitudes, emotional maladjustments, and interests) which give correlations of very appreciable size (about as large as those obtained between tests of intelligence and marks) with academic success.²²

The most significant factor next to estimated intelligence in its association with scholarship appeared to be the quality, or composite of qualities, defined as school attitude.²³

It would appear from all available data that the relationship between educational interests and abilities as expressed in school grades is represented by average correlations between + .20 and + .40.²⁴

The major groups of causes of scholastic deficiency were found to appear in the following order of significance.²⁵

	Significance scores
Motivation and interests.....	265
Intellectual factors.....	265
Emotional factors.....	221
Educational factors.....	202
Environmental factors.....	148
Study habits and methods.....	113
Physical factors.....	90
Teaching methods and content.....	32
Motor factors.....	28

The data give a minus third-order correlation between general health and school marks, and a relatively low correlation between preparation and marks; the high correlation of "school attitude" with marks is the striking feature of the situation.²⁶

¹⁹Herriott, *op. cit.*, p. 42-43.

²⁰*Ibid.*, p. 44.

²¹Gates, *op. cit.*, p. 20.

²²Chambers, O. R. "Measurement of Personality Traits," *Research Adventures in University Teaching*. Bloomington, Illinois: Public School Publishing Company, 1927, p. 76.

²³Fleming, C. W. "A Detailed Analysis of Achievement in the High School," *Teachers College, Columbia University Contributions to Education*, No. 196. New York: Bureau of Publications, Teachers College, Columbia University, 1925, p. 185.

²⁴Fryer, Douglas. "Interest and Ability in Educational Guidance," *Journal of Educational Research*, 16:36, June, 1927.

²⁵Ohmann, O. A. "A Study of the Causes of Scholastic Deficiencies in Engineering by the Individual Case Method," *University of Iowa Studies in Education*, Vol. 3, No. 7. Iowa City: University of Iowa, 1927, p. 56-57.

²⁶Pressey, S. L. "An Attempt to Measure the Comparative Importance of General Intelligence and Certain Character Traits in Contributing to Success in School," *Elementary School Journal*, 21:229, November, 1920.

In the light of these investigations and several others that might be mentioned, it appears that personality traits form a significant characteristic of pupils when considered as learners in school subjects.

6. Severe illness and certain types of physical defects, such as blindness, or deafness, are handicaps in learning, but the significance of all aspects of a pupil's *physical condition* is not known. That such physical defects as adenoids, enlarged tonsils, deafness, and poor vision are factors in school achievement is indicated in the following quotations:

In every case, except in that of vision, the children rated as "dull" are found to be suffering from physical defects to greater degree than the "normal" or "bright" children.²⁷

No evidence is apparent that the good or bad condition of the tonsils had any effect on intelligence, but those children who showed improvement in condition of tonsils had the highest rate in school achievement.²⁸

The conclusions of Sandwick,²⁹ Hall and Crosby,³⁰ Sumner,³¹ and Mallory³² concur with the above in emphasizing the importance of physical condition as a factor in school achievement. The conclusions of two recent investigations are not in harmony with those just given, since the claim is made that physical defects are a minor factor in scholastic success:

Physical defects were not much more prevalent among the retarded group than among the normally progressing group, and therefore, nonpromotion could not be attributed to that cause.³³

It is obvious, of course, that very serious defects will handicap a child in learning. Their influence is probably both direct and indirect. But lesser defects do not appear to have any causal connection with poor scholarship. In fact no association of any kind appears in these data between physical health and achievement. Even comparatively serious defects do not necessarily entail poor achievement.³⁴

The writers are inclined to favor the view expressed in the conclusion just quoted from Westenberger. The experimenter is justified in considering physical condition, so far as the experiment is concerned, a minor factor in learning activity. Extreme cases of ill

²⁷Ayres, L. P. *Laggards in Our Schools*. New York: The Russell Sage Foundation, 1909, p. 125.

²⁸Hoefler, Carolyn and Hardy, M. C. "The Influence of Improvement in Physical Condition on Intelligence and Educational Achievement," *Twenty-Seventh Yearbook of the National Society for the Study of Education*, Part I. Bloomington, Illinois: Public School Publishing Company, 1928, p. 387.

²⁹Sandwick, R. L. "Correlation of Physical Health and Mental Efficiency," *Journal of Educational Research*, 1:199-203, March, 1920.

³⁰Hall, Irene, and Crosby, Amy. "A Study of the Causes of Inferior Scholarship of Pupils in Low First Grade," *Journal of Educational Research*, 14:375-83, December, 1926.

³¹Sumner, H. W. "Health and Home Factors in Non-Promotions," *Chicago School Journal*, 9:101-103, November, 1926.

³²Mallory, J. N. "A Study of the Relation of Some Physical Defects to Achievement in the Elementary School," *George Peabody College for Teachers, Contributions to Education*, No. 9. Nashville: George Peabody College for Teachers, 1922. 78 p.

³³Stalnaker, E. M. and Roller, R. D., Jr. "A Study of One Hundred Nonpromoted Children," *Journal of Educational Research*, 16:270, November, 1927.

³⁴Westenberger, E. J. "A Study of the Influence of Physical Defects upon Intelligence and achievement," *The Catholic University of America, Educational Research Bulletin*, Vol. 2, No. 9. Washington: The Catholic Education Press, 1927, p. 45.

health, or physical defect, tend to eliminate themselves from the ordinary schoolroom. It is probable that both groups will contain approximately the same number of children with defects, due to the operation of chance, and even if they do not, the inequality would have to be considerable to influence appreciably the mean achievement of the group.

7. Both Thorndike³⁵ and Starch³⁶ have concluded on the basis of the findings of several investigations reviewed by them that *sex* is a very minor factor in learning. The conclusions of Minnick³⁷ and Touton³⁸ are in agreement with those of Thorndike and Starch, but a recent investigation by Webb shows that when boys and girls of the same general intelligence are compared with respect to achievement in geometry, the boys exceed the girls on the lower mental levels but are exceeded by them on the higher.³⁹ He states "that those studies of sex differences, which neglect to take into account the factor of mental ability, fail to discover significant differences between sex groups which may exist at one mental age level, but not at another."⁴⁰

Fitzgerald and Ludeman have reported the results of an investigation in which it was found that in the sixth and seventh grades boys achieved more in history than girls, but in the eighth grade the greater achievement was shown by the girls.⁴¹ Van Wagenen found sex differences in learning American history to be great enough to warrant establishing two sets of norms for his "American History Scales."⁴² Fisher discovered that a loss of efficiency in mechanical learning takes place a year earlier in girls than in boys.⁴³

From these more recent studies the conclusion may be drawn that sex is a factor of less importance than those described in the preceding pages, but it should not be neglected by the educational experimenter who seeks highly dependable results.

³⁵Thorndike, E. L. *Educational Psychology*, Vol. III. New York: Teachers College, Columbia University, 1914, p. 169-205.

³⁶Starch, Daniel. *Educational Psychology*. New York: The Macmillan Company, 1919, p. 63-72.

³⁷Minnick, J. H. "A Comparative Study of the Mathematical Ability of Boys and Girls," *School Review*, 23:73-84, February, 1915.

³⁸Touton, F. C. "Sex Differences in Geometric Abilities," *Journal of Educational Psychology*, 15:246-47, April, 1924.

³⁹Webb, P. E. "A Study of Geometric Abilities Among Boys and Girls of Equal Mental Abilities," *Journal of Educational Research*, 15:256-62, April, 1927.

⁴⁰*Ibid.*, p. 262.

⁴¹Fitzgerald, J. A. and Ludeman, W. W. "Sex Differences in History Ability," *Peabody Journal of Education*, 6:175-81, November, 1928.

⁴²Van Wagenen, M. J. "Historical Information and Judgment in Pupils of Elementary Schools," *Teachers College, Columbia University Contributions to Education*, No. 101. New York: Bureau of Publications, Teachers College, Columbia University, 1919. 74 p.

⁴³Fisher, V. E. "A Few Notes on Age and Sex Differences in Mechanical Learning," *Journal of Educational Psychology*, 18:562-564, November, 1927.

8. The importance of *race* as a factor in learning is difficult to determine.⁴⁴ Various other factors, such as language handicap, social status, parental occupation, and other environmental influences, tend to obscure its significance. Although it may be impossible to ascribe differences between racial groups to something inherent in their respective races, it is none the less evident that actual differences exist in school achievement. The following are typical of the conclusions reached by investigators.

Statistical data carefully collected and presented in the foregoing study indicate rather conclusively that primary French-speaking children in certain Louisiana parishes are lower in achievement than English-speaking children and are seriously retarded.⁴⁵

The authors ascribe the lower achievement of the French-speaking children to the language handicap.

The number of months by which the median educational age of the entire group of white children exceeds the median of the negro group was found to be 16.7 months. It was found further that only 14.5 per cent of the negro children reach or exceed the median educational age of the white children.⁴⁶

A recent study of the retardation of seventeen hundred children of immigrants in two cities of northern Michigan shows that retardation according to nationality follows very closely the median intelligence quotients of the nationalities.⁴⁷

The conclusions may be expressed that whether there are inherent differences in race or not, characteristics distinctive of racial groups, such as language ability, social status, parental occupations, customs, prejudices, attitudes, and the like, are of enough significance in school work that race must be considered by the educational experimenter who desires to set up equivalent groups.

Significant characteristics of pupil material not independent traits.

Several of the characteristics of pupil material described in the preceding pages are not independent. The correlation between general intelligence and previous achievement in any one school subject for a single school grade and general intelligence is likely to fall between .11 and .69.⁴⁸ The interdependence of general intelligence and general school achievement is expressed in the following statement from

⁴⁴Race may influence achievement indirectly through intelligence since races will differ in capacity to learn as they vary in intelligence. This aspect of the race factor does not concern us here since equating pupils with respect to intelligence will take care of it. We are interested in the more direct influences of racial characteristics on learning.

⁴⁵Brouillette, J. W., Foote, I. P., Robert, E. B., and Terrebonne, L. P. *A Comparative Study of the School Progress of Foreign-Speaking and English-Speaking Children in the Early Elementary Grades*. Chicago: Scott, Foresman and Company, 1928, p. 62-63.

⁴⁶Witty, P. A. and Decker, A. I. "A Comparative Study of the Educational Attainment of Negro and White Children," *Journal of Educational Psychology*, 18:498-99, October, 1927.

⁴⁷Brown, G. L. "Intelligence as Related to Nationality," *Journal of Educational Research*, 5:326, April, 1922.

⁴⁸Gates, A. I. "The Correlations of Achievement in School Subjects with Intelligence Tests and Other Variables," *Journal of Educational Psychology*, 13:280, May, 1922.

Kelley: "On the average, in the neighborhood of .90 of the capacity measured by an all-round achievement battery score,—reading, arithmetic, science, history, etc.,—and of the capacity measured by a general intelligence test is one and the same."⁴⁹ General intelligence is also positively related to study habits and personality traits. Butterweck has shown that the brighter pupils in his investigation tended to employ the better study habits.⁵⁰ Herriott, in the research already referred to, found a small though significant positive relationship between general intelligence and the personality traits listed on page 23. In typical grade groups the correlation between chronological age and mental age is negative. Brighter children tend to be accelerated, while duller children are retarded, so that in a given school class, the relatively brighter are the younger, and the relatively duller are the older. Terman reports a high negative correlation ($-.74$) between the I. Q. and the chronological age of a group of children entering high school as freshmen.⁵¹ Baldwin has shown that, in general, children who are gifted mentally are also superior physically.⁵² Hoefler and Hardy state that children whose physical condition is good have a more rapid mental growth than children whose physical condition is fair.⁵³ Sex and race are not to be thought of as variables that may be correlated with intelligence. They are the least significant of the factors listed as characteristics of pupil material and are the most independent.

The fact that several of the characteristics are positively correlated with general intelligence means that if two groups of pupils are equivalent with respect to mental age, or intelligence-test scores, under typical conditions, they are likely to approach equivalence with respect to previous achievement, study habits, personality traits, and physical condition.

Securing equivalent groups for a controlled experiment.⁵⁴ It is relatively easy to assemble two groups that are equivalent with reference to a given characteristic, provided that characteristic can be measured accurately. For example, pupils may be paired on the basis of mental age, or intelligence-test scores, so that for each pupil in one

⁴⁹Kelley, T. L. *Interpretation of Educational Measurements*. Yonkers-on-Hudson, New York: World Book Company, 1927, p. 21.

⁵⁰Butterweck, J. S. "The How to Study Problem," *Journal of Educational Research*, 18:66-76, June, 1928.

⁵¹Terman, L. M. *The Intelligence of School Children*. Boston: Houghton Mifflin and Company, 1919, p. 82.

⁵²Baldwin, B. T. "Anthropometric Measurements," *Genetic Studies of Genius*, Vol. 1. Stanford University, California: Stanford University Press, 1925, p. 135-71.

⁵³Hoefler and Hardy, *op. cit.*, p. 371-87.

⁵⁴One group of pupils is equivalent to another with respect to a given characteristic when for each pupil in one group there is a mate in the second who possesses the same amount of the characteristic. An approach to equivalence is secured when the central tendency and variability of one group with respect to a given characteristic are equal to these measures of the other.

group there will be a mate in the second group having the same mental age or test score. Obviously, it would be difficult, if not impossible, under typical conditions to assemble two groups by locating pairs of pupils that are equivalent in respect to all significant characteristics. Hence, in assembling equivalent groups by pairing, the experimenter usually considers only one or at most two characteristics. When the groups have been assembled, they should be checked for equivalence with respect to the remaining significant characteristics. For example, if two groups have been assembled by pairing pupils having the same mental ages, or intelligence-test scores, the mean and standard deviation of each group should be calculated for chronological age, and previous achievement, when it is significant. If the mean and standard deviation of one group are not approximately equal to those of the other group, adjustments should be made to secure approximate equality or the pair of groups rejected for experimentation. If adequate measuring instruments were available, it would be desirable also to check the equivalence of the groups with respect to study habits and personality traits in the same way. The equivalence of the groups with respect to sex and race should be checked to make certain the groups exhibit no marked differences with respect to these characteristics. The experimenter should also make certain that the two groups involve no serious differences in physical condition.

A technique seems to be evolving for selecting pairs of pupils on the basis of a composite measure of characteristics. According to one technique, if it is desired to pair children on the basis of two characteristics, such as intelligence and previous achievement, a correlation chart of the test scores for intelligence and achievement may be constructed. The position of each child with respect to both of these characteristics is shown by a single dot on the chart. The experimenter selects his pairs by locating dots which are closest together. An illustration of the use of this technique is to be found in the report of an experiment by Butterweck.⁵⁵ Another technique is that of combining measures of different characteristics into one composite measure representing all of them. The children are then paired on the basis of the composite scores. The use of this technique is illustrated in the experiment of Douglass on the relative effectiveness of two sequences in supervised study.⁵⁶ Before this technique can be commended as the

⁵⁵Butterweck, J. S. "The Problem of Teaching High-School Pupils How to Study," *Teachers College, Columbia University Contributions to Education*, No. 237. New York: Bureau of Publications, Teachers College, Columbia University, 1926. 116 p.

⁵⁶Douglass, H. R. "The Experimental Comparison of the Relative Effectiveness of Two Sequences in Supervised Study," *University of Oregon Publication*, Education Series, Vol. 1, No. 4. Eugene, Oregon: University of Oregon, 1927, p. 173-218.

best technique to use, research must determine the weights to be given each characteristic in the composite score.

Melby and Lien⁵⁷ have reported on a technique for controlling pupil factors which does not involve the use of pairing procedures. The initial status of three or more available groups is determined with respect to intelligence and previous achievement. After the order of superiority of the groups has been determined from comparison of their initial status, the experimenter selects one of the average groups as the experimental group. The assumption is made that if the mediocre experimental group exceeds in final achievement the initially superior group, then superiority of the experimental factor is dependably indicated. If, however, the final achievement of the mediocre experimental group falls below that of the initially inferior group, then the inferiority of the experimental factor is shown. The technique is commendable in that it permits the use of ordinary school classes without modification. It cannot be regarded, however, as anything more than a "practicable technique," as it is labeled by Melby and Lien. The technique lacks precision in that the difference in gains in achievement is not ascribable to educative factors alone, as is the case when equivalent groups are used. Since it lacks this precision, it is difficult to see how clear-cut conclusions may be drawn from an experiment in which it is used.

The educative factors that affect pupil achievement. The educative factors that affect pupil achievement are grouped here under the following heads:

- I. Teacher factors
- II. General school factors
- III. Extra-school factors

I. Teacher factors that affect pupil achievement. Amount of training, teaching experience, intellectual status, personality, physical condition, sex, and age are usually listed as important teacher factors, but they influence pupil achievement for the most part indirectly through their contributions to more immediate factors. For example, training, amount of teaching experience, and intellectual status contribute to the teacher's instructional techniques and to his skill in the use of them; hence, these factors influence pupil achievement indirectly. Our problem here is to determine the teacher factors that influence pupil achievement directly.

⁵⁷Melby, E. O. and Lien, Agnes. "A Practicable Technique for Determining the Relative Effectiveness of Different Methods of Teaching," *Journal of Educational Research*, 19:255-264, April, 1929.

Credit is given to Professor John G. Rockwell of the University of Minnesota for devising this technique and using it in an experiment on thyroid deficiency.

In *The Commonwealth Teacher-Training Study*, Charters and Waples determined a list of twenty-five teacher traits⁵⁸ by interviewing a number of persons considered competent and by listing the "trait names" and "trait actions" mentioned by these persons as characteristics of good teachers. In view of the comprehensiveness of this study, it might be argued that the twenty-five traits listed, or at least those of highest rank, should be taken as the teacher factors that affect pupil achievement and, hence, as the teacher factors that should be controlled in an experiment. It does not appear satisfactory to do this. The list is too long and does not include instructional techniques or classroom-management procedures. Consequently, the present writers propose the following list of teacher factors for consideration. Evidence of the potency of each of these factors is presented as a basis for a conclusion in regard to the ones that must be controlled in experimentation in order to avoid introducing a serious error in the results.

1. Instructional techniques
 - a. Learning exercises
 - b. Motivation procedures
 - c. Directive procedures
 - d. Diagnostic procedures
2. Classroom-management procedures
3. Skill in carrying out instructional techniques and classroom-management procedures
4. Zeal of the teacher with reference to experimental factor
5. Personality traits
6. Physical condition
7. Sex
8. Age

1. The more influential *instructional techniques* may be classified under four heads: (a) learning exercises; (b) motivation procedures; (c) directive procedures; (d) diagnostic procedures. The attention given to methods courses in the professional training of teachers is evidence of the conviction that the instructional techniques employed by a teacher affect the achievements of his pupils. Hence, it is not necessary to present evidence in justification of them as important teacher factors.⁵⁹ It should be noted, however, that the influence of

⁵⁸Charters, W. W. and Waples, Douglas. *The Commonwealth Teacher-Training Study*. Chicago: University of Chicago Press, 1929, p. 18.

⁵⁹Some indirect evidence is afforded by investigations of the relation between achievement in the field of methods of teaching and teaching ability. In one such investigation the partial correlation (between "ability to pass a professional test" and "general teaching ability") was found to be +.570.

Knight, F. B. "Qualities Related to Success in Teaching," *Teachers College, Columbia University Contributions to Education*, No. 120. New York: Bureau of Publications, Teachers College, Columbia University, 1922, p. 42.

a given technique depends on its appropriateness. In order to be most effective, a given technique must be suited to the pupils, compatible with the objectives to be attained, and supplemented by other techniques. For example, a learning exercise suitable for pupils on the lower levels of intelligence is not likely to be a good one for bright pupils. Certain types of drill exercises in arithmetic have been demonstrated to be effective relative to the attainment of certain objectives, but they are not effective when other objectives are to be attained. A "good" learning exercise is likely to be relatively ineffective unless it is supplemented by appropriate directive and diagnostic procedures. The rule that practice should be distributed rather than concentrated is further evidence that the influence of a given instructional technique depends upon factors other than its intrinsic character.

2. *Classroom-management procedures* include such items as taking the roll, distributing and collecting materials, starting the work of the period, and dismissing the class in case the pupils go to another room at the end of the period, and dealing with disciplinary cases. The importance of these procedures is generally recognized. In fact, until recent years the teacher's ability as a disciplinarian was considered to be the most important of his qualifications. While other aspects of teaching are now considered of more importance than the mere maintenance of order, adequate attention to routine matters of classroom management, inclusive of discipline, is regarded as essential for the promotion of the most suitable environment for learning. If, however, distinctly undesirable practices are avoided, it appears likely that variations in classroom-management procedures will not affect pupil achievement to a significant extent.

3. The effectiveness of an instructional technique or a classroom-management procedure depends upon the *skill* with which it is carried out. This factor was implied in the discussion of instructional techniques, but its importance justifies more specific consideration. Although we have no means of securing precise measures of teaching skill, it is obvious that some teachers are more skillful in carrying out certain instructional techniques than are other teachers. When a new technique is being compared with a familiar one, it is likely that the new one will be applied less skillfully. For example, suppose an experiment is devised to determine the effect of supervised study in comparison with study without supervision. Suppose further that the plan of supervising study has been worked out so that the procedure is specified in detail. If a teacher, who has become a skillful instructor under a plan that does not involve supervised study,

but who has not had experience in supervising study, attempts to teach one class employing supervised study and another without supervised study, it is reasonable to expect that he will be considerably more skillful in teaching the second class. If this is the case, the experiment would furnish a comparison between skillful teaching without supervised study and teaching with supervised study somewhat crudely carried out. Hence, the experiment would not yield satisfactory evidence of the relative merits of skillful teaching with supervised study and skillful teaching without supervised study.

An illustration of the recognition of the importance of skill as an educative factor is afforded by the Newark Phonics Experiment. The teachers of the experimental classes, the principals of the schools involved, and the members of the Experimental Committee met together and formulated a detailed working plan. Then the plan was tried out for a semester before the real experiment was begun.⁶⁰

4. The *zeal* that a teacher exhibits in carrying out the instructional techniques he is employing is a subtle factor. It is closely related to the factor of skill, and perhaps the two overlap to some extent, but there is evidence that indicates the presence of an important educative factor that differs in some respects from skill. It is reasonable to expect that a teacher will exhibit greater zeal when employing a method that he believes in than when employing one that he does not like. The influence upon pupil achievement of the teacher's preference in regard to methods is indicated in an unpublished report⁶¹ of an experiment to determine the relative merits of instructional procedures that may be designated as Method A and Method B. Several teachers cooperated in the experiment, each one teaching one class according to Method A and another class according to Method B. The following results were secured.

	Number	Mean Score	Mean Scholastic Grade
Pupils taught by Method A.....	417	71.5	83.9
Pupils taught by Method B.....	440	69.5	83.8
Gain in favor of Method A.....		2.0	

The teachers were asked to indicate which method they preferred. The following results were obtained when the data were tabulated according to the preference of the teachers.

⁶⁰Sexton, E. K. and Herron, J. S. "The Newark Phonics Experiment," *Elementary School Journal*, 28:690-701, May, 1928.

⁶¹The writers are indebted to Dr. Rosalie M. Parr, of the University of Illinois, for these data. A report of this study is to be published in the *Journal of Chemical Education*.

TEACHERS PREFERRING METHOD A

	<i>Number</i>	<i>Mean Score</i>	<i>Mean Scholastic Grade</i>
Pupils taught by Method A.....	131	75.0	84.8
Pupils taught by Method B.....	140	59.3	82.4
Gain in favor of Method A.....		15.7	

TEACHERS PREFERRING METHOD B

	<i>Number</i>	<i>Mean Score</i>	<i>Mean Scholastic Grade</i>
Pupils taught by Method A.....	180	67.2	85.4
Pupils taught by Method B.....	178	72.2	85.2
Gain in favor of Method B.....		5.0	

TEACHERS HAVING NO PREFERENCE

	<i>Number</i>	<i>Mean Score</i>	<i>Mean Scholastic Grade</i>
Pupils taught by Method A.....	80	67.0	82.7
Pupils taught by Method B.....	89	67.2	83.0
Gain in favor of Method B.....		0.2	

The experiment was carried on during the latter part of the semester and the scholastic grades are probably a fair index of the equivalence of the two groups of pupils. According to this criterion, the paired groups are approximately equivalent except in the case of those taught by teachers preferring Method A. The difference for this pair of groups, however, is small in comparison with the difference between the mean scores. Furthermore, it may be that the scholastic grades of these pupils were influenced by their performances during the experiment.⁶² The differences between the mean scores of the several pairs of groups strongly suggest that the preference of the teachers in regard to the method of teaching affected the achievements of the pupils. If it is assumed that the preference in regard to methods affected the zeal of the teachers, it follows that this characteristic of teaching is an important educative factor and, hence, must be controlled in order to secure dependable results.

Douglass⁶³ has reported data that may appear to be in opposition to the conclusion just stated. At the close of an experiment to determine the relative effectiveness of two sequences of supervised study,

⁶²The papers of the test given at the close of the experiment were scored by a central committee, and, consequently, the teachers did not know the results until after the grades had been assigned. But it is likely that the teachers had some idea of the achievements of the pupils during the experiment, and since the pupils taught by Method A by teachers preferring this method did much better work than the pupils taught by Method B by the same teachers, it is not unlikely that the difference in mean scholastic grades is due in part to this fact. If this hypothesis is correct, these two groups were more nearly equivalent than the mean semester grades indicate.

⁶³Douglass, *op. cit.*, p. 173-213.

all instructors involved in the investigation were asked to express an opinion in regard to the relative merit of the two instructional procedures. Nine of the fourteen opinions expressed were contrary to the experimental results in the pair of classes taught by the teacher giving the opinion. The conditions of this investigation differ in certain significant respects from the one described in the preceding paragraphs. In the first place, the Douglass experiment was carried on in the University High School at the University of Oregon. The other experiment was cooperative and involved about as many different schools as there were teachers. Another difference is that Douglass asked his teachers to express an opinion in regard to the results of the experiment, whereas in the other experiment the teachers were asked to indicate the method they preferred. Finally, the teachers in a University High School are likely to be more scientifically-minded than teachers in typical high schools and, hence, would be less likely to have strong preferences and more likely to be equally zealous in carrying out both of the methods.

The statement is made in the report of the experiment by Melby and Lien that, "The teacher, in fact, was secretly hoping that the results would reflect credit on the experimental method Yet results favored the control groups."⁶⁴ This should not be interpreted to mean that the data of this experiment are such as to minimize the importance of zeal as a factor. It is evident from the description of the procedures employed in the instruction of the control pupils that considerable zeal was exercised in spite of the teacher's dislike for these procedures. The inference that may be drawn from the report of this experiment is that this teacher was also sufficiently scientifically-minded to control the factor or zeal. It, therefore, appears that the Melby and Lien experiment and the one by Douglass are not necessarily in opposition to the one previously described. This conclusion is supported by evidence from other investigations.

The influence of some teacher factor, which probably was zeal, is revealed in the Newark Phonics Experiment. ". . . . the results show conclusively that there is immeasurably less difference between classes taught with and without phonics than between different schools. Where the results were unusually good in a class taught by a teacher using phonics, they were unusually good when the same teacher taught without phonics. On the other hand, poor results were secured in both phonic and non-phonic groups taught by the same teacher."⁶⁵

⁶⁴Melby and Lien, *op. cit.*, p. 264.

⁶⁵Sexton and Herron, *op. cit.*, p. 701.

In the experiment by Collings⁶⁶ the children taught by the project method achieved more than those taught by the traditional method, but it appears from Collings' report that these teachers worked much harder at their task than did the teachers in the control schools. In view of this fact, it does not appear justifiable to ascribe the superior achievement of the project-method group entirely to the method of instruction. The unusual zeal of the teachers undoubtedly contributed a large portion of the superiority in achievement. This conclusion is supported by an investigation reported by Gates.⁶⁷ The account of the experiment indicates that the teachers employing the "modern systematic method" exhibited as much zeal as those employing the "opportunistic method." The results favor the former. Although this experiment differs in several respects from the one conducted by Collings, they are sufficiently alike to justify the conclusion that the zeal of the teacher is a potent educative factor.

More direct evidence of the effect of a high degree of zeal upon achievement is furnished by an investigation by Pittman.⁶⁸ The description of the activities of the teachers in the experimental group of schools makes it apparent that they exhibited a very high degree of zeal. For example, it is stated: "The teachers under professional supervision did approximately four times as much professional reading as they themselves had done during the previous year, or as the unsupervised group, with which they were compared, did during the year of the experiment."⁶⁹ The description of this experiment would not be seriously distorted if the zeal of the teachers was designated as the experimental factor. Hence, the distinct superiority in achievement of the pupils in the experimental schools was undoubtedly due in a large measure, either directly or indirectly, to the zeal of the teachers.

After a relatively elaborate and careful study of the factors related to teaching success, Knight concluded that ". . . general factor of interest in one's work becomes the dominant factor in determining one's success in teaching . . . it is reasonable to suppose that genuine interest in one's work accounts for a large part of teaching success."⁷⁰

5. The term *personality traits* is used here as a name for a complex of subtle teacher factors that are commonly designated by such terms as "general appearance," "voice," "self-control," "tact," "sympathy," "sense of justice," and "loyalty." Such traits have not been defined

⁶⁶Collings, Ellsworth. *An Experiment with a Project Curriculum*. New York: The Macmillan Company, 1923. 346 p.

⁶⁷Gates, A. I., et al. "A Modern Systematic Versus an Opportunistic Method of Teaching," *Teachers College Record*. 27:679-700, April, 1926.

⁶⁸Pittman, M. S. *The Value of School Supervision*. Baltimore: Warwick and York, Inc., 1921. 129 p.

⁶⁹*Ibid.*, p. 101.

⁷⁰Knight, *op. cit.*, p. 9.

so that satisfactory measurement is possible, and, consequently, we do not have any definite measure of their effect upon pupil achievement. There is, however, a wide-spread conviction⁷¹ that the "personality"⁷² of the teacher is an important educative factor. This conviction is supported by some evidence. Morris has reported a partial coefficient of correlation of .463 between success in practice teaching and "trait index."⁷³ Hence, it seems safe to conclude that "personality traits" do affect pupil achievement to such an extent that they cannot be safely neglected in an experiment.

6, 7, and 8. The teacher's *physical condition, sex, and age* have an indirect influence on school achievement in so far as they condition the zeal with which a teacher employs instructional procedures and the skill with which he uses them. The teacher's physical condition, sex, and age may influence directly the achievement of children by engendering attitudes that may be beneficial, or detrimental, to learning. The relation of the teacher's physical condition to teaching efficiency is indicated in studies made of teacher failure. In a study by Buellesfield poor health takes twelfth rank as a chief cause and twentieth rank as a contributory cause.⁷⁴ Littler ranks poor health lowest of seven causes of teacher failure.⁷⁵ Moses places it eleventh in point of frequency; there are no less frequent causes mentioned.⁷⁶ In a recent study of teacher failure Madsen reports physical condition as the cause in only two out of thirty-one cases, in one case deafness and in the other case general physical disability.⁷⁷

Some correlation coefficients have been determined in an effort to indicate the relation of health or physical condition to teaching efficiency. Bradley states that the correlation between "general merit" and physical efficiency is .59, the lowest of several listed by him.⁷⁸

⁷¹The *Commonwealth Teacher-Training Study* referred to on page 31 affords conclusive evidence of this statement. For senior high-school teachers the ten traits considered most important are: breadth of interest, self-control, good judgment, leadership, scholarship (intellectual curiosity), forcefulness, honesty, adaptability, enthusiasm, and open-mindedness.

⁷²Although this term has not been defined with precision, as it is commonly used, it undoubtedly includes zeal (as the term has been used in the preceding pages) and probably overlaps with skill. Hence, the term "personality traits," as it is used here, is not synonymous with "personality."

⁷³Morris, E. H. "Personal Traits and Success in Teaching," *Teachers College, Columbia University Contributions to Education*, No. 342. New York: Bureau of Publications, Teachers College, Columbia University, 1929, p. 49.

This "trait index" is defined as a composite of likes and dislikes, resourcefulness and insight, tact, degree of positiveness of judgment, and characteristic feeling-attitudes. *Ibid.*, p. 18.

⁷⁴Buellesfield, Henry. "Causes of Failure Among Teachers," *Educational Administration and Supervision*, 1:451, September, 1915.

⁷⁵Littler, Sherman. "Causes of Failure Among Elementary-School Teachers," *School and Home Education*, 33:255-256, March, 1914.

⁷⁶Moses, C. V. "Why High-School Teachers Fail," *School and Home Education*, 33:166-169, January, 1914.

⁷⁷Madsen, I. N. "The Prediction of Teaching Success," *Educational Administration and Supervision*, 13:44-45, January, 1927.

⁷⁸Bradley, J. H. "A Study of the Relative Importance of the Qualities of a Teacher and Her Teaching in Their Relation to General Merit," *Educational Administration and Supervision*, 4:359, September, 1918.

Boyce gives a smaller correlation coefficient, .18, between health and general merit.⁷⁹ Ruediger and Strayer report the correlation to be .04 between general merit and health.⁸⁰ The recent study of Whitney gives a coefficient of .124 between physique and teaching success after graduation.⁸¹ However, it should be stated that this is a greater relationship than was found for intelligence and success after graduation. Whitney places physique as the fourth most important item in the prediction of success in teaching. It follows student teaching, professional marks, and academic marks. The weight of the evidence is in favor of regarding the physical condition of the teacher, so long as extremes are avoided, as a minor factor in the achievement of the pupils.

There is no evidence to be found in the literature which would indicate that the teacher's sex is an important factor in the learning activity of the pupils. It is said that the pre-adolescent boy prefers men teachers to women teachers, but it is yet to be proven that this prejudice, even assuming it to be universally existent, is sufficient to decrease his achievement significantly.

The age of the teacher is not usually a significant factor in successful teaching. After stating that the correlation of teaching skill with age was negligible for a group of Massachusetts teachers, Knight goes on to say:

We know there is *some* correlation between age in general and teaching ability. A five-year-old child could not teach, and excessive old age would no doubt be negatively correlated. But within those age limits during which men and women ordinarily teach, age does not appear to be correlated with teaching skill. The younger teachers are not the best as a current superstition would lead us to think; nor do years of tenure make material additions to skill.⁸²

This estimate of the importance of age as a factor in teaching is concurred in by Whitney who states, "Age is not a particularly important element in good teaching."⁸³ It seems safe to assume, therefore, that so long as extremes are avoided, the teacher's age is not a significant factor in the learning activity of the pupils.

The control of teacher factors. The preceding consideration of teacher factors has demonstrated the necessity of controlling, i. e. keeping the same, at least four teacher factors: (1) instructional procedures, (2) skill in carrying out these procedures, (3) zeal of the teacher, and (4) personality traits. Control of instructional techniques may be approached by giving the teachers participating in the

⁷⁹Boyce, A. C. "Qualities of Merit in Secondary School Teachers," *Journal of Educational Psychology*, 3:154, March, 1912.

⁸⁰Ruediger, W. C. and Strayer, G. D. "The Qualities of Merit in Teachers," *Journal of Educational Psychology*, 1:275, March, 1910.

⁸¹Whitney, F. L. "The Prediction of Teaching Success," *Journal of Educational Research Monographs*, No. 6. Bloomington, Illinois: Public School Publishing Company, 1924, p. 20.

⁸²Knight, F. B. "Qualities Related to Success in Elementary School Teaching," *Journal of Educational Research*, 5:212, March, 1922.

⁸³Whitney, *op. cit.*, p. 63.

experiment detailed instructions with regard to the conduct of their classes during the experiment. An example of such an attempt to control this factor is illustrated in the following quotation from the report of an experiment by Coryell:

In order to secure uniformity of procedure and the consistent carrying out of the prescribed methods, the teachers who were to collaborate in the experiment met from time to time in conference for devising ways and means. The class work for the first week was planned in the minutest detail. The same questions were actually used by all three teachers and the lesson plans were followed as exactly as was humanly possible while conducting a live recitation. Many other plans were made and used in common, and where no detailed lesson plan was drafted for all three teachers to use, the work to be covered each day was broadly outlined.⁸⁴

From one point of view, it is not sufficient to secure equivalence of instructional techniques. They should be representative of sound educational practice. This requires, among other things, that there be adaptation of techniques to the needs and purposes of the pupils as they are revealed in the course of the instruction. Hence, if control of instructional techniques is carried too far, the requirement of sound educational practice may be violated.

The control of skill and of zeal is much more difficult. A specified degree of skill or of zeal cannot be secured by asking teachers to follow certain instructions. Any direct request to be more skillful, and especially to be more zealous, may produce the opposite result. By exercising care in selecting teachers and by making use of indirect devices, a skillful experimenter may secure approximate equivalence of these two teacher factors, but, since neither can be measured objectively, he cannot be certain that he has done so.

A method frequently employed to control these factors is the rotation of teachers at the mid-point of the experimental period. The teacher who has been teaching the experimental group exchanges with one who has been teaching the control group.⁸⁵ Thus, any difference in skill or zeal on the part of two teachers is expected to be corrected by the fact that both the experimental and control pupils have received an equal amount of stimulation from both teachers. The experiment of Douglass is an example of the use of this technique.⁸⁶ This procedure will be successful in securing equivalence of these factors when each teacher is equally skillful in carrying out both forms of the experimental factor and is equally zealous in doing so. A teacher might teach with equal skill and zeal in employing two different techniques, but it appears likely that most teachers, because of a lack of

⁸⁴Coryell, N. G. "An Evaluation of Extensive and Intensive Teaching of Literature," *Teachers College, Columbia University Contributions to Education*, No. 275. New York: Bureau of Publications, Teachers College, Columbia University, 1927, p. 13.

⁸⁵This exchange involves also a change of teachers relative to the experimental factor.

⁸⁶Douglass, *op. cit.*, p. 187.

familiarity with, or a dislike for, one of the procedures, might teach with less skill and zeal in one of the groups than in the other. When this occurs, the rotation procedure will not succeed in securing control of skill and zeal except by chance.

Another plan for securing control of these factors is to have the same teacher teach an experimental and a control group. The success of this method will depend on the degree to which the teacher carries out both the experimental and control instructional procedures with equal skill and zeal. In order to approach control when a single teacher is used, or when two teachers exchange groups at the mid-point of the experiment, it might be suggested that teachers be practiced in the experimental procedures before the start of the experiment, and enough of the scientific attitude be engendered in them to overcome the operation of prejudice for any particular method. When a teacher develops a preference for one of the procedures, he becomes disqualified for the experiment.

Another difficulty is introduced when we recognize the requirement that the teaching represent sound educational practice. This requirement involves the provision that the teacher believe in the method he is employing rather than be indifferent or even open-minded toward it. In fact, sound educational practice probably requires that the teacher be prejudiced in favor of the method he is employing. If this is admitted, it is apparent that an experimenter should not expect to secure equivalence of skill and zeal by the rotation method or by having a teacher teach both an experimental group and a control group.

Since "personality traits" can not be measured satisfactorily, it is impossible to determine accurately the status of teachers with reference to this factor. Marked differences probably can be discovered, but in general it is not possible with our present techniques to select teachers who are equivalent with reference to "personality traits." Control may be secured by the rotation method or by having the same teacher instruct both a control group and an experimental group. Although these procedures will secure control of "personality traits," they are not completely satisfactory for reasons pointed out in the preceding paragraphs.

II. General school factors that affect pupil achievement. Pupil achievement is affected directly or indirectly by several general school factors. For example, it is generally assumed that the textbook used in a course influences the achievement of the pupils. Much of this influence is indirect. The character of the text influences the learn-

ing exercises assigned which in turn influence achievement. In the following list of general school factors no attempt is made to indicate whether a factor functions directly or indirectly.

1. Instructional materials (textbooks, library, maps, laboratory apparatus, etc.)
2. Time devoted to learning activity
3. Characteristics of the class as a group
4. Size of class
5. Size of school
6. School organization
7. Administration and supervision
8. School building, especially lighting, heating, and ventilation

1. *Instructional materials*, such as textbooks, libraries, and other school equipment, influence the learning activity of pupils through the learning exercises that they furnish or make possible. Texts in arithmetic, algebra, language, physics, and most of the other subjects furnish a number of learning exercises. Texts and other books make possible other learning exercises, such as requests to study certain pages or questions whose answers may be found by reading. In a similar way charts, maps, moving picture machines, laboratory apparatus, and the like affect the number and type of learning exercises that may be assigned. Hence, the achievement of the pupils is likely to be affected by the instructional materials used with a class.

The intimate relation between instructional materials and learning exercises may make it impossible to have the former constant when the latter are greatly different. It should be noted, however, that certain types of learning exercises require certain instructional materials. Hence, if the purpose of an experiment is to compare two types of learning exercises, such as the lecture-demonstration method of teaching a science and individual-laboratory work, the materials must differ. In such cases, the difference in instructional materials is essentially a phase of the experimental factor.

2. If the *time devoted to learning activity* is assumed to be an approximate index of the amount of exercise of modifiable connections, it is apparently an important educative factor. In considering its importance in experimental investigations two cases should be noted: (1) the long-time experiment in which the difference in time spent is due to absences of certain pupils, and (2) the experiment in which the difference in time spent is a difference in the length of the class period or in the total time devoted to study.

In a study reported by Odell, a slight positive correlation was found between average school marks and per cent of time in attendance.⁸⁷ The relationship of length of attendance to educational age is shown in the coefficient of correlation of $.30 \pm 2$ reported by Denworth.⁸⁸ In view of these relatively low correlations it would seem plausible to say that if there were no extreme cases of inattendance, or irregular attendance, and if the absences were approximately balanced, attendance would be an insignificant factor.

In the second case, it seems reasonable to expect that the time spent in learning activity is an important educative factor. Experimentation on the distribution of practice in learning has shown that there are optimum lengths of practice periods for different types of learning. Pyle⁸⁹ in his substitution experiment used fifteen-, thirty-, forty-five-, and sixty-minute practice periods. His results were in favor of the thirty-minute period. The experiments of Hahn and Thorndike,⁹⁰ Kirby,⁹¹ Starch,⁹² and Lyon⁹³ indicate that the length of the practice period is a factor in learning. The evidence just cited tends to show that more time spent in learning activity does not necessarily imply more learning. Up to a certain point, increase of the learning period may be beneficial to learning; beyond this point, increase may be detrimental. It is probable that this sort of thing operated in the investigations of Rice,⁹⁴ Heck,⁹⁵ Jones and Ruch,⁹⁶ and Barnes and Douglass⁹⁷ who found little relation between time spent in learning activity and achievement. It should be noted that "time spent in learning activity" needs definition in this connection. Thinking and talking about an assignment probably should be included as well as formal study, either at school or at home. If this thesis is

⁸⁷Odell, C. W. "The Effect of Attendance Upon School Achievement," *Journal of Educational Research*, 8:422-32, December, 1923.

⁸⁸Denworth, K. M. "The Effect of Length of School Attendance Upon Mental and Educational Ages," *Twenty-Seventh Yearbook of the National Society for the Study of Education*, Part II. Bloomington, Illinois: Public School Publishing Company, 1928, p. 80.

⁸⁹Pyle, W. H. *The Psychology of Learning*. Baltimore: Warwick and York, Inc., 1928, p. 44. (Revised edition.)

⁹⁰Hahn, H. H. and Thorndike, E. L. "Some Results of Practice in Addition under School Conditions," *Journal of Educational Psychology*, 5:65-84, February, 1914.

⁹¹Kirby, T. J. "Practice in the Case of School Children," *Teachers College, Columbia University Contributions to Education*, No. 58. New York: Bureau of Publications, Teachers College, Columbia University, 1913. 98 p.

⁹²Starch, Daniel. "Periods of Work in Learning," *Journal of Educational Psychology*, 3:209-213, April, 1912.

⁹³Lyon, D. O. "The Relation of the Length of Material to the Time Taken for Learning and the Optimum Distribution of Time," *Journal of Educational Psychology*, 5:1-9, 85-91, 155-163, January, February, and March, 1914.

⁹⁴Rice, J. M. "The Futility of the Spelling Grind," *Forum*, 23:163-72, 409-19, April, June, 1897.

⁹⁵Heck, W. H. "Correlation Between Amounts of Home Study and Class Marks," *School Review*, 24:533-549, September, 1916.

⁹⁶Jones, Lonzo and Ruch, G. M. "Achievement as Affected by Amount of Time Spent in Study," *Twenty-Seventh Yearbook of the National Society for the Study of Education*, Part II. Bloomington, Illinois: Public School Publishing Company, 1928, p. 131-134.

⁹⁷Barnes, D. G. and Douglass, H. R. "The Value of Extra Quiz Sections in the Teaching of History," *University of Oregon Publication*, Education Series, Vol. 1, No. 7. Eugene: University of Oregon, 1929, p. 276-284.

accepted, the control of the time spent in learning activity frequently will be difficult. The evidence cited from the experimentation on distribution of practice is sufficient to warrant the assertion that the experimenter should use whatever means are available to secure, so far as possible, an *equal* amount of time each day to be spent in learning activity, both in recitation and study, by the experimental and control pupils.

3. The phrase *characteristics of a class as a group* is used to designate a factor that is difficult to define. It includes what is commonly called *esprit de corps*. It does not include general intelligence and the other factors listed on pages 19 and 20 since the two groups are expected to be equivalent in these respects. Rivalry among certain members of a class may stimulate the entire group to greater effort. Because the pupils like each other or because of outside associations a teacher may prefer to instruct one class rather than another and, hence, may exhibit unusual zeal in his work. On the other hand, the members of a class may not like each other. There may even be petty jealousies and enmities that make the teacher's task unusually difficult. The characteristics of the class as a group constitute a very intangible factor that operates in subtle ways. It is, however, of sufficient importance to warrant the consideration of the careful experimenter.

4. The *size of the class* disappears as an educative factor in an experiment where equivalent groups are secured by pairing, since this procedure secures classes of equal size. If the two groups are not equal in size, small differences do not appear to be significant because within fairly wide limits size of class does not appear to be an important educative factor.⁹⁸

Incidentally, it may be noted that when generalizing from an experiment with classes of a given size, the conclusions may be expected to be applicable to classes of other sizes within a considerable range, provided the size of the class does not affect other educative factors.

5. The *size of the school* indirectly affects the achievement of its pupils. Larger schools tend to possess superior organizations, better qualified administrative, supervisory, and instructional staffs, and a

⁹⁸Breed, F. S. and McCarthy, G. D. "Size of Class and Efficiency of Teaching," *School and Society*, 4:965-971, December 23, 1916.

Edmonson, J. B. and Mulder, F. J. "Size of Class as a Factor in University Instruction," *Journal of Educational Research*, 9:1-12, January, 1924.

Hudelson, Earl. *Class Size at the College Level*. Minneapolis: The University of Minnesota Press, 1928. 300 p.

Stevenson, P. R. "Class-Size in the Elementary School," *Bureau of Educational Research Monograph*, No. 3. Columbus: Ohio State University, 1925. 35 p.

Stevenson, P. R. "Smaller Classes or Larger: A Study of the Relation of Class-Size to the Efficiency of Teaching," *Journal of Educational Research Monographs*, No. 4. Bloomington, Illinois: Public School Publishing Company, 1923. 127 p.

Bureau of Educational Research. "Relation of Size of Class to School Efficiency." *University of Illinois Bulletin*, Vol. 19, No. 45, Bureau of Educational Research Bulletin No. 10. Urbana: University of Illinois, 1922, p. 39.

greater diversity of school equipment and, hence, probably provide a better environment for learning. Inferiority in these things has caused Ruffi⁹⁹ to question the efficiency of the small high school. However, in spite of the diversity between small and large schools in these things, it is yet to be proven that school size is anything more than a minor factor in the achievement of the pupils. Gowen and Gooch¹⁰⁰ compared the average college grades of students from large high schools with those of students from small high schools and failed to find a significant difference. Size of school, in itself, does not seem to be anything but a very minor factor in learning activity. As long as small size does not mean different organization, less qualified administrators and teachers, or lack of the materials of instruction prescribed in the experiment, it may be neglected by the experimenter even when the two groups compared are in different schools.

6. It seems reasonable that the *organization of the school* is important enough to be considered when experimental and control groups are in different schools. For example, schools in which there is individual instruction, ability grouping, or a platoon system are not appropriate environments for experimental groups unless the control groups are subject to the same conditions. To illustrate the importance of school organization, the conclusion of an investigation in which the achievement of equivalent rural- and city-school children was compared may be given:

The results of this study indicate that the progress of graded-school pupils was approximately one-half school year in advance of that of the pupils with whom they were paired from the rural schools.¹⁰¹

Since the pupils were equivalent in intelligence and chronological age, the difference in achievement must be ascribed, at least in part, to the superior organization of the graded city schools.

7. The *administration and supervision* of a school must be an important factor in learning activity if the attention given to these fields in teacher-training institutions is any criterion. However, it is difficult, if not impossible, to find any quantitative evidence in regard to the contribution of this factor to classroom learning. The reason for this seems to lie in the fact that any influence exerted by administration or supervision must be an indirect one operating through the teacher, the course of study, the organization of classes,

⁹⁹Ruffi, John. "The Small High School." *Teachers College, Columbia University Contributions to Education*, No. 236. New York: Bureau of Publications, Teachers College, Columbia University, 1926, p. 141.

¹⁰⁰Gowen, J. W. and Gooch, Marjorie. "The Mental Attainments of College Students in Relation to the Preparatory School and Heredity," *Journal of Educational Psychology*, 17:408-418, September, 1926.

¹⁰¹Stone, C. W. and Curtis, J. W. "Progress of Equivalent One-Room and Graded-School Pupils," *Journal of Educational Research*, 16:264, November, 1927.

the provision of school equipment, and so on. It seems logical to assume that the educational experimenter who has controlled these more direct factors will have taken care of administration and supervision.

8. The *school building* is a factor in school achievement in that it provides an environment that may be beneficial or detrimental to learning. While there is no experimental evidence, it seems logical to assume that learning takes place more readily in the beautiful and appropriate school buildings now in existence than in the ugly and inconvenient structures of a generation ago. The importance of lighting is recognized in the weight given to it in building score cards.¹⁰² While the chemical composition of the air is usually so constant as to be unimportant, its temperature, humidity, and movement spell comfort or discomfort to pupils in a classroom and through this influence achievement.¹⁰³ Two investigations in which Thorndike has participated minimize these factors of ventilation.¹⁰⁴ However, lest the inference be made that ventilation is a wholly unimportant factor because of these findings, Sandiford has made the following comment:

Apparently we can, if we will, work as hard under adverse conditions of heat and humidity as under favorable ones. Even summer school in New York or Timbuctoo need not daunt us! What should be noted, however, is that these distressing conditions are uncomfortable, and if we subject children to them the likelihood is that their attention will be distracted from work.¹⁰⁵

It may be stated that light, heat, and ventilation become important factors to the educational experimenter only when grossly abnormal conditions prevail. When such conditions are avoided, they probably are not factors of sufficient importance to warrant the attention of the experimenter.

The control of general school factors. When the two groups of pupils are within the same school, the most significant general school factors appear to be (1) instructional materials, and (2) time devoted to learning activity. The control of instructional materials as a non-experimental factor is accomplished by securing an identity

¹⁰²For example see:

Strayer, G. D. Score Card for City School Buildings. *Teachers College Bulletin*, Seventh Series, No. 12. New York: Teachers College, Columbia University, 1916, p. 6.

¹⁰³Burnham, W. H. "The Optimum Temperature for Mental Work," *Pedagogical Seminary*, 24:69, March, 1917.

Burnham, W. H. "The Optimum Humidity for Mental Work," *Pedagogical Seminary*, 26:328, December, 1919.

McLure, J. R. "The Ventilation of School Buildings," *Teachers College, Columbia University Contributions to Education*, No. 157. New York: Bureau of Publications, Teachers College, Columbia University, 1924, p. 109.

¹⁰⁴Thorndike, E. L., McCall, W. A., and Chapman, J. C. "Ventilation in Relation to Mental Work," *Teachers College, Columbia University Contributions to Education*, No. 78. New York: Bureau of Publications, Teachers College, Columbia University, 1916. 33 p.

Thorndike, E. L. and Kruse, P. J. "The Effect of Humidification of a Room Upon the Intellectual Progress of the Pupils," *School and Society*, 5:657-660, June 2, 1917.

¹⁰⁵Sandiford, Peter. *Educational Psychology*. New York: Longmans, Green and Company, 1929, p. 271-72.

of instructional materials for both the experimental and control groups. Reeder¹⁰⁶ secured such identity by using the same textbook in both groups and permitting access to the textbook only during the class period. It goes without saying that pupils should not only have the same opportunities with respect to a textbook, but should have equal access to supplementary material, such as reference books, maps, charts, and museums, as well. This statement, of course, applies only to those instructional materials that are not involved in the experimental factor.

Securing equivalence of time spent in learning activity demands that the length of the class period and the number of periods spent on the experimental learning be the same for both groups. It necessitates also that the experimental and control pupils spend an equal amount of time in study whether in school or at home. This is probably best accomplished by having all the pupils study the experimental learning for the same length of time under the supervision of the study-hall director, or possibly their classroom teacher, or teachers. If the experimental and control pupils are to engage in the experimental learning at home, all of them should do so. The control of the time factor in this case may be approached by securing the cooperation of the parents of the children. Equivalence with respect to the time factor also demands that at the close of the experiment, the experimenter should check the amount of and regularity of attendance of the pupils in the experimental and the control groups. When the attendance of either pupil of a pair is grossly deficient or irregular, the pair probably should be discarded.

The characteristics of the class as a group should not be neglected, but little can be done to control this factor. If equivalent groups are secured and the teacher factors are adequately controlled, it is likely that the two classes will not differ greatly in their characteristics as a group. For example, if the pupils are equivalent with respect to intelligence and previous achievement, both the experimental and control pupils will have the same advantages in profiting from the recitations of their fellows. If the teacher, or teachers, instruct the two groups with equal skill and zeal, similar group attitudes should be engendered. The experimenter, however, should endeavor to evaluate the two classes with respect to their group characteristics, and, if differences are apparent, the fact should be recognized in interpreting the results of the experiment.

¹⁰⁶Reeder, E. H. "A Method of Directing Children's Study of Geography," *Teachers College, Columbia University Contributions to Education*, No. 193. New York: Bureau of Publications, Teachers College, Columbia University, 1925. 98 p.

The size of the class is automatically controlled as a factor when classes are formed by pairing pupils. The presence of pupils not actually included in the experiment, as is sometimes the case when groups are selected without interfering with the composition of regular school classes, will not interfere with the experiment unless the number of such pupils is large. Where a number of paired groups are used, it is probable that none of the pairs should differ greatly in size if the results are to be combined or compared.

The size of the school does not influence one group more than the other, if both are within that school. When experimental and control groups are to be in different schools, the experimenter should select schools that are approximately the same size. In the control of this and other general school factors in cooperative experimentation where several schools participate, a measure of control is attained by having each experimental *group* paired with a control *group* in the same school. The cooperative experiment of Breed in which fourteen schools cooperated is an example of this.¹⁰⁷

III. **Extra-school factors that affect pupil achievement.** Pupil achievement is affected by several factors that have not been included in the preceding lists. The following appear to deserve consideration:

1. Participation in extra-curricular activities
2. The pupil's home life
3. Community interest in and attitude toward the school

1. Carefully supervised *participation in extra-curricular activities* tends to be beneficial rather than detrimental to learning.¹⁰⁸ The pupil who engages in some such activity frequently becomes more interested in his regular school work. Dramatic, scientific, technical, literary, and debating clubs not only add interest to the school subjects to which they are related but they also may contribute directly to achievement. It is likely, however, that for each child, there is an optimum amount of participation above which his school achievement will suffer.

2. The *child's home life* may influence his school achievement in many ways. Studying at home under parental supervision and with parental sympathy, listening to conversation of parents and other members of the family, reading periodicals and books that the home

¹⁰⁷Breed, F. S. "Measured Results of Supervised Study," *School Review*, 27:186-204, 262-84, March, April, 1919.

¹⁰⁸This contention is supported in the following reports of research:

Tremper, G. N. "The Effect of Participation in Extra-Curricular Activities on the Scholarship of the Participants in the Kenosha, Wisconsin, Senior High School." A thesis submitted for the degree of Master of Arts in Education. Urbana: University of Illinois, 1923. 63 p.

Crawford, C. E. "The Effect of Participation in Extra-Class-Room Activities on Scholarship of High School Pupils." A thesis submitted for the degree of Master of Arts in Education. Urbana: University of Illinois, 1929. 64 p.

affords, traveling with members of the family, and the like are activities that sometimes make large contributions in the fields of school achievement. The following quotations indicate that the parents of children have been regarded as an important factor in school achievement, particularly achievement that results from home study:

Home environment is a factor in the formation of study habits. Its influence may be either for good or for bad. . . .

Home study is desirable because it acts as a check on the formation of habits out of school that would be negative in their influence on habits in school.¹⁰⁹

The survey of the fourth, fifth, and sixth grades seems to justify this conclusion: Where the parents are capable of guiding the child and are inclined to supervise the home study, their children succeed in school. But where parents are illiterate or for other reasons are unable or unwilling to supervise the home study, their children as a rule either make slow progress or are failures entirely when measured by the progress of their companions in school.¹¹⁰

It is probable that the conclusion of Brooks exaggerates the importance of supervision of home study by parents since Heck¹¹¹ has presented data to show that it is immaterial whether students study at home or at school. The inconclusiveness of the research on supervised study would lead one to question the value of the type of supervision most parents are capable of giving. It is possible that the attitude parents take toward the school as an educative agency is a more potent influence than any supervision they may administer. For example, Hurlock¹¹² has shown that praise of pupils by teachers is much more beneficial to achievement than reproof or indifference. It is probable that the same is true of praise or reproof on the part of parents relative to the school work of their children. Reavis¹¹³ has described several interesting cases in which failure in school achievement was due to detrimental parental attitudes whose correction, effected by enlisting the cooperation of the parents, resulted in the change from failure to success.

Listening to conversation of parents and other members of the family, reading periodicals and books that the home affords, and traveling with members of the family are activities that may contribute to the experimental learning. These experiences help provide the background of information for the learning that is to take

¹⁰⁹Reavis, W. C. "Some Factors That Determine the Habits of Study of Grade Pupils," *Elementary School Teacher*, 12:81, October, 1911.

¹¹⁰Brooks, E. C. "The Value of Home Study Under Parental Supervision," *Elementary School Journal*, 17:193, November, 1916.

¹¹¹Heck, W. H. "Comparative Tests of Home Work and School Work," *Journal of Educational Psychology*, 10:153-62, March, 1919.

¹¹²Hurlock, E. B. "An Evaluation of Certain Incentives Used in School Work," *Journal of Educational Psychology*, 16:145-59, March, 1925.

¹¹³Reavis, W. C. "Constructive Student Accounting in the Secondary School, A. Administering the Maladjusted Student," *Supplementary Educational Monographs*, No. 24. Chicago: University of Chicago, 1923, p. 20-33.

place during the experiment. Topics in history, civics, biology, literature, and economics are more meaningful to the pupil who has had related experiences through conversation with members of his family, or through travel. It is impossible to estimate the extent to which school achievement is influenced by these out-of-school experiences.

Several recent studies have minimized the importance of home environment with respect to school achievement. For example, Heilman's statement that "57% of the variation in educational age was due to mental age or such hereditary factors as had been measured; about 7% of the variation was due to the influences of school training and socio-economic status combined, or such environmental factors as had not been measured"¹¹⁴ would lead one to believe that home environment is not a significant factor. However, it is yet to be proven that the information acquired and the ideals and attitudes engendered in the home influence school achievement, as represented in the experimental learning, so little that the experimenter is justified in neglecting this factor.

3. School achievement is influenced by *community interest in and attitude toward the school*. If the community is high in the socio-economic scale, the members of the community are likely to show much interest in school affairs and to cooperate with the principal and teachers in attaining the best conditions for school work. For example, the parents of such a community may cooperate with the school faculty in providing more adequate library facilities. In other cases the community may be permeated with attitudes antagonistic toward the school administration. Such attitudes among parents tend to be acquired by pupils. Thus, community attitudes and interest may exert a subtle though powerful influence on school learning.

The control of extra-school factors. The participation in extra-curricular activities should be checked by means of information secured from teachers, school records, or the pupils themselves. The experimenter probably has controlled this factor satisfactorily when he has insured that there is no great excess of participating students in either group. This factor appears to be a minor one; therefore small differences in participation may be neglected.

The relationships found to exist between the intelligence of children as measured by typical intelligence tests and parental occupation and social status would lead one to believe that pairing children on the basis of intelligence scores helps to secure equivalence with respect

¹¹⁴Heilman, J. D. "Factors Determining Achievement and Grade Location," *The Pedagogical Seminary and Journal of Genetic Psychology*, 36:453, September, 1929.

to the home life of the children used in the experiment.¹¹⁵ Control of this factor is also aided by securing equivalence with respect to previous achievement. If the initial achievement tests are valid and reliable with respect to the experimental learning, and if the groups are equivalent with respect to the mean scores on the initial achievement test, then it seems probable that the groups will be equivalent so far as influence from information obtained out of school is concerned. The inference should not be made that the child's home life is considered an insignificant factor, or that securing of equivalence with respect to intelligence is all that needs to be done. The experimenter should be on the alert to detect cases in which abnormal home environment, particularly detrimental parental attitudes, are handicapping the learning of individual children in the groups.

The factor of community interest in and attitude toward the school will not usually demand attention if the experiment is confined to one school, or if there is a pair of groups in each one of several schools. If the experimental and control groups are in different schools, in different communities, this factor should receive attention. In such cases, however, it seems probable that pairing pupils on the basis of intelligence will do much to insure control of this factor, since it is likely that community interests and attitudes tend to vary with the intellectual level of the children. It is probably desirable that the presence of the experiment should not be given too much publicity, lest the parents and the other members of community take an unwelcome interest in the experimental or control procedures. It is desirable, for the sake of generalization, that the communities in which experiments are conducted be typical of communities to which the results are to be applied.

Summary with reference to control of educative factors. In the light of the preceding discussion and of practical considerations, it appears that equivalence¹¹⁶ of at least the following educative fac-

¹¹⁵Book, W. F. *The Intelligence of High-School Seniors*. New York: The Macmillan Company, 1922. 371 p.

Bridges, J. W. and Coler, L. E. "The Relation of Intelligence to Social Status," *Psychological Review*, 24:1-31, January, 1917.

Dexter, E. S. "The Relation Between Occupation of Parent and Intelligence of Children," *School and Society*, 17:612-14, June 2, 1923.

English, H. B. "An Experimental Study of Mental Capacities of School Children, Correlated With Social Status," (Yale Psychological Studies) *Psychological Monographs*, 23:266-331, 1917.

Haggerty, M. E. and Nash, H. B. "Mental Capacity of Children and Paternal Occupation," *Journal of Educational Psychology*, 15:559-72, December, 1924.

Pressey, S. L. and Ralston, Ruth. "The Relation of the General Intelligence of School Children to the Occupation of Their Fathers," *Journal of Applied Psychology*, 3:366-73, December, 1919.

Terman, L. M., et al. "Racial and Social Origin," *Genetic Studies of Genius*, Vol. 1. Stanford University, California: Stanford University Press, 1925, p. 55-83.

¹¹⁶Control of factors by securing equivalence means that conditions have been so arranged by the experimenter that the factors operate equally in both the experimental and the control groups.

tors must be secured, or appropriate allowance for non-equivalence must be made in interpreting the difference in gains.

1. Instructional techniques
2. Skill in carrying out instructional techniques
3. Zeal of the teacher
4. Personality traits of teachers
5. Instructional materials
6. Time spent in learning activity

It should be noted that frequently the experimental factor is a phase of instructional techniques. When this is the case, the requirement of equivalence applies only to the remaining phases. A similar comment applies to instructional materials.

Control of instructional techniques and of instructional materials can be secured by careful planning and by giving attention to details during the experiment. When the learning is restricted to the classroom, the control of time spent and of the other environmental factors is easily secured, but when the learning involves home study, satisfactory control is more difficult. The greatest difficulty is in securing satisfactory control of skill, zeal, and personality traits of teachers.

In addition to controlling the factors enumerated, the other educative factors should be investigated to make certain that no marked differences exist. If there are significant differences, either the experiment should be organized so that they neutralize each other, or their possible influence must be estimated and corrected for in the interpretation of the data.¹¹⁷

The problem of controlling educative factors in different types of experimentation. 1. *Single group experiments.* Since it is impossible to equate non-experimental factors in a single group experiment, control of factors must depend on estimations of their influences on the experimental learning. The effect due to the application of the experimental factor must be singled out from the effect due to all the other factors. This is usually done by comparing the achievement of the pupils under the influence of the experimental factor with their achievement prior to the application of the factor. It is obvious that this procedure will not often secure dependable quantitative results. In addition to the improvement that must be ascribed to the change that has occurred in the intellectual and educational status of the pupils, some of the improvement may be due to the less difficult instructional material or to the greater zeal and effort shown by the teacher because of the novelty of the experimental method, or both.

¹¹⁷Factors not equated may be said to be controlled when their variation is determined and the effect recognized in the difference in gains.

The single group experiment is a desirable activity for the classroom teacher to engage in since it is likely to be stimulating, but in general it cannot be expected to result in dependable answers to educational problems.

2. *Experiments in which two equivalent groups are taught by the same teacher.* The control of the non-experimental factors in experiments in which two equivalent groups are taught by the same teacher is dependent on the degree to which conditions are arranged so that these factors operate equally in both groups. The procedures to be used for securing such equivalence of factors have been suggested in the preceding pages. Since this type of experiment is conducted by a single teacher in one school, personality traits, sex, age, and physical condition of the teacher, size of school, school organization, administration and supervision, school building, and community attitude and interest in the school do not need to receive the attention of the experimenter, since these are the same for both groups.¹¹⁸ It is evident, however, that the other factors listed may be of unequal influence on achievement unless conditions are arranged with care.

After equivalent groups of pupils have been secured, control of the teacher factors of skill and zeal is very important. The teacher should be equally familiar with the instructional procedures and materials used in the experimental group and with the instructional procedures and materials used in the control group. The attitudes of the teacher with respect to these instructional procedures and materials should be such that the teaching is done with equal zeal in both groups. In addition, the teacher must exercise constant care throughout the experiment in order to maintain an identity of classroom-management procedures and time spent in learning activity. The teacher must be able to adapt herself readily when teaching one group immediately after the other.

The rotation technique is often employed to secure control of pupil factors when the groups are only approximately equivalent.¹¹⁹ The instructional procedures and materials of the experimental and control groups are exchanged at the mid-point of the experiment. In computing the results, the gain credited to the experimental factor is the sum of the gains of both groups while under its influence. The gain credited to the control procedures is the sum of the gains of

¹¹⁸This statement applies only to the control of factors so that the difference may be a significant difference for the groups concerned. If the results are to form the basis of generalizations, these factors must be typical of the schools to which the generalizations are to be applied.

¹¹⁹If Group A acts first as experimental and second as the control group, while Group B acts first as control and second as experimental group, two hypothetically equivalent groups are secured. Group A (as experimental) plus Group B (as experimental) is equivalent to Group A (as control) plus Group B (as control). In other words, the pupils are equivalent to themselves.

both groups while acting in the capacity of controls. The use of this technique, however, may introduce errors of more significance than those it would seek to eliminate. The group that first receives the benefit of the experimental factor is likely to acquire abilities, such as study habits, that will carry over and function when the group is acting as control. What is likely to happen is shown in the following illustrations in which the true gain is assumed to be 8 when the experimental instructional procedures are used. The true gain is assumed to be 4 when the control procedures are used. The experimental instructional procedure is labeled Method X, and the control procedure, Method Y. Then for the hypothetical ideal situation the gains are as follows:

<i>Gain</i>			
Group A	8	with Method X	
Group B	4	with Method Y	
Group B	8	with Method X	(after rotation)
Group A	4	with Method Y	(after rotation)
Difference = $(8 + 8) - (4 + 4) = 8$ in favor of Method X			

Assume that the carry over of study habits by Group A introduces an error of 3:

<i>Gain</i>			
Group A	8	with Method X	
Group B	4	with Method Y	
Group B	8	with Method X	
Group A	7	$(4 + 3)$	with Method Y plus study habits
Difference = $(8 + 8) - (4 + 7) = 5$ in favor of Method X			

The effect of this error plus the effects of others combining with it in unknown ways may be sufficient to destroy the significance of results, especially if the computed difference in gains happens to be small. If the teacher varies in skill or zeal, the use of the technique of rotating pupils is not likely to eliminate the errors created. Let us assume that the teacher prefers the experimental factor to the extent that the error introduced is equal to one-half the influence due to the experimental factor, or Method X, and at the same time let us assume that her dislike for the control procedures, or Method Y, is also sufficient to cause an error equal to one-half of the influence due to the Method Y.

<i>Gain</i>			
Group A	12	$(8 + 4)$	with Method X plus preference
Group B	2	$(4 - 2)$	with Method Y plus dislike
Group B	12	$(8 + 4)$	with Method X plus preference
Group A	2	$(4 - 2)$	with Method Y plus dislike
Difference = $(12 + 12) - (2 + 2) = 20$ in favor of Method X			

For the sake of comparison let us assume that instead of preferring the experimental instructional procedure, Method X, the

teacher dislikes it and prefers the control procedure, Method Y. Further, let us assume that the dislike of Method X removes half of its effectiveness and the preference for Method Y doubles its effectiveness. Then:

		<i>Gain</i>	
Group A	4	(8 - 4)	with Method X plus dislike
Group B	6	(4 + 2)	with Method Y plus preference
Group B	4	(8 - 4)	with Method X plus dislike
Group A	6	(4 + 2)	with Method Y plus preference
Difference =		(4 + 4) - (6 + 6)	= -4 in favor of Method Y

Thus the failure of the rotation technique to control the teacher factor may result in exaggerating the influence of this experimental instructional procedure, or it may result in creating an apparent difference in favor of the really less desirable control instructional procedures. When it is remembered that failure to control the teacher factor may be accompanied by error due to carry over of abilities, it will be recognized that the rotation technique does not insure dependable results.

3. *Experiments in which two equivalent groups are taught by different teachers in the same school.* The control of non-experimental factors when equivalent groups are taught by different teachers in the same school is very similar to the control of factors when both groups are taught by the same teacher. There is no need to give attention to such factors as size of school, school organization, administration and supervision, school building, and community attitude and interest, since these are the same for both groups.¹²⁰ The fact that different teachers are used increases the importance of the teacher factors. In order that both teachers will be equal in their influence on achievement, irrespective of the experimental factor, they must teach with equal "skill" and "zeal" in carrying out instructional techniques and classroom-management procedures. The experimenter may seek to secure equality of these teacher factors by selecting teachers who have approximately the same intelligence, training, and experience, and who are not widely different in age or physical condition. After teachers have been selected on the basis of equality, or similarity, in the above characteristics, more adequate control of skill and zeal may be attempted by practicing the teachers in the instructional procedures and materials of the experiment. In doing this, the experimenter should be especially careful to engender scientific attitudes toward the instructional procedures and materials in an

¹²⁰A partial exception should be noted with reference to the last two factors. School-rooms within a building vary and care should be exercised to insure that the rooms in the experiment as being carried on are not significantly different. Care should also be exercised to insure that the community attitude toward the experiment is neutral.

effort to minimize the influence of teacher preferences, or dislikes, for methods or materials. Finally, the use of detailed lesson plans in both groups during the experiment should be effective as an aid in the control of the teacher factors.

“Personality traits” must receive attention; however, it is difficult to select teachers who are equivalent with respect to this factor. A principal or other supervisor who is intimately acquainted with the teachers of a school may select two teachers who are approximately equivalent, but since we have no satisfactory means of measuring this fact, the degree of equivalence cannot be determined.

The rotation of teachers is a technique frequently used to control the teacher factors. At the mid-point of the experiment the teachers exchange groups and procedures. Group A with Method X continues with Method X but with the new teacher. Group B with Method Y continues with Method Y but with a new teacher.¹²¹ It is probable that the use of this technique eliminates such lesser teacher factors as age, sex, physical condition, and personal idiosyncrasies. It does not seem likely, however, that rotation of teachers adds anything to the control of the important factors, skill and zeal. For example, let 8 be the gain due to Method X for one-half the experimental period, and let 4 be the gain due to Method Y for one-half of the experimental period. Then for the ideal case the gains are as follows:

Gain

Group A	8	with Method X
Group B	4	with Method Y
Group B	4	with Method Y (continued)
Group A	8	with Method X (continued)
Difference = $(8 + 8) - (4 + 4) = 8$ the “true” difference due to the superiority of Method X.		

Now let us assume that both teachers are unskilled in the use of Method X so that half of its effectiveness is lost.

Gain

Group A	4	$(8 - 4)$ with Method X plus lack of skill
Group B	4	with Method Y
Group B	4	with Method Y
Group A	4	$(8 - 4)$ with Method X plus lack of skill
Difference = $(4 + 4) - (4 + 4) = 0$		

Again, let us assume that both teachers are equally more zealous for the experimental factor, although they are equally skilled in the use of both methods.

¹²¹Both teachers and pupils might be rotated, in which case the hazards described in the previous discussion of rotation of pupils would be accompanied by those described in the present discussion.

<i>Gain</i>			
Group A	12	(8 + 4)	with Method X plus zeal
Group B	4		with Method Y
Group B	4		with Method Y
Group A	12	(8 + 4)	with Method X plus zeal
Difference = (12 + 12) - (4 + 4) = 16			

Finally, let us assume that, although the two teachers are equally skilled in both methods, one is zealous for the experimental factor, while the other is equally prejudiced against it.

<i>Gain</i>			
Group A	12	(8 + 4)	with Method X plus zeal
Group B	4		with Method Y
Group B	4		with Method Y
Group A	4	(8 - 4)	with Method X plus prejudice
Difference = (12 + 4) - (4 + 4) = 8. This happens to be the true difference, but notice the conditions necessary for these two factors to eliminate themselves.			

Thus the rotation of teachers may fail to eliminate the error due to the teacher factors—skill in the use of instructional procedures and zeal of the teacher with reference to the experimental factor. The rotation technique, whether of pupils, or teachers, or both, is of doubtful desirability since its use does not give more certain control than when rotation is not used. It is a dangerous technique to employ in any form, since it may engender a false idea that by its use non-experimental factors are controlled, and because rotation of pupils or teachers, except at the end of a term or semester, creates an abnormal situation.

4. *Experiments in which two equivalent groups are taught by different teachers in different schools.* When the experimental group is in one school and the control group is in another under a different teacher, the general school and extra-school factors become significant.¹²² The fact that different schools are used introduces possible differences in instructional materials, size of school, time spent in learning activity (class periods and study periods), school organizations, school administration and supervision, school buildings, community interests in and attitudes towards the schools, children's home lives, attitudes of homes toward the schools, home facilities for study, home duties performed by pupils, and the participation in extra-curricular activities. The most effective means of securing control of these factors rests in selecting schools that appear to be as much alike as possible.¹²³ In other words, schools should be selected that

¹²²The control of pupil and teacher factors is no less important than when a single school is used.

¹²³If the results are to serve later as the basis of generalization the school selected should also be typical of those to which the generalizations are to apply.

are approximately the same size and in communities of much the same social and economic status. It would not be desirable to select one school that employs ability grouping while the other selected does not. It would not be advisable to select one school that has elaborate library and laboratory facilities, while the other school selected does not have these advantages. The control of instructional materials and time spent in learning activity is most effectively accomplished by preparing detailed lesson plans for both the experimental and control groups. The possible variation in other factors, such as home facilities for study and extra duties performed by the pupils, should be investigated and the differences observed used as the basis of a limitation placed on the experimental conclusions.

5. *Cooperative experiments.* Cooperative experiments may be conducted in the same, or in different schools. It is considered the most desirable practice to have the cooperating teachers instruct pairs of experimental and control groups. In the interpretation of the data obtained by cooperative experimentation one of two methods may be used. Each pair of groups may be regarded as a sub-experiment, and the conclusions of these sub-experiments may be compared with one another. The other method of interpretation is that which depends on the concept of the cooperative experiment as a single experiment. All of the experimental pupils are regarded as composing one large experimental group; all of the control pupils are considered as a single large control group.¹²⁴ The difference in gains obtained will, of course, be the average of the differences in gains for all of the paired groups. The increase in size of the experimental and control groups by this combining of data will reduce considerably the variable errors of measurement, validity, and sampling that existed for the individual pairs of groups.¹²⁵ It is probable that systematic errors, since they are likely to vary from one pair of groups to another, will tend to offset each other. In other words, they may become variable errors and, hence, be more easily accounted for in the statistical treatment of results. It is probable that such combination of experimental and control pupils will aid in securing more perfect equivalence with respect to pupil factors and more perfect control of non-experimental factors, since departures from control in the several pairs of groups may tend to balance each other. It is probable, also, that the combined group of experimental pupils and the combined group of control

¹²⁴It may be desirable to exclude one or more pairs of groups because of gross errors. For an example of such exclusion, see:

Douglass, H. R., *et al.* "The Relative Effectiveness of the Problem and Lecture Methods of Instruction in Principles of Economics," *University of Oregon Publication*, Vol. 1, No. 7. Eugene, Oregon: University of Oregon, 1929, p. 290.

¹²⁵The next chapter describes the interpretation of experimental data with reference to these errors.

pupils will be more representative of the pupils to whom the generalizations are to apply than one of the small groups would be. The combining of data in this way does not guarantee all this, however, since it is easily possible for a systematic error of measurement, validity, or sampling, or a lack of control of some non-experimental factor to run through the measures of all the groups and thus bias the combined results. For example, all of the teachers might be zealous for the new method of procedure that constitutes the experimental factor, since to be zealous for it is the mode. Again, all of the teachers might be unskilled in the use of the method because of its newness. If the cooperative group were all selected from rural schools, or all from city schools, representativeness with respect to all children would not be increased by combining results. Experimentation by cooperation of teachers and schools is eminently desirable, but in order to secure dependable results, data from the cooperating groups should be combined with care if summation of faults is to be avoided.

CHAPTER III

THE INTERPRETATION OF DIFFERENCES IN GAINS

The general plan of handling experimental data. The general plan of handling experimental data may be illustrated by considering an experiment involving two groups—one an experimental group and the other a control group. The administration of the achievement test at the beginning of the experimental period¹ yields scores as follows:

For the experimental group $e_1, e_2, e_3 \dots \dots e_n$ whose mean is E_1 .

For the control group $c_1, c_2, c_3 \dots \dots c_n$ whose mean is C_1 .

The administration of the test at the end of the experimental period yields a second set of scores:

For the experimental group $e_1', e_2', e_3', \dots \dots e_n'$ whose mean is E_2 .

For the control group $c_1', c_2', c_3' \dots \dots c_n'$ whose mean is C_2 .

The mean gain in achievement made by the experimental group is $E_2 - E_1$ and is designated by the symbol, "Gain E."² The mean gain in achievement made by the control group is $C_2 - C_1$ and is labeled, "Gain C." The difference in gains, D , is equal to Gain E - Gain C.

The problem of interpretation. The problem of interpretation is to determine the extent to which the difference in gains, D , may be due to imperfections in the experimental procedure and in the measures of achievement and, consequently, to determine the extent to which the experimenter is justified in interpreting D as indicating the merit of the experimental factor. The errors introduced in the measures of achievement by the imperfections of the experimental procedure and of the measuring instruments are of two kinds: variable and systematic. The effect of the variable errors is described in terms of the chances that, if they were eliminated, the difference would have the opposite sign. For example, assume that the obtained difference, D , is equal to 2.5. If the variable errors were eliminated, D would be different—possibly 4.2, possibly 6.4, possibly 0.7, possibly -1.2, possibly other values. The correct value cannot be calculated, but, if we have certain information about the magnitude of the variable errors, we can calculate the chances that the true value of D will fall within any interval. In view of the fact that D is an index of the

¹Under certain conditions it is appropriate to omit this initial test.

²This may also be obtained by calculating the individual gains and averaging them.

merit of the experimental factor, it is obvious that we are primarily concerned with the chances that the true D may be negative. If the calculated D is positive, the true D is more likely to be positive than negative; hence, the experimental factor is more likely to be superior than inferior. However the experimenter cannot make a very strong claim for the superiority of this factor unless the chances for the true D being positive are much greater than for it being negative. How many times greater they should be in order to justify a claim for the superiority is a matter of opinion. The chances are 3 to 1 in favor of the true difference being positive when the obtained difference, D , is equal to the probable error of measurement, $P.E._{Meas.D}$, and slightly greater than 10 to 1 when D is equal to twice $P.E._{Meas.D}$. It may seem that these chances, especially 10 to 1, are sufficient betting odds to justify a rather strong claim for the superiority of the experimental factor. Undoubtedly they do justify some claim for superiority, but it is a common practice to require that they be at least 369 to 1 in order to call the difference *statistically significant*³ with reference to the variable errors being considered. This condition is fulfilled when the difference is equal to or greater than 2.78 times the standard error of the difference or approximately 4.4 times the probable error of the difference.

It should be noted that in addition to determining the degree of significance of the obtained difference, D , as indicated in the preceding paragraph, it is necessary to consider the effect of the systematic errors due to imperfections in the experimental procedure and to imperfections in the measuring instruments used. When the experimenter desires to generalize from his results, he must consider also the extent to which the two groups of pupils are representative of the larger group for which he desires to express conclusions.

If the experimental group is assumed to be equivalent to the control group, the specific questions to be considered are:

1. What allowance⁴ must be made for variable errors in the measures of achievement?
2. What allowance must be made for systematic errors of measurement not common to all groups of measures of achievement?
3. What allowance must be made for variable errors of validity in the measures of achievement?

³"Significant" and "significance" are technical terms in the field of statistics.

⁴The allowance for variable errors in the measures of achievement will be expressed in terms of a standard error of measurement, $\sigma_{meas.}$ or of a probable error of measurement, $P.E._{meas.}$. The allowance for chance errors of validity will be expressed in a similar way. The allowance for systematic errors of measurement or for lack of control of important educative factors will be expressed as an amount to be added to or subtracted from the obtained difference.

4. What allowance must be made for lack of control of important educative factors?
5. In generalizing what allowance must be made for possible non-representativeness of the groups of pupils?

Although it is necessary to answer these questions separately, the final interpretation of the difference, D , must be based on the combined allowance for all causes. For example, if D is equal to or greater than ten times $P.E._{Meas.D}$, it does not necessarily follow that the experimental factor is superior, because the other allowances must also be considered, and it might happen that the combined effect of these would be to reverse the sign of D , or at least to make its reversal not improbable.

The allowance for variable errors of measurement. An approximate index of the variable errors⁵ of measurement in a group of scores may be determined by giving the test twice⁶ to the pupils included in the experiment and computing the coefficient of correlation between the two sets of scores.⁷ This result is called the coefficient of reliability of the test and is designated in the following discussion by the symbol⁸ r_{12} . The coefficient of reliability has been determined for several tests and this value of r_{12} may be used, provided the standard deviation of the scores on which its determination is based is approximately equal to the standard deviations of the scores in the experiment.

If the standard deviations of the distributions of experimental scores are not approximately equal to each other, or to the standard deviations of the distributions on which the reliability coefficient was based, then the coefficient of reliability should be corrected by means of Kelley's formula for the relation between ranges in obtained scores and reliability coefficients.⁹ The magnitude of the variable errors of measurement in the individual scores is indicated by the standard or probable error of measurement of a score.¹⁰ If one is using the obtained test scores the following formula should be used to compute the standard error of measurement.¹¹

⁵Variable errors of measurement differ for different members of a group not only in magnitude but in direction as well. Such errors tend to distribute themselves about zero as a mean according to the normal distribution, or curve of chance. The fact that variable errors group themselves in this way justifies one in saying that the chances are greatest that the true mean is close to the observed mean; the chances of the true mean being anything very different from the observed mean decrease the further we get from the observed mean.

⁶It is desirable that two parallel forms be used.

⁷Another means of determining this coefficient is by correlating the scores for odd and even items on one application of the test and correcting the obtained coefficient with the Spearman-Brown formula.

⁸A number of statisticians including Kelley make use of the symbol r_{11} in place of r_{12} .

⁹Kelley, T. L. *Statistical Method*. New York: The Macmillan Company, 1923, p. 222.

¹⁰The standard and probable errors of a score apply to all the individual scores from which they were calculated, *collectively*. They do not apply without this interpretation to any given individual score.

¹¹The probable error of any measure may always be obtained by multiplying the standard error by the constant .6744897 or .6745. This statement applies particularly to the formulae for $\sigma_{Meas.M}$, $\sigma_{(m+v)_M}$, and $\sigma_{(m+s)_M}$ which are changed to $P.E._{Meas.M}$, and so on by merely inserting the constant, .6745, before the expression for the standard error.

$$\sigma_{\text{Meas. Score}} = \sigma_{\text{Dist.}} \sqrt{1 - r_{12}}$$

It is known that, because of variable errors, obtained test scores tend to vary more widely from the mean than do true test scores. Kelley¹² has shown that this variability may be reduced and some of the error eliminated by computing "estimated true scores" by means of the following regression equation:

$$\bar{X}_{\infty} = r_{12}X + (1 - r_{12})M_x$$

$$\bar{X}_{\infty} = \text{regressed or estimated true score}$$

X = obtained test score

M_x = mean of the distribution of scores

r_{12} = the coefficient of reliability (Kelley uses r_{11} for the same thing.)

After obtained scores have been changed to "estimated true scores," the magnitude of the variable error in the individual scores is indicated by the standard error of measurement found by the following formula:¹³

$$\sigma_{\text{Meas. Est. True Score}} = \sigma_{\text{Dist.}} \sqrt{1 - r_{12}^2}$$

The variable error of measurement of the mean of a group of test scores is obtained by dividing the appropriate formula above by the square root of the size of the group. If one is dealing with obtained scores, the formula becomes:

$$\sigma_{\text{Meas. M}} = \frac{\sigma_{\text{Dist.}} \sqrt{1 - r_{12}^2}}{\sqrt{N}}$$

If one is dealing with regressed or "estimated true scores," the formula is:

$$\sigma_{\text{Meas. M}} = \frac{\sigma_{\text{Dist.}} \sqrt{1 - r_{12}^2}}{\sqrt{N}}$$

These formulae are the appropriate ones to use in determining the standard errors of measurement of the means E_1 , E_2 , C_1 , and C_2 . The $\sigma_{\text{Dist.}}$ used is that obtained by calculation from the distribution of obtained initial or final test scores corresponding to E_1 , E_2 , C_1 , or C_2

¹²Kelley, T. L. *Interpretation of Educational Measurements*. Yonkers-on-Hudson, New York: World Book Company, 1927, p. 178.

¹³For a description and derivation of this formula, see: Kelley, *op. cit.*, p. 176-77. $\sigma_{\text{Dist.}}$ = σ of the distribution of the obtained scores for which the error is being computed.

for which the error is being computed and may be designated by the symbols $\sigma_{Dist.e}$, $\sigma_{Dist.e'}$, $\sigma_{Dist.e}$, or $\sigma_{Dist.e'}$.

In determining the standard error of measurement of Gain E or Gain C, one should insert the values of the standard errors of measurement of E_1 and E_2 or C_1 and C_2 , obtained by the use of the above formulae in the formulae below.¹⁴ If one has determined the probable errors of measurement of E_1 and E_2 , or C_1 and C_2 , the formulae to be used are similar. In place of each standard error substitute the corresponding probable error in order to obtain the probable error of measurement of Gain E or Gain C.¹⁵

$$\sigma_{Meas.Gain E} = \sqrt{\sigma_{Meas.E_1}^2 + \sigma_{Meas.E_2}^2 - 2r_{E_1E_2} \cdot \sigma_{Meas.E_1} \cdot \sigma_{Meas.E_2}}$$

$$\sigma_{Meas.Gain C} = \sqrt{\sigma_{Meas.C_1}^2 + \sigma_{Meas.C_2}^2 - 2r_{C_1C_2} \cdot \sigma_{Meas.C_1} \cdot \sigma_{Meas.C_2}}$$

Standard errors of measurement of the mean gains may be computed in another way with equivalent results. To do so requires calculation of the individual gains by subtracting e_1 from e_1' , e_2 from e_2' , c_1 from c_1' , and so on for all the individuals participating in the experiment. The standard error of the mean gain is then calculated by the appropriate formula below:¹⁶

$$\sigma_{Meas.Gain E or C} = \frac{\sigma_{Distribution of Individual Gains} \sqrt{1 - r_{12}}}{\sqrt{N}}$$

$$\sigma_{Meas.Gain E or C} = \frac{\sigma_{Distribution of Individual Gains} \sqrt{r_{12} - r_{12}^2}}{\sqrt{N}}$$

¹⁴This procedure is justified only when the same test, or equivalent forms, are administered at the beginning and end of the experiment. When different tests are given, there is opportunity for difference in units and zero points that prevents computation of gains. When the use of the same test, or equivalent forms, is not feasible, comparison must be restricted to the final test means, E_2 and C_2 , and the standard error of difference between these means computed by the formula:

$$\sigma_{Meas.D} = \sqrt{\sigma_{Meas.E_2}^2 + \sigma_{Meas.C_2}^2} \\ (E_2 - C_2)$$

¹⁵This statement applies also to formulae given later for $\sigma_{Meas.D}$, $\sigma_{(m+v) Gain E}$, $\sigma_{(m+v) Gain C}$, $\sigma_{(m+s) Gain E}$, $\sigma_{(m+s) Gain C}$, and $\sigma_{(m+s) D}$.

The coefficients of correlation used in these formulae are, theoretically, those between the mean of initial test scores (E_1) and the mean of final test scores (E_2) of a large number of similar experimental groups. The same is true for the control groups. Practically, the coefficient used is obtained by correlating the initial and final test scores of the experimental group to give $r_{E_1E_2}$ and the initial and final test scores of the control group to obtain $r_{C_1C_2}$. For justification of this, see:

Kelley, T. L. *Statistical Method*. New York: The Macmillan Company, 1923, p. 178.
¹⁶ r_{12} should be corrected to correspond with the standard deviation of the individual gains used. See page 61.

To determine the standard error of measurement of the difference in gains, D, one should insert the values of the standard errors of measurement of the Gains E and C in the formula below.¹⁷

$$\sigma_{\text{Meas. D}} = \sqrt{\sigma_{\text{Meas. Gain E}}^2 + \sigma_{\text{Meas. Gain C}}^2}$$

The following hypothetical example illustrates the use of the preceding formulae. It is assumed that equivalent forms of an achievement test were used whose coefficient of reliability, r_{12} , is equal to .85. It is also assumed that the correlations between the initial and final test scores have been computed for both groups, and the means and standard deviations of the four distributions obtained. These hypothetical values are:

r_{12}	= .85	E_1	= 73.32	$\sigma_{\text{Dist. } e}$	= 7.60
$r_{E_1 E_2}$	= .71	E_2	= 76.25	$\sigma_{\text{Dist. } e'}$	= 7.44
$r_{C_1 C_2}$	= .65	C_1	= 73.20	$\sigma_{\text{Dist. } c}$	= 7.56
N	= 25	C_2	= 74.12	$\sigma_{\text{Dist. } c'}$	= 7.84

$$\text{Gain E} = E_2 - E_1 = 76.25 - 73.32 = 2.93$$

$$\text{Gain C} = C_2 - C_1 = 74.12 - 73.20 = .92$$

D, the difference in gains = Gain E - Gain C = 2.93 - .92 = 2.01

$$\sigma_{\text{Meas. E}_1} = \frac{7.60\sqrt{1-.85}}{\sqrt{25}} = .5887$$

¹⁷One step of the total procedure may be eliminated by the use of the following formula:

$$\sigma_{\text{Meas. D}} = \sqrt{\sigma_{\text{Meas. E}_1}^2 + \sigma_{\text{Meas. E}_2}^2 + \sigma_{\text{Meas. C}_1}^2 + \sigma_{\text{Meas. C}_2}^2 - 2r_{E_1 E_2} \cdot \sigma_{\text{Meas. E}_1} \cdot \sigma_{\text{Meas. E}_2} - 2r_{C_1 C_2} \cdot \sigma_{\text{Meas. C}_1} \cdot \sigma_{\text{Meas. C}_2}}$$

For a derivation of this formula with respect to errors of sampling, see: Lindquist, E. F. and Foster, R. R. "On the Determination of Reliability in Comparing the Final Mean-Scores of Matched Groups," *Journal of Educational Psychology*, 20:102-106, February, 1929.

The comment might be made that these formulae neglect the correlation that may exist between the gains of the paired pupils. In other words, the expression, $-2r_{gegc} \sigma_{\text{Meas. Gain E}} \sigma_{\text{Meas. Gain C}}$ where r_{gegc} is the coefficient obtained by correlating the distribution of individual gains of the experimental pupils with the distribution of individual gains of the control pupils, should also be included under the radical of the formula given above, or under the radical in the long formula just given. The authors just referred to justify its exclusion by the statement, "But since there can be no real correlation between the scores of one group and those of another may be omitted from the equation" p. 105.

Coefficients of correlation are regularly obtained by correlating two distributions of measures of the same individuals. The uncertain conclusions of research on the effect of practice on individual differences would cause one to question the dependability of a coefficient obtained by correlating gains of paired individuals. Owing to the uncertainty of this correlation the probable and standard errors obtained with the above formula are interpreted as "limits beyond which the true error cannot fall." For arguments in favor of the inclusion of this expression, see:

Walker, H. M. "Concerning the Standard Error of a Difference," *Journal of Educational Psychology*, 20:53-60, January, 1929.

$$\sigma_{\text{Meas. E}_2} = \frac{7.44\sqrt{1-.85}}{\sqrt{25}} = .5763$$

$$\sigma_{\text{Meas. C}_1} = \frac{7.56\sqrt{1-.85}}{\sqrt{25}} = .5856$$

$$\sigma_{\text{Meas. C}_2} = \frac{7.84\sqrt{1-.85}}{\sqrt{25}} = .6073$$

$$\sigma_{\text{Meas. Gain E}} = \sqrt{(.5887)^2 + (.5763)^2 - 2 \times .71 \times .5887 \times .5763} = .4438$$

$$\sigma_{\text{Meas. Gain C}} = \sqrt{(.5856)^2 + (.6073)^2 - 2 \times .65 \times .5856 \times .6073} = .4994$$

$$\sigma_{\text{Meas. D}} = \sqrt{(.4438)^2 + (.4994)^2} = .6681 \text{ or } .67$$

Since the difference in gains, which is 2.01, is three times as large as the standard error of measurement of the difference, which is .67, the following interpretation is justified. Considering only the variable error of measurement and assuming that errors due to faulty equivalence, failure to control external non-experimental factors, and departure from validity of the measuring instruments have been eliminated, or otherwise accounted for, then for the groups concerned, *and only for the groups concerned*, the difference in achievement indicates the superiority of the status of the experimental factor prevailing in the experimental group. Subject to the limitations just expressed, the probability that the observed difference has the same sign, or is in the same direction, as the true difference is greater than the ratio 740 to 1.¹⁹ Stated in another way, if the experiment could be repeated with the *same* groups, under the same conditions, the chances of obtaining another observed difference of the same sign, or in the same direction, are greater than the ratio 740 to 1.²⁰

The example given illustrates the calculation when obtained scores and the standard errors are used. Other examples might have been given using regressed scores with the standard errors, obtained scores with the probable errors, or regressed scores with the probable errors. Although the calculation of these is similar, care must be

¹⁹As has already been explained, the standard or probable error obtained by the procedure outlined is regarded as a limit. If it were feasible to obtain a reliable coefficient for the small amount of correlation that may exist between the gains of the paired pupils and thus to arrive at a more accurate and an always smaller standard error, the chances of statistical significance would of course be greater.

²⁰The comment might be made in regard to this interpretation that repetition under the same experimental conditions with the same groups should secure differences not only of the same sign, but of the same magnitude. Identical differences would be secured with identical conditions and groups, if it were not for the unreliability of the initial and final tests. The standard and probable errors of measurement of a difference allow for this unreliability and nothing else.

taken to use the appropriate formulae. The following table gives the chances of statistical significance of differences that are a given number of times larger than the standard or probable error of the differences. The second column gives the chances of the true difference falling within the range, plus and minus, of the probable or standard error of the difference. This interpretation is less applicable to experimentation than that given in the third column. The experimenter is most interested, not in the magnitude of the observed difference, but in the probability that the observed difference has the same sign as the true difference. When these chances are great, 369 to 1 or better, the experimenter is justified in asserting *that the variable errors of measurement do not destroy the dependability of a conclusion in favor of the superiority of the experimental factor.*²¹

TABLE I.
CHANCES OF STATISTICAL SIGNIFICANCE OF A DIFFERENCE

	The chances that the true difference does not differ from the observed difference by more than the indicated amount.	The chances that the true difference has the same sign, or is in the same direction, as the observed difference.
D = σ_D	2.15 to 1	5.3 to 1
D = 2 σ_D	21 to 1	43 to 1
D = 2.78 σ_D *	184 to 1	369 to 1
D = 3 σ_D	369 to 1	740 to 1
D = 4 σ_D	15,772 to 1	31,545 to 1
D = P.E. _D	1 to 1	3 to 1
D = 2 P.E. _D	4.6 to 1	10.3 to 1
D = 3 P.E. _D	22 to 1	45 to 1
D = 4 P.E. _D	142 to 1	286 to 1
D = 5 P.E. _D	1,341 to 1	2,684 to 1

*This multiple of the standard error of difference appears in McCall's formula for the experimental coefficient:

$$E.C. = \frac{\text{Difference}}{2.78 \times \sigma_{\text{Difference}}}$$

When the expression is equal to 1.0, the chances that the true difference has the same sign are in the ratio of 369 to 1. McCall uses this as the critical point below which differences should not be recognized as significant. If the chances are greater than 369 to 1 then the difference is to be recognized as significant. The statement, "An experimental coefficient of 1.0 is just exactly practical certainty. An experimental coefficient of .5 means half certainty, one of 2.0 means double certainty and so on," is not very meaningful since it is impossible to multiply certainty. See:

McCall, W. A. *How to Measure in Education*. New York: The Macmillan Company, 1922, p. 404-405.

The allowance for variable errors of validity. Achievement is not a unitary thing. It includes three types of controls of conduct: (1) specific habits; (2) knowledge; (3) general patterns of conduct. In a given case the achievement to be considered may be restricted to only certain elements under one of the rubrics. For example, in an experi-

²¹See footnote on page 60.

ment to determine the relative merits of two methods of teaching addition, the achievement to be measured might be restricted to the skills (specific habits) that function in doing examples of addition of a specified type. In an experiment to determine the relative effect of certain methods of teaching English literature in the high school, the achievement to be measured might be restricted to changes in the interest of the pupils in reading literature of a specified type. On the other hand, when the problem of an experiment asks concerning the effect of an educative factor without any restrictions, there is the implied requirement for measuring *all* elements of the resulting pupil achievement which may include specific habits, knowledge, and general patterns of conduct.

In order for an achievement test or a group of such tests to yield results that are valid for a given experiment, it must measure, either directly or indirectly, all of the elements of the achievement or a representative sample of all of the achievements specified or implied by the statement of the problem of the experiment.

The allowance for the variable errors of validity can be calculated if the coefficient of validity is known. In order to obtain this coefficient it will, of course, be necessary to have a valid criterion measure of the achievement specified by the problem of the experiment. If such measures were available for the pupils in the experiment, they would be used, and the question of validity would be eliminated. This will seldom be the case, but it may happen that the test used has been validated previous to its use in the experiment by calculating the coefficient of correlation between the scores it yields and the valid criterion measures. If this coefficient, r_{1C} , is known and the standard deviations on which it is based are approximately equal to the standard deviation of the obtained scores, then the gross²² allowance for variable errors of validity (validity and measurement) may be calculated by the following formulae:

$$\sigma_{(m+v)E_1, E_2, C_1, \text{ or } C_2} = \frac{\sigma_{\text{Dist.}} \sqrt{1 - r_{1C}^2}}{\sqrt{N}}$$

$$\sigma_{(m+v)\text{Gain E}} = \sqrt{\sigma_{(m+v)E_1}^2 + \sigma_{(m+v)E_2}^2 - 2r_{E_1E_2} \sigma_{(m+v)E_1} \cdot \sigma_{(m+v)E_2}}$$

$$\sigma_{(m+v)\text{Gain C}} = \sqrt{\sigma_{(m+v)C_1}^2 + \sigma_{(m+v)C_2}^2 - 2r_{C_1C_2} \sigma_{(m+v)C_1} \cdot \sigma_{(m+v)C_2}}$$

$$\sigma_{(m+v)D} = \sqrt{\sigma_{(m+v)\text{Gain E}}^2 + \sigma_{(m+v)\text{Gain C}}^2}$$

²²The allowance indicated by the use of these formulae will be too large, because the criterion measures, as well as the measures being validated, include variable errors of measurement. It is useful, however, as a limit beyond which the probable error of validity cannot go.

The standard or probable error of the difference between the mean gains secured by the use of these, or similar, formulae is to be interpreted in the same way as the standard or probable error of the difference between the mean gains due to variable errors of measurement alone. In the interpretation, however, it should be pointed out that the measure of error secured covers the variable errors of validity of the test used, the variable errors of measurement of the test used, and the variable errors of measurement of the criterion. To the extent that the criterion departs from validity itself, the measure also includes variable errors of validity of the criterion. Therefore, the index secured should be dealt with as a limit rather than as an accurate means of allowing for variable errors of validity. If the difference between the means, or mean gains, is 2.78 or more times the standard error of this difference as found by the above formulae, one is justified in stating that it is statistically significant since the calculated measure of error is known to be somewhat larger than the true measure of error.²³

It will seldom be possible to calculate, even approximately, the allowance for chance errors of validity in the measures of achievement and, hence, usually the experimenter must make an estimate. In making this estimate, the specifications, both explicit and implied, of the problem of the experiment in regard to the achievement to be measured must be clearly recognized. In many cases the quality of permanency is implied, and when there is this implication, it must be considered. No rules can be specified for estimating the allowance for chance errors of validity, but it is probably true that, except when the achievement is confined to specific habits or is relatively narrow, this allowance is likely to be equal to or greater than the allowance for chance errors of measurement.

The allowance for systematic errors of measurement or validity, lack of equivalence of the groups, and non-equality of significant non-experimental factors. In considering the allowance for systematic²⁴ errors of measurement or validity it should be noted that the effect of a systematic error is eliminated from the mean gain in achievement when it is the same in the two sets of scores from which the gain is computed.²⁵ Similarly, the effect is eliminated in the difference when

²³See footnote on page 64.

²⁴Errors which are present in all the scores of a group, not necessarily of the same magnitude, but always in the same direction are called *systematic* errors. For example, if distances of 9, 12, and 15 feet were measured by a yardstick one-half inch too long, systematic errors would occur of $1\frac{1}{2}$, 2, and $2\frac{1}{2}$ inches. The older term "constant error" is less desirable, since it implies that the individual errors of the members of the group be of the same magnitude.

²⁵Let $M_1 = 80$ (the initial test mean of the group)

Let $M_2 = 84$ (the final test mean of the group)

Then if each has a systematic error of +2, and if this is deducted, $M_1 = 78$ and $M_2 = 82$; but the difference between them, or gain, is still 4.

it is the same for the two gains that are compared.²⁶ Hence, it is necessary to consider only the cases in which these conditions do not prevail.

Systematic errors of measurement may result from the failure to control conditions at the time of measurement. For example, the teacher of the experimental group may permit the students of that group to spend a few more minutes on the final test than were allowed the control students. The control group may be given the test with more ample directions than are given to the experimental group. In interpreting a difference, an experimenter should inquire whether possible systematic errors of measurement or validity in the measures of achievement have been eliminated in the manipulation of the data. In case it does not seem highly probable that they have been eliminated, he must estimate their probable effect upon the difference of gains.

Systematic errors of validity may result from failure to use tests which are equally valid with respect to the achievement of both groups. For example, in a comparison of the project and the assignment methods, the tests used may favor the specific abilities possibly more favorably engendered by the assignment method and, hence, may not evaluate adequately the more general abilities acquired by the project pupils. Thus, a systematic error of validity may cause the difference to be interpreted in favor of the assignment method, whereas, if the values obtained by the project pupils had all been measured, the opposite conclusion might have been reached.

The allowances to be made for systematic errors of measurement and validity cannot be calculated in quantitative terms. Estimates must be determined and applied as limitations in the interpretation of results.

It is difficult, if not impossible, to estimate accurately the effect of *lack of equivalence* of groups upon the difference between the gains. For example, consider an experiment in which the attempt is made to determine the relative effectiveness of two techniques of drill on arithmetical calculation—Technique X and Technique Y. Suppose the two groups differ slightly in mean mental age, that for Group A being 12.45 and that for Group B being 12.68. Assuming equivalence in all other respects, what allowance for this non-equivalence should be made in interpreting the difference between the gains made by the two groups? In order to answer this question accurately, it would be

²⁶Let $G_1 = 6$ (the gain in achievement for the experimental group)

Let $G_2 = 2$ (the gain in achievement for the control group)

Then if each has a systematic error of -3 , and if this is deducted (algebraically), $G_1 = 9$ and $G_2 = 5$; but the difference between them is still 4.

necessary to know the influence of mental age upon the achievement concerned under the conditions of the experiment. In general, this information is not available; therefore, the experimenter can only estimate roughly the probable effect of the lack of equivalence in mean mental ages.

It should be noted that the influence of non-equivalence of groups upon the difference of the gains may either be positive or negative. This fact expressed in equation form would be:

Calculated difference = true difference \pm effect of non-equivalence

Hence, the true difference may be exaggerated, minimized, or negated by the lack of equivalence of the groups.

Failure to keep equivalent one or more of the important *non-experimental* factors will affect the difference of the gains in achievement. For example, when a teacher of the experimental group believes in the method of teaching that forms the experimental factor and is zealous in carrying it out, the gain of the experimental group is likely to be greater than it would be under a teacher who is prejudiced against the method, or one who is neutral with reference to it. Hence, a lack of equivalence of "teacher zeal" will introduce a systematic error into one of the gains. It is difficult to determine the magnitude of such an error and, hence, to correct for it. An approximation to its magnitude may be obtained by performing a supplementary experiment in which the uncontrolled factor becomes the experimental factor. This is not often feasible. Usually, the experimenter goes to the literature for experimental evidence to prove that this factor has a negligible effect on learning, or he gives reasons based on observations made during his own experiment to show that the influence of the factor is not sufficient to destroy the significance of his findings. It is evident that failure to control important educative factors, with consequent introduction of constant errors of unknown magnitude and direction, will render small differences in gains insignificant.

In concluding this consideration of systematic errors in experimental studies, it should be emphasized that the allowance to be made for them in interpreting the difference in gains is likely to be much larger than that to be made for the variable errors of measurement and of validity. Since the allowance for variable errors is inversely proportional to the square root of the number of scores, this allowance will become relatively small when the size of the experi-

mental and control groups is large. The size of these groups does not affect the systematic errors, if other conditions remain the same. This qualifying clause is added because when conditions are varied in increasing the size of the groups, as is likely to be the case in a cooperative experiment, the systematic errors may be decreased. When this occurs, the variable errors are increased, and, consequently, the statement that the allowance for these errors is inversely proportional to the square root of the number of scores is not always true. However, it is impossible to know precisely what happens in a particular case; therefore, it is a safe plan to assume that the size of the group does not affect the systematic errors.

The allowance in generalizing for non-representativeness of groups used in the experiment. In generalizing the interpretation of a difference between gains, the representativeness of the groups of pupils used in the experiment must be considered. In doing this, two cases are encountered: (1) the pupils selected in a random manner such that the groups may be expected to be representative except for the operation of chance; (2) the pupils not selected in a random manner and the non-representativeness of the groups due to factors other than chance. The allowance for probable non-representativeness in the first case can be calculated by certain formulae, which are described below; but this case is not the one usually encountered. In fact, it would be practically impossible to select by a random method two groups of pupils for an experiment. For example, if it were desired to select two random groups from the fourth-grade pupils in a city-school system, the names of all such children could be arranged alphabetically and the names taken from this list in a random manner. It is apparent, however, that such a procedure would seldom, if ever, be feasible. Hence, in generalizing the interpretation of a difference between gains, an experimenter has to deal with the second case, and for this case, no formulae are applicable.

1. *Generalizing from a random sample.* A random sample is not necessarily a perfectly representative one.²⁷ Chance is operative, and the mean of the sample may or may not coincide with the mean of the entire population from which the sample was drawn. The larger the sample the greater the chances that the mean of the sample will be somewhere close to the mean of the population. It may fall above or below. Where it is likely to fall is shown by determining the stand-

²⁷For an excellent discussion of sampling, see: Walker, H. M. "The Sampling Problem in Educational Research," *Teachers College Record*, 30:760-74, May, 1929.

ard or probable error of the mean due to sampling. The formula²³ to be used is:

$$\sigma_{\text{Sampling}_M} = \frac{\sigma_{\text{Dist.}} \sqrt{r_{12}}}{\sqrt{N}}$$

Since the experimenter needs to know the magnitude of the combined effect of sampling and the variable errors of measurement, the following formula may be used when dealing with random groups.

$$\sigma_{(s+m)_M} = \frac{\sigma_{\text{Dist.}}}{\sqrt{N}}$$

The values obtained from this formula for the standard errors of means E_1 , E_2 , C_1 , and C_2 should be inserted in the following formulae²⁹ to determine the standard errors of measurement and sampling of the mean Gains E and C .

$$\sigma_{\text{Gain } E}^{(s+m)} = \sqrt{\sigma_{(s+m)E_1}^2 + \sigma_{(s+m)E_2}^2 - 2r_{E_1E_2} \cdot \sigma_{(s+m)E_1} \cdot \sigma_{(s+m)E_2}}$$

$$\sigma_{\text{Gain } C}^{(s+m)} = \sqrt{\sigma_{(s+m)C_1}^2 + \sigma_{(s+m)C_2}^2 - 2r_{C_1C_2} \cdot \sigma_{(s+m)C_1} \cdot \sigma_{(s+m)C_2}}$$

If the individual gains have been determined, the standard errors of measurement and sampling of the mean gains may be determined by the following formula:

$$\sigma_{\text{Gain } E \text{ or Gain } C}^{(m+s)} = \frac{\sigma_{\text{Distribution of Individual Gains}}}{\sqrt{N}}$$

²³It has been shown that the standard error of the mean due to sampling *alone* is equal to $\frac{\sigma_{\text{Dist.}} \sqrt{r_{12}}}{\sqrt{N}}$, where $\sqrt{r_{12}}$ is the index of reliability of the measuring instrument used, and is not equal to $\frac{\sigma_{\text{Dist.}}}{\sqrt{N}}$, the formula generally used, and which, for example, is given in:

Holzinger, K. J. *Statistical Methods for Students in Education*. Boston: Ginn and Company, 1928, p. 232.

Kelley, and Huffaker and Douglass have shown that $\frac{\sigma}{\sqrt{N}}$, ordinarily supposed to stand for merely the errors of sampling, includes both the errors of sampling and of measurement, *see*:

Kelley, T. L. "Note Upon Holzinger's Formula for the Probable Error," *Journal of Educational Psychology*, 14:376-77, September, 1923.

Huffaker, C. L. and Douglass, H. R. "On the Standard Errors of the Mean Due to Sampling and to Measurement," *Journal of Educational Psychology*, 19:643-49, December, 1928.

²⁹One is justified in using these formulae only when the same test, or equivalent forms of it, are administered at the beginning and end of the experiment. When this is not feasible, comparison must be restricted to the final test means, E_2 and C_2 , and the standard error of measurement and sampling computed by the following formula:

$$\sigma_{(E_2 - C_2)}^{(m+s)_D} = \sqrt{\sigma_{(m+s)E_2}^2 + \sigma_{(m+s)C_2}^2}$$

See footnote on page 63.

The standard errors of the mean gains may then be inserted in the following formula³⁰ to obtain the standard error of measurement and of sampling of the difference, D:

$$\sigma_{(s+m)D} = \sqrt{\sigma_{(s+m)Gain E}^2 + \sigma_{(s+m)Gain C}^2}$$

The expression $-2r_{ge gc} \sigma_{(s+m)Gain E} \cdot \sigma_{(s+m)Gain C}$ is not included in the above formula for the reason given in the footnote on page 64. In order to illustrate the use of these formulae, the data used in the example given to illustrate the calculation of the standard errors of measurement of a difference will be used again in calculating the standard error of measurement and sampling of the difference 2.01. The illustration applies to the procedure in which the standard errors of the initial and final test means are first computed.

$$\sigma_{(s+m)E_1} = \frac{7.60}{\sqrt{25}} = 1.520 \qquad \sigma_{(s+m)E_2} = \frac{7.44}{\sqrt{25}} = 1.488$$

$$\sigma_{(s+m)C_1} = \frac{7.56}{\sqrt{25}} = 1.512 \qquad \sigma_{(s+m)C_2} = \frac{7.84}{\sqrt{25}} = 1.568$$

$$\sigma_{(s+m)Gain E} = \sqrt{(1.520)^2 + (1.488)^2 - 2 \times .71 \times 1.520 \times 1.488} = 1.1458$$

$$\sigma_{(s+m)Gain C} = \sqrt{(1.512)^2 + (1.568)^2 - 2 \times .65 \times 1.512 \times 1.568} = 1.2894$$

$$\sigma_{(s+m)D} = \sqrt{(1.146)^2 + (1.289)^2} = 1.7263 \text{ or } 1.73$$

The chances that the true difference has the same sign, or is in the same direction, may be obtained approximately from the second column of Table I. Calculated more accurately—since the difference 2.01 is 1.16 times the standard error of the difference—the chances are slightly greater than 7 to 1 (interpreting the standard error as a limit) that the true difference will have the same sign. Authorities recommend that the experimenter should not consider

³⁰If the errors of E₁, E₂, C₁, and C₂ have been computed the formulae given may be combined into:

$$\sigma_{(s+m)D} = \sqrt{\sigma_{(s+m)E_1}^2 + \sigma_{(s+m)E_2}^2 + \sigma_{(s+m)C_1}^2 + \sigma_{(s+m)C_2}^2 - 2r_{E_1 E_2} \sigma_{(s+m)E_1} \sigma_{(s+m)E_2} - 2r_{C_1 C_2} \sigma_{(s+m)C_1} \sigma_{(s+m)C_2}}$$

such a difference as statistically significant when attempting to generalize from his data. The combined effect of variable errors of measurement and of variable errors of sampling, alone, is sufficient to render the statistical significance of the difference inferior to that customarily demanded. The findings may be considered significant so far as the groups, themselves, are concerned, since other factors were assumed to be controlled.

If the chances had been near or greater than 369 to 1, as will frequently be the case, generalizations should still be made with considerable caution. The use of these formulae is justified only when the sample is random; and the chances given in the table refer only to the allowance to be made for variable errors of measurement and variable errors of sampling. They do not guarantee that the difference is significant in spite of faulty equivalence, poor control of experimental conditions, or even carelessness in computation. In other words, the difference is to be regarded as significant and worthy of being used as a basis of generalization when the following conditions have been satisfied:

1. The groups are equivalent at the start of the experiment, or the departures from equivalence have been shown to be insignificant.
2. All of the non-experimental factors have been controlled during the course of the experiment, or failure to control has been shown to be negligible in effect on the difference in gains.
3. The measuring instruments used have been shown to possess high validity. Measures of achievement have been secured not only for specific abilities but for general abilities as well.
4. The testing conditions have been the same for both experimental and control groups, or systematic errors of measurement resulting from failure to secure identical testing conditions have been shown to be insignificant.
5. The sample has been shown to be random, i.e. selected without bias from the population to which the generalization is applied.
6. The difference of the gains is equal to or greater than 2.78 times the standard error of difference due to variable errors of measurement and sampling.³¹

When the above conditions have not been met, the conclusions must be appropriately restricted. In addition to making allowance for

$$\sigma_{(s+m)D} \quad \text{OR} \quad P.E._{(s+m)D},$$

³¹See page 60.

it is necessary to allow for the estimated effect upon the difference between the gains due to failure to secure the conditions listed above.

2. *Generalizing from a sample that is not random.* As pointed out on page 71, the experimenter usually works with groups that are not random samples of a larger population and, hence, must estimate rather than calculate, the allowance to be made for probable non-representativeness of the groups. For making this estimate, no specific rules can be stated. As a general procedure the experimenter should consider all available evidence relative to the traits of the groups concerned. For example, the intelligence test scores will be known, and the experimenter should show how the mean and standard deviation of these scores compare with the corresponding measures of the larger population. If the available evidence indicates that the groups are highly representative of the larger population, he may generalize with considerable confidence; if the evidence indicates that the groups are not reasonably representative of the larger population, he must refrain from generalizing or appropriately limit his statements.

Concluding statement. The preceding discussions of the significance of differences in gains should make it clear that caution must be exercised in interpreting a small difference and that the interpretation cannot be accomplished by the application of any formula or group of formulae. In general, it is necessary to inquire carefully and critically into the conditions of the experiment; then the best that can be done is to *estimate* the allowance that should be made for imperfections in the data. Since an estimate must be considered only an approximation, it follows that the interpretation of a relatively small difference in gains must be somewhat uncertain. When the difference is relatively large, definite conclusions may be justified, but even in this case they must be restricted to the conditions of the experiment. For example, in the experiment with the project method Collings³² obtained very large differences in gains, much larger than any reasonable estimate of the total allowance for non-equivalence of groups, variable and chance errors of measurement and of validity, and failure to control non-experimental factors, provided zeal is included in the experimental factor. Hence, he is justified in asserting that under the conditions of the experiment the project method as applied is distinctly superior to usual methods of teaching as exemplified in the control schools. He is not justified in any statement, except with appropriate qualifications, concerning the relative merits

³²See page 36.

of the project method in general or as it might be applied by other teachers.

In regard to generalizing from an experiment, it should be noted that application of the formula:

$$\sigma_{(s+m)D} = \sqrt{\sigma^2_{(s+m)\text{Gain E}} + \sigma^2_{(s+m)\text{Gain C}}}$$

is justified only when it is reasonably certain that any non-representativeness is due to chance. If factors other than chance may have operated, this formula cannot yield the allowance that should be made for non-representativeness. Furthermore, it should be noted that this formula should be used only when generalization is attempted. If the conclusions are restricted to the groups of pupils concerned, its use is superfluous.

CHAPTER IV

A CRITICAL EVALUATION OF EXPERIMENTAL STUDIES RELATING TO SUPERVISED STUDY

In the following description of the requirements for a precise experiment it is assumed that the objective of supervised study is not primarily that of increasing achievement in the particular school subject studied, but instead is that of engendering habits of study which will increase the effectiveness of the learner's efforts generally, not only at the time of supervision, but in the future. Hence, the purpose of the experiment is to evaluate, in terms of the attainment of this objective, the group of techniques employed by the teacher during the supervised-study period by comparing the attainment of the supervised pupils with that of pupils whose study is characterized by a different type of supervision, or more usually, by the absence of supervision.

Requirements for the conduct of an experiment in supervised study.

1. *Specification of supervised study as an experimental factor.* The term "supervised study" is used to designate a variety of instructional procedures and combinations of instructional procedures employed by the teacher in connection with the studying of his pupils and designed to guide them to the acquisition of efficient study procedures. These instructional procedures include making the assignment; giving general rules for study; making suggestions for the doing of learning exercises; answering questions that members of the group desire to ask; inspecting work as it is being done, and calling attention to faults; giving direct assistance; suggesting supplementary learning exercises to individual pupils; providing aids for study such as reference books, maps, pictures, etc.; and maintaining a place suitable for study. Definition of supervised study as an experimental factor requires detailed specification of the particular instructional procedures to be employed during the study period of the experimental group. Where comparison is to be made between two types of supervised study, the instructional procedures for the study periods of both groups must be specified in detail. This means that the supervised-study procedure to be employed must be described in writing, or at least a detailed record must be kept of what was done.

2. *Equivalent groups.* The groups of pupils used in the experiment should be equivalent in all respects that affect methods of study and

learning in the subject to be studied. This requirement can be approximated by pairing pupils on the basis of intelligence test scores and then comparing the groups thus formed with respect to chronological age, previous achievement in the school subject, and study habits. If the differences between the means and the standard deviations of the groups with respect to these three characteristics are relatively small, the groups may be considered approximately equivalent. It is desirable that the groups also be approximately equivalent with respect to personality traits, physical conditions, sex, and race.

3. *Control of teacher factors.* Control of the teacher factors involves maintaining the same status of all factors under this head in both the experimental and the control groups, except the procedures specified as supervised study and the corresponding procedures employed with the control groups; or if the same status is not maintained, the amount of non-equivalence must be measured and its effect upon the engendering of study habits and the experimental learning must be determined. The factors whose control in supervised-study experiments appear to be the most important are: (1) instructional techniques employed during the recitation period, especially those relating to the assignment; (2) motivation techniques; (3) skill of teacher in carrying out instructional techniques and classroom-management procedures; (4) zeal of teacher; (5) personality traits. In addition, care should be exercised to avoid marked differences in the minor teacher factors.

4. *Control of general and extra-school factors.* The important factors under this head are: (1) materials of instruction, (2) environment in which pupils study, and (3) time per day devoted to study. The pupils in the control group should be given as convenient an access to reference books, maps, charts, and other aids to study as the members of the experimental group; and desks, chairs, light, heat, ventilation, and other aspects of the study environment should be identical for both groups. This means that the pupils of the control group should study in a classroom during the period devoted to supervising the study of the other groups and in the presence of their teacher, but without any other supervision¹ than that required to maintain order and to hold them to their tasks. The other general school factors and the extra-school factors should be controlled by means of the methods described on pages 45-47 and 49-50.

5. *Measurement and interpretation of differences in gains in achievement and in the acquisition of study habits.* Equivalent forms

¹Unless two types of supervised study are being compared.

of an achievement test of high reliability and validity should be used as initial and final tests. An objective measurement of study habits should be made at the beginning and end of the experimental period so that the relative acquisition of study habits may be compared. Both groups should study in the unsupervised fashion for a time after the close of the experiment. After a period of some months the pupils should be tested for achievement in the subject matter in which they are at the time engaged, and for the possession of good study habits. Superiority in achievement for the experimental pupils, if shown after this lapse of time, will constitute a weighty argument for the supervised-study procedures. Retention of study habits, after the removal of supervision, as indicated by superiority in achievement and by their direct measurement means the attainment of the objective of supervised study and, in consequence, a favorable conclusion for the experimental factor.

In the interpretation of differences in gains in achievement, or in the acquisition of study habits, the techniques described in Chapter III of this bulletin should be used.

6. *Generalization.* The groups of pupils should be representative of the population to which the generalizations are to be applied. Unless the pupils involved in the experiment have been selected by a process of random sampling, evidence should be presented to show the extent to which the groups are typical. If the groups are not typical, the conclusions must be restricted accordingly.

The teachers selected to conduct the experiment should be typical of those teaching the subject in general, and the instructional and classroom-management procedures they use in the recitation should be representative of sound educational practice.

Descriptions of experiments in supervised study. 1.² Earhart, in 1906, conducted the pioneer experimentation in this field.³ The first of the experiments described in her monograph was, in a sense, controlled. Five sixth-grade and four seventh-grade classes were trained in finding the subject of the lesson, in organizing the subject-matter, in verifying the authors' statements, and in supplementing the lesson. The possession of the abilities to perform these activities was measured, before and after training, by means of sample

²The experiments described are all those which have been reported in the more easily accessible literature of education. In addition to those described, several experiments have been reported merely as unpublished masters' and doctors' theses in education.

³Earhart, L. B. "Systematic Study in the Elementary Schools," *Teachers College, Columbia University Contributions to Education*, No. 18. New York: Bureau of Publications, Teachers College, Columbia University, 1908, p. 67-97.

The second of the experiments described is also reported in:
Earhart, L. B. "Experiment in Teaching Children How to Study," *Education*, 30:236-44, December, 1909.

lessons for which the pupils were asked to indicate the steps they would take in finding the subject, and so on. The measurement was completed by a series of tests in which the pupils were requested to perform the activities listed above. Fifteen sixth-grade and four seventh-grade classes that had not received the training were given the same tests and the results compared with those of the trained classes. The second of the experiments reported in the monograph was not controlled. A single group of twenty fourth-grade pupils was used. The supervised study consisted of giving the pupils an aim for each of sixteen lessons and encouraging them to ask questions as they studied literature in the presence of the teacher. No quantitative results are given for this second experiment. The following conclusion is stated for both of the experiments:

The results of this series of lessons, coupled with the results of the tests in geography given to the sixth and seventh grades, indicate strongly that pupils in the elementary schools in grades including the fourth as well as higher classes, are able not only to employ the factors of logical study, but also that by means of systematic efforts, they can be made to improve in their employment of them.⁴

2. Breslich⁵ is to be credited with having conducted the first carefully controlled experiment in supervised study. Two groups of high-school pupils of unreported size were selected of approximately equal ability in algebra as shown by their final examination grades of the preceding semester. The control or unsupervised group recited in the traditional manner for forty-five minutes and prepared the advance assignment during the study hour or at home. The experimental group recited during one period and used the next for study in the presence of the teacher who employed the following instructional procedures: passing about the room, watching the pupils at work, offering suggestions, giving no help until a serious effort had been made by the pupil, and stopping the whole class for discussion of mistakes that might become general. The technique used by the teacher is characterized by the adaptation of the instructional procedures to meet the needs of the moment or of the individual. The results given in terms of school grades at the end of fourteen weeks show the supervised group to have achieved slightly more. It is stated that the poorer pupils profited most and that the brighter students seemed to have suffered some loss. A rotation of the groups at the end of the fourteen weeks and a continuation of the experimentation for six lessons resulted in the former supervised group maintaining its superiority of

⁴Earhart, *op. cit.*, p. 79.

⁵Breslich, E. R. "Teaching High-School Pupils How to Study," *School Review*, 20:505-15, October, 1912.

achievement, thus indicating that study habits acquired under supervision continued to function after the supervision had been removed.

3. Minnick⁶ conducted an experiment in supervised study in which two groups of eighteen tenth-grade pupils in plane geometry were approximately equivalent with respect to means and measures of variability of school grades in algebra. The experimental group met for forty minutes of recitation and forty minutes of study of the advance assignment, during which the teacher answered questions and made suggestions. Additional work was given to the brighter students to keep them busy during the hour. The control group recited and studied in the usual fashion. Both groups had the privilege of asking the instructor questions during his consultation period. After a period of fifteen weeks, the results were reported in terms of daily recitation grades, grades on six-weeks tests, and grades on the final examination. On the basis of the consistent superiority of the supervised group as shown by these grades, the following conclusion is reported: ". . . students under such instruction not only master the text more thoroughly but are more able to take the initiative in new work than are the students under the unsupervised plan."⁷

4. In an uncontrolled experiment by White,⁸ all the classes, with the exception of those in shop or laboratory courses, of a four-year high school were given thirty minutes of directed study during the sixty-five minute divided periods. The chief supervised study procedure was that of help and encouragement of backward pupils. After a trial of eight weeks, it is stated that the use of supervised study resulted in lower costs, less withdrawal from school, more work for the principal, and dislike for the plan by some of the teachers and parents. Although the author states early in his report that, "It is worth much more to a pupil to have an instructor teach him how to study than to teach him Latin or algebra,"⁹ nothing is said as to whether or not this belief was justified by the experiment.

5. Dunn¹⁰ used two groups of eleven fourth-grade pupils in language. Approximate equivalence was determined on the basis of scores on a standardized language test. The pupils of the supervised group were given directions for outlining and for studying by wholes rather than by parts during the study period of thirty minutes. The control pupils studied in a classroom for the same length of time

⁶Minnick, J. H. "An Experiment in the Supervised Study of Mathematics," *School Review*, 21:670-673, December, 1913.

⁷*Ibid.*, p. 673.

⁸White, E. A. "An Experiment in Supervised Study," *Educational Administration and Supervision*, 1:257-62, April, 1915.

⁹*Ibid.*, p. 258.

¹⁰Dunn, G. A. "The Value of Supervised Study," *Teachers College Record*, 13:430-437, November, 1917.

without such supervision. At the end of the experimental period of four weeks, the recitation grades of the pupils were compared, and on the basis of this comparison the following conclusion is reported: ". . . there is a decided difference in the results produced by the two different methods of study, and, furthermore, the data would suggest that the directed-study period is of vastly more value to children than is the undirected-study period."¹¹

6. Heck¹² has reported an experiment in arithmetic in which 141 fifth-, sixth-, seventh-, and eighth-grade pupils participated. The pupils were divided into two groups of approximately seventy pupils each. The method used is best explained in the words of the experimenter, "The half-year grades tested were 5A, 6B, 6A, 7B, and 8 . . . These six grades were divided into two groups, as equally balanced as possible; the first group was composed of grades 8, 7B, and 6B, and the second group of grades 7A, 6A, and 5A."¹³ During the study period the pupils of one of these groups worked out the examples in computation and reasoning contained in one of the forms of the Curtis Arithmetic Tests. The pupils of the other group worked out the same problems at home, without help, the same evening. On the following day the procedure was reversed for the two groups and another form of the same tests was used. The combined results for both groups indicated no significant differences between school environment and home environment as a factor in study. Somewhat the same procedure was employed by the author during an experiment in English composition in which ninety-five high-school students of all classes participated. One group wrote a theme at school; the other wrote a theme at home. The groups were reversed, and a second theme was written. Again, no significant difference was found between school and home environment as a factor in study.

7. Breed¹⁴ directed a cooperative experiment in fourteen schools. In each, a group of ninth-grade pupils in a given subject was divided into an experimental and a control group on the basis of scores on an informal preliminary test, or on previous school marks. Each of the pairs of groups was taught by a single teacher. The experimental factor was twenty minutes of directed study during the fifty-minute divided period and had as its essential element guidance of pupils in applying study rules. The supervised-study groups recited for thirty minutes, while the unsupervised, or control, groups recited for fifty

¹¹Dunn, *op. cit.*, p. 437.

¹²Heck, W. H. "Comparative Tests of Home Work and School Work," *Journal of Educational Psychology*, 10:153-62, March, 1919.

¹³*Ibid.*, p. 153.

¹⁴Breed, F. S. "Measured Results of Supervised Study," *School Review*, 27:186-204, 262-284, March, April, 1919.

minutes each day. The control group studied in the traditional fashion. After six weeks the groups were reversed, and the experiment continued another six weeks. The conclusions state that, on the average, supervision of study resulted in less efficient learning of algebra and English composition, but in more efficient learning of Latin. Breslich is substantiated in that supervised study favored the poorer students, but hindered the brighter. For this reason, Breed advocates a differential plan of study supervision. The results of this comprehensive and fairly well controlled experiment are interesting in that they fail to confirm the claims made for supervised study by many enthusiasts of that time.

8. Heckert¹⁵ has reported an experiment using a modified double-period plan in English composition in which no definite portion of the period was devoted to supervised study. Two equivalent groups of seventeen pupils each were selected on the basis of their mental-test scores and ratings of their compositions. The instructional procedures of the supervised-study group were as follows: diagnosis of individual difficulties in composition writing, aid in overcoming these difficulties, analysis of compositions, instruction in outlining and in the use of outlines, and attempts to arouse enthusiasm for composition writing. Although the instructional procedures of the control group are not described, it is possible that they may have included some of the above. The statement is made that both groups had the same teacher for the recitation, but that the teacher was aided during the study period of the supervised group by the author. An attempt was made to arouse an equal amount of enthusiasm for composition writing among the control pupils who were also prevented from devoting more time to study than those of the supervised group, or from receiving assistance at home. At the end of twenty-five periods of one hour each the achievement of these pupils was again determined by means of composition ratings. The conclusions state that supervised study in English composition is "eminently worth while" and that ". . . under fairly skillful direction the brighter children of a supervised class not only make better progress than the brighter members of equal ability of the unsupervised group but that they also make better progress than the slower children of the supervised group."¹⁶

9. Beauchamp¹⁷ has reported the results of an experiment in physical science in which two approximately equivalent groups of twenty-

¹⁵Heckert, J. W. "The Effects of Supervised Study in English Composition," *Journal of Educational Research*, 5:368-80, May, 1922.

¹⁶*Ibid.*, p. 378.

¹⁷Beauchamp, W. L. "A Preliminary Experimental Study of Technique in the Mastery of Subject-Matter in Elementary Physical Science," *Supplementary Educational Monographs*, No. 24. Chicago: University of Chicago Press, 1923, p. 47-87.

six pupils each were selected on the basis of age, intelligence, reading rate, and reading comprehension. Both of the groups received instruction in methods of study during the first of six "units" of work. At the end of the first unit, which continued for four weeks, a change was made in the instructional procedures. The supervised-study group received specific instruction in studying a paragraph to determine its central idea, in finding and answering questions on the material assigned, in reading through the entire assignment before beginning an analytical study of its parts, and in solving thought questions. The unsupervised group continued to use the general methods of study suggested to them during the first unit of work. The amount of time given to study, the material studied, and the environment for the studying were the same for both groups. At the end of each unit of work the achievement of the pupils was evaluated by means of written reports, completion tests, and thought questions on the material covered in the unit. The acquisition of study habits of the supervised and unsupervised pupils was estimated by analysis and comparison of their study notes. The reading ability of the pupils was tested again at the end of the seven months of experimentation.

The analysis of the study notes of the two groups of pupils showed that the pupils of the supervised group had acquired more effective habits of study. The other conclusions drawn by Beauchamp are given below:

1. Specific training in finding the central thought of a paragraph, determining the questions one must be able to answer in order to obtain an adequate understanding of a topic, and reading an entire block of material through for its general plan, results in a more thorough comprehension of the subject-matter than undirected study on the same material.

2. Specific training and practice in answering thought questions based on the application of some scientific principle are more efficient than incidental training in answering thought questions.

3. Training the pupil to make various types of analyses of the subject-matter increases the ability of the pupil to interpret and reproduce what he reads.

4. The gain in rate of silent reading is greater if the pupil is not required to make an analysis of what he reads.¹⁸

10. Brown and Worthington¹⁹ have reported the results of a cooperative experiment in algebra, English, and United States history in which five high schools participated. Seven pairs of equivalent groups, varying in size from twenty-three to thirty pupils, were selected on the following bases: school subject marks (three pairs of groups); intelligence quotients (one pair of groups); intelligence scores, ratings

¹⁸Beauchamp, *op. cit.*, p. 87.

¹⁹Brown, W. W. and Worthington, J. E. "Supervised Study in Wisconsin High Schools," *School Review*, 32:603-12, October, 1924.

on composition and spelling scales, and scores on a standardized achievement test (one pair of groups); intelligence scores and school-subject marks (one pair of groups); and intelligence and reading scores (one pair of groups). Each of these pairs of groups was taught by a single teacher. The supervised study consisted of "directing the mental operations, whether they are reciting, being assigned a lesson, or working out an assignment" during a sixty-minute period divided approximately into "twenty minutes for discussion and review of previous work through recitation, examinations, etc.; fifteen minutes for assignment of new problem; twenty-five minutes for working out new problem."²⁰ The control group spent fifteen minutes less time per day with the teacher, since they met for a forty-five minute recitation. Neither group was limited in time that might be spent in study outside of class.

The achievements of the pupils were evaluated in various ways: school marks (three pairs of groups); ratings on composition scale (one pair of groups); true-false test, general examination prepared by teacher, (one pair of groups); standardized test and semester marks (one pair of groups); and standardized test (one pair of groups). At the end of an experimental period of one semester the following conclusions were derived from the results:

1. . . . two pairs, the algebra classes in School C and the English classes in School D, showed rather definitely that greater progress was made in the supervised-study groups; four pairs showed slight variation in progress, favorable to supervised study; and one pair, the United States history classes in School A, indicated that the recitation plan was superior as a method of instruction.
2. Three supervised-study classes had fewer failures than the parallel recitation classes; one had the same number; one had more; and in the other two cases the number of failures could not be determined from the data submitted.
3. In general, then, the objective data indicated a superiority of the supervised-study plan over the recitation plan as a method of instruction. However, the objective data were not conclusive.²¹

11. Johnson²² has reported an experiment in eighth-grade arithmetic in which two groups approximately equivalent on the basis of intelligence scores and past scholastic grades were used. The supervised study was a composite of the following elements: a sixty-minute period, of which at least twenty-five minutes were given to study of the next day's assignment in the presence of the teacher; ability grouping with differentiated assignments in the supervised class; and suggestive directions, given in mimeograph form and referred to dur-

²⁰Brown and Worthington, *op. cit.*, p. 604.

²¹*Ibid.*, p. 612.

²²Johnson, A. W. "The Effectiveness of Directed Study," *Elementary School Journal*, 26:132-36, October, 1925.

ing the study period. The members of the control group were not divided on the basis of ability, recited in the traditional manner, and studied during the study period or at home. At the end of six weeks the groups were reversed and the experiment continued for another six weeks. The results are expressed in terms of the number of problems attempted and the number of problems correctly solved on tests given at intervals of three weeks. The following conclusion is derived from these findings: ". . . that in order to train in efficient and economical study habits the studying must be done under the teacher's immediate direction."²³

12. Reeder²⁴ has reported an experiment in seventh-grade geography in which two groups of twenty-three pupils each were rotated to secure equivalence. In addition, the groups were considered equivalent because the pupils had had no previous acquaintance with the textbook studied during the experiment. No initial achievement test was administered, first, because the pupils had no specific knowledge of the subject-matter to be studied on which they could be tested, and second, because the administration of an initial test would give the control pupils some direction in what to study.²⁵

The study of the experimental pupils differed from that of the control pupils in that the former received mimeographed sheets of study questions with each assignment. These study questions resembled very much in form the items on various kinds of new-type tests. The pupils of both groups studied in the same school environment and had access to the textbook only during the study period. At the end of two study periods, and after twenty minutes of review of the text and study sheets by the experimental pupils, and of the text alone by the control pupils, the books and study sheets were collected and a true-false test administered. The groups were rotated at the end of the week^{25a} and the experiment continued so that the former unsupervised group received supervision and the former supervised group went unsupervised. The results were combined for two week units. That is to say, the means of the scores on the final true-false tests for the unsupervised weeks of both groups were added together, and the result subtracted from the combined means for the supervised weeks of both groups. The duration of the entire experiment at the Speyer School was six weeks. Reeder supplemented his results with similar experiments at other schools, one of which lasted

²³Johnson, *op. cit.*, p. 135.

²⁴Reeder, E. H. "A Method of Directing Children's Study of Geography," *Teachers College, Columbia University Contributions to Education*, No. 193. New York: Bureau of Publications, Teachers College, Columbia University, 1925. 98 p.

²⁵*Ibid.*, p. 33.

^{25a}The classes met for only three periods per week.

six weeks, and three which lasted but two. The data were interpreted for each of the two week units of all of the experiments in terms of the experimental coefficient, and summaries are given for the results at different schools and for all of the experimentation. Reeder reports a final experimental coefficient of 1.62 indicating that the chances in favor of the superiority of his method of supervision as compared with its absence are as 65,000 to 1.

13. Douglass²⁶ has reported an experiment in which ten pairs of groups, averaging fourteen pupils each, were selected and carefully paired on the basis of age and a composite of intelligence test score and achievement test score. Some of the groups were taught by members of the regular high-school staff which consisted "largely of mature, progressive, and somewhat superior teachers."²⁷ The majority of the groups were taught by practice teachers. The author claims that there was nothing about the teaching staff, the personnel of the groups used in the experiment, or the equipment of the school that would render the results not typical of those which could be obtained in the ordinary high school.

The problem was the determination of the relative effectiveness of the study-recite sequence in supervised study as compared with the recite-study sequence. The pupils of each pair of groups were subjected to the same supervised-study procedures during the portion of the period devoted to study. They had the same instructional procedures and materials during the recitation. The only difference was that in one group the pupils recited and then studied their lesson for the next day, while in the other group they studied the lesson first, and followed their study with immediate recitation. In order to equate teacher factors and those of room environment, teachers and rooms were exchanged at the mid-point of the experiment. At the end of eleven weeks, final achievement tests were administered. These tests were similar in form and content to the subject-matter tests administered at the beginning of the experiment, and in one case the same test was repeated. The coefficients of reliability were determined for all of the tests used and were found to range from .614 to .943. In the interpretation of data, appropriate statistical procedures were used. The following conclusions are reported in the monograph:

1. It cannot be said that either an R-S or an S-R sequence is more effective than the other for all classes or types of work.

²⁶Douglass, H. R. "The Experimental Comparison of the Relative Effectiveness of Two Sequences in Supervised Study," *University of Oregon Publications*, Vol. 1, No. 4. Eugene: University of Oregon, 1927, p. 173-218.

For a briefer account, see:

Douglass, H. R. "An Experimental Investigation of the Relative Effectiveness of Two Plans of Supervised Study," *Journal of Educational Research*, 18:239-45, October, 1928.

²⁷*Ibid.*, p. 173.

2. The S-R sequence is more effective in classes in history and social science and in literature than the R-S sequence.
3. The R-S sequence is more effective in classes in mathematics and science, though the superiority may not be manifest in each class.
4. The relative effectiveness of the S-R sequence for history and social science and English classes is greater for classes in grades above the eighth than for seventh and eighth grade classes.
5. Teachers' judgments of the relative effectiveness of two methods are not to be taken too seriously.
6. Neither sequence operates to produce greater variability in progress than the other, generally or in any particular subject or school grade.
7. Neither sequence is peculiarly favorable for more capable or for the less capable student.²⁸

Evaluation of the techniques employed in the conduct of the experiments in supervised study. 1. *Specification of supervised study as an experimental factor.* There has been little agreement among the experimenters in this field as to just what particular procedures constitute supervised study. An examination of the reports shows that some have emphasized guidance of pupils while studying as the chief procedure. An example of this is to be found in the experiment of Breslich. The teacher passed about the room watching pupils at work, making suggestions but rarely answering questions directly, giving no help until a serious effort had been made by the pupil, stopping the whole class when mistakes were discovered which might become general, and adapting the guidance to meet the needs of the moment or the individual.²⁹ Other experimenters have emphasized instruction in the techniques of learning, or methods of study as the chief procedure of supervised study. Beauchamp gave the experimental group training in determining the central idea of a paragraph, in organizing their thinking about the central idea, in finding and answering questions relative to the material assigned, in reading through a whole assignment before beginning an analytical study of its parts, and in solving thought questions.³⁰

In many of the reports of the experiments in this field the descriptions of the procedures used are inadequate. For example, instruction in methods of study was the chief procedure employed by the teachers in the cooperative experiment of Breed, but the report does not present information with respect to just what these methods were other than that they were taken from Whipple.³¹ The results of the different experiments may not be compared, or a general conclusion to the experimentation in this field synthesized, because of the variety, complexity, and inadequacy of description of the experimental factors. In

²⁸Douglass, *op. cit.*, p. 218.

²⁹Breslich, *op. cit.*, p. 508.

³⁰Beauchamp, *op. cit.*, p. 49.

³¹Breed, *op. cit.*, p. 196.

many cases it is impossible to ascribe the results obtained to specific supervised-study procedures.

2. *Equivalence of groups.* A variety of techniques were employed in the *controlled* supervised-study experiments in an effort to secure equivalent groups. These techniques range from shifting pupils so that the average of the two groups were as nearly the same as possible with respect to school-subject marks of the preceding semester, or scores on an informal preliminary test, in the experiment of Breed³² to the pairing of pupils on the basis of chronological age and a composite of regressed intelligence and achievement test scores in the experiment of Douglass.³³ Beauchamp recognized such criteria as intelligence test scores and scores on informal reading rate and comprehension tests in showing the equivalence of his groups.³⁴ Reeder assumed that his groups were equivalent because they were rotated at the mid-point of the experiment and because the pupils had no specific initial knowledge of the subject-matter to be studied.

Beauchamp is the only experimenter who came near to recognizing the importance of study habits as one of the criteria of equivalence.³⁵ Douglass³⁶ approached closest to the requirements of a precise experiment in that he made use of three of the criteria, or characteristics of pupil material, and did so on the basis of regressed scores.³⁷ None of the experimenters paired pupils on the basis of intelligence test scores and later checked the equivalence thus secured with respect to previous achievement, chronological age, study habits, personality traits, physical condition, sex, and race in the manner suggested in the discussion of the control of pupil factors. It is probable that the earlier experiments should not be condemned too severely for this, since these techniques had not been developed.

3. *Control of teacher factors.* Some of the techniques employed for the control of instructional procedures are excellent. For example, Breed provided the teachers who participated in his experiment with mimeographed copies of the general directions to be followed in the conduct of the experiment.³⁸ Douglass insured a clear understanding on the part of the teachers of the procedures to be employed by holding a meeting at which the procedures were explained and questions relative to them answered.³⁹ The teachers were also given a summary of instructions for the conduct of the experiment and were asked to

³²Breed, *op. cit.*, p. 192.

³³Douglass, *op. cit.*, p. 177-83.

³⁴Beauchamp, *op. cit.*, p. 50-54.

³⁵Beauchamp, *op. cit.*, p. 51.

³⁶Douglass, *op. cit.*, p. 177-83.

³⁷See page 87.

³⁸Breed, *op. cit.*, p. 191.

³⁹Douglass, *op. cit.*, p. 184-185.

keep an experimental log in which was kept a record of all absences, interruptions of class work, and special distractions which might influence the progress of the pupils. Reeder⁴⁰ eliminated instructional procedures by administering supervision of study in the form of mimeographed sheets of study questions with each assignment. The control pupils, of course, did not receive these study questions.⁴¹ An attempt to control zeal and effort is indicated in the following quotation from Heckert:

In both groups, the teacher attempted to arouse as much enthusiasm for the work as possible in order that every child might do his best. In this she succeeded for we never had keener interest in the writing of compositions.⁴²

There is abundant evidence of failure to control certain teacher factors in many of the experiments. In the experiment of Heckert⁴³ the teacher was assisted by the experimenter in the administration of the supervised-study procedures. It is logical to assume that the pupils were stimulated not only by the supervised-study procedures, but also by the superior skill and presence of the experimenter. Among the directions given the cooperating principals by Breed is the following:

Select for the supervision of study teachers who are known to be interested (a) in undertaking the experiment, and (b) in teaching pupils how to study.⁴⁴ This suggestion could not help but introduce faulty control of the teacher factor of zeal and perhaps also of skill. *The teachers were selected for their bias in favor of the experimental factor.* The critical reader secures the impression from a careful examination of the reports of these experiments that the zeal of the teachers for the novel supervised-study procedures was in most cases an uncontrolled factor of sufficient influence to produce the apparent superiority of the experimental factor.

While the teachers in the experiment of Breed were selected for their interest in supervised study, there is no other evidence, in this case, that they were not typical of high-school teachers in general. The teachers of Douglass' experiment do not seem to be representative of the profession, since they were either practice teachers or regular members of the staff of a university high school.⁴⁵ The same statement may be made with respect to the experiments of Breslich and Beauchamp, although no practice teachers were used. In the rest of the reports the information concerning teachers is too meager to permit judgments relative to the representativeness of the teachers.

⁴⁰Reeder, *op. cit.*, p. 33.

⁴¹*Ibid.*, p. 33.

⁴²Heckert, *op. cit.*, p. 371.

⁴³*Ibid.*, p. 371.

⁴⁴Breed, *op. cit.*, p. 192.

⁴⁵Douglass, *op. cit.*, p. 175.

4. *Control of general and extra-school factors.* Evidence is presented in the reports of Reeder and Douglass to show that care was taken to see that the experimental and control pupils had identical materials of instruction. For example, Douglass states in the instructions to teachers: "Not only should the two sections cover the same material during the experimental period, but they should be kept together and should progress at the same rate."⁴⁶ Reeder permitted the experimental and control pupils to have access to the textbook only during the study periods.⁴⁷ It is probable that some of the other experimenters exercised similar care, although no information was given to prove that this was the case. The failure to have both groups of pupils study in comparable⁴⁸ environments, as is evident in the experiments of Breed,⁴⁹ Breslich,⁵⁰ Brown and Worthington,⁵¹ Johnson,⁵² and Minnick,⁵³ is likely to have caused variation in this factor. The experimenters just referred to are to be criticized for not controlling adequately the factor of materials of instruction and for failure to eliminate the general and extra-school factors inherent in non-comparable study environments.

The study environment of the experimental pupils, in that they studied in the presence of their teacher, was not comparable, in other respects, to the study environments of the control pupils who studied in a study hall, or at home. The experimental pupils probably were stimulated by the mere presence of the teacher. The control pupils who studied in a study hall had this stimulation to a lesser amount, but the control pupils who studied at home had it, or didn't have it, according to the character of their parents. Another aspect of the failure to control the study environment is its effect on the time factor. Where the experimental and control pupils studied in comparable environments, the time factor was also controlled. In the experiments of Beauchamp,⁵⁴ Douglass,⁵⁵ Dunn,⁵⁶ Heckert,⁵⁷ and Reeder⁵⁸ the same amount of time was set aside each day for study on the part of the experimental and control pupils. In the experiment

⁴⁶Douglass, *op. cit.*, p. 185.

⁴⁷Reeder, *op. cit.*, p. 33.

⁴⁸When the control pupils study in a classroom in the presence of their teacher, they study in an environment that is comparable to those of the experimental children, providing the teacher gives no other supervision than that required to maintain order and to hold them to their tasks. See the requirements for a precise experiment in supervised study, p. 73.

⁴⁹Breed, *op. cit.*, p. 191-92.

⁵⁰Breslich, *op. cit.*, p. 508.

⁵¹Brown and Worthington, *op. cit.*, p. 604.

⁵²Johnson, *op. cit.*, p. 133.

⁵³Minnick, *op. cit.*, p. 671.

⁵⁴Beauchamp, *op. cit.*, p. 52.

⁵⁵Douglass, *op. cit.*, p. 185.

⁵⁶Dunn, *op. cit.*, p. 433.

⁵⁷Heckert, *op. cit.*, p. 377.

⁵⁸Reeder, *op. cit.*, p. 33.

of Heckert⁵⁹ the control children studied at home, but care was taken to prevent them from spending more time at study than the supervised pupils. In the experiment of Reeder⁶⁰ both groups studied for the same amount of time at school. Additional study was prevented by collection of the textbooks at the close of the study period. With the possible exception of the experiment of Reeder, one cannot be sure that the members of the *supervised* group did not do some of the studying outside of the classroom. In the experiment of Johnson the failure to control a factor of school organization is indicated by the information that the experimental pupils were divided into ability groups, while the control pupils were not.⁶¹

Failure to control the general and extra-school factors satisfactorily might be justified by the experimenters on the grounds that they were comparing supervised-study procedures with the traditional absence of supervision and, for this reason, should not have had the control pupils study anywhere else but in the study hall or at home. There is some justification in this view, since the conditions of experimentation are more comparable to ordinary school conditions. However, such conclusions are not as valuable as results secured under more precise control of conditions. In the precise experiment the results may be ascribed to the supervised-study procedures alone, while in the experiments described as lacking this precision, the difference in gains must be ascribed to a complex of factors including supervised-study procedures, instructional materials, presence of the teacher, and home conditions.

Evidence has already been presented of failure to control important non-experimental factors. For example, Heckert did not recognize that he was introducing an uncontrolled experimental factor by participating in the instruction of the supervised pupils. Johnson did not recognize the presence of an uncontrolled factor when the experimental group was divided into ability groups, while the control group was not. Minnick is to be criticized for permitting his control pupils to consult the teacher during his conference hours, thus obtaining a measure of supervision and at the same time introducing an unrecognized uncontrolled factor. The experimenters who failed to control the time factor also failed to recognize its importance in interpreting the data. It has already been stated that failure to control such important teacher factors as skill and zeal is evident "between the lines" in many of the reports of the experiments. None of

⁵⁹Heckert, *op. cit.*, p. 377.

⁶⁰Reeder, *op. cit.*, p. 33.

⁶¹Johnson, *op. cit.*, p. 132-33.

the experimenters recognized that the superiority shown for the supervised-study procedures may have been due to the enthusiasm of the teacher for new procedures.⁶² None of the experimenters can be credited with recognizing lack-of-control to an adequate extent.

In view of the evidence the potency of the educative factors designated as "important,"⁶³ especially the skill and zeal of the teacher with reference to the experimental factor, it seems reasonable to say that failure to control completely the non-experimental factors should have received explicit recognition in interpreting the difference in gains in each of the experiments in this group.

5. *Measurement and interpretation of differences in gains in achievement and in the acquisition of study habits.* A variety of tests were used to measure achievement in the school subjects in which supervised study was tried as an experimental factor. Brown and Worthington, Heck, Heckert, and Douglass were the only experimenters to use standardized tests, or tests of known reliability.⁶⁴ Beauchamp and Reeder used new-type tests of their own construction but made no effort to show that these tests were reliable. The rest of the experimenters depended on ordinary school marks on recitations, monthly quizzes, or traditional examinations.

Earhart and Beauchamp were the only experimenters to attempt a direct measurement of study habits. The former administered sample lessons with questions relative to procedures that would be employed by the child in studying them; the latter made an examination of the study notes of his supervised and unsupervised pupils. Breed, Breslich, and Johnson sought to measure the acquisition of study habits by measurement of achievement after rotation on the assumption that retention of superiority by the former supervised pupils would indicate acquisition of good study habits. Minnick measured the ability to solve new problems for the same purpose. Brown and Worthington, Douglass, Dunn, Earhart, Heck, Heckert, and White made no effort to measure the acquisition of study habits.

None of the experimenters made use of the formulae⁶⁵ that yield indices of the variable errors of measurement. Reeder⁶⁶ and Douglass⁶⁷ made use of the formula⁶⁸ that gives the combined allowance for variable errors of measurement and for sampling and may be credited

⁶²Douglass presents some evidence to show that the preference of a teacher for a procedure does not mean that the teacher will be more successful with that procedure in terms of pupil achievement. However, he is comparing two procedures for which teachers have not strong preferences or prejudices. See page 87.

⁶³See page 51.

⁶⁴See footnote on page 11.

⁶⁵See pages 61-64.

⁶⁶Reeder, *op. cit.*, p. 89.

⁶⁷Douglass, *op. cit.*, p. 204.

⁶⁸See pages 71-73.

with having recognized and allowed for the variable errors of measurement, even though they did not differentiate the variable errors of measurement from those of sampling. In none of the other experiments was any attention given to the variable errors of measurement. It is probable that the earlier experimenters should not be criticized severely for this omission since the formulae may not have been accessible to them. It is certain, however, that the findings of most of the experiments in supervised study are less dependable because of the neglect to minimize, or account for, variable errors of measurement.

Douglass is the only experimenter in this field who has included reference to systematic errors in his report. He states:

Those familiar with statistical procedure realize that the obtained gains from studies of this sort may be the results of a real difference in the experimental factors, chance errors of measurement and of sampling, or systematic errors from those sources such as might occur in the selection of experimental groups or in the failure to rule out or hold constant some factor favoring one group or the other. As has been pointed out, great care was exercised to prevent the operation of such systematic errors.⁶⁹

No mention is made in the reports of any of the experiments concerning the precautions taken to insure identical testing conditions, unless the above quotation of Douglass is taken to mean that he maintained such conditions. The neglect to mention such precautions leads to the inference that identical testing conditions were not maintained in the other experiments. The probable consequence is that unrecognized systematic errors of measurement have influenced the results in unknown, though probably small, amounts.

No attention was given by any of the experimenters to variable or systematic errors of validity. The significance of this fact becomes apparent when we note that with the exception of the experiments of Earhart and Beauchamp no attempt was made to measure the acquisition of study habits directly. Breed, Breslich, Johnson, and Minnick sought to measure the acquisition of these abilities indirectly, but it is reasonably certain that such indirect measures were grossly lacking in validity.⁷⁰ If the merit of supervised study is judged, as in most of the experiments described, merely on the basis of differences in school achievement in the immediate presence or absence of supervised study, and no account is taken of differences in acquisition or retention of study habits, or of the differences in school achievement long after supervision has been removed, the conclusion arrived at is likely to be unjust. Systematic errors of validity of this nature have not been

⁶⁹Douglass, *op. cit.*, p. 204.

⁷⁰See the discussion on pages 66-69.

adequately accounted for in these supervised-study experiments and because of this fault, alone, many of the conclusions seem undependable.

6. *Generalization.* No evidence is given in the reports of the experiments to show that the groups were selected in a random fashion from the population of school children to which the generalizations are applied. It is evident from an examination of the reports that all the groups were ordinary school classes; therefore, whatever non-representativeness was present may not be ascribed to random sampling. It is probable that the groups used by Breed and by Brown and Worthington may be considered more typical of high-school children in general than the groups used in any of the other experiments. Fourteen high schools in cities of various sizes participated in the experiment of Breed,⁷¹ while four high schools in cities of moderate size cooperated in the experiment of Brown and Worthington.⁷² The experiments of Breslich,⁷³ Beauchamp,⁷⁴ and Douglass⁷⁵ were conducted in a university high school. Douglass sought to show that in spite of the character of the high school in which his experiment took place, the pupils were typical. He states:

This school is a laboratory for the school of education of the University. It has the six-year type of organization, comprising grades seven through twelve. Pupils are accepted for registration on the basis of priority of application. Great effort is expended to maintain a representative student personnel. Not more than two-thirds of any one grade may be of one sex. Athletics and other activities are maintained with a view to attracting a representative group of young people. The entrance requirements are identical with the junior and senior high schools of Eugene. The school is in no way a special preparatory school for the University, and the average age and range of age-ability is approximately equal to that of the typical Oregon high school.⁷⁶

Douglass is to be highly commended for presenting this information in the report of his experiment, but the authors of this study cannot feel that he is quite justified in making the following statement:

There is nothing about the student or teaching personnel or about the equipment of the school which would give ground for a belief that whatever experimental results have been obtained could not be expected to be typical of results obtained in the ordinary high school of over two or three teachers.⁷⁷

The fact that the pupils had to apply for admission, pay fees, and attend school in proximity to a university is sufficient to render them somewhat non-representative of high-school pupils in general. Such environmental influences could not help but engender attitudes toward

⁷¹Breed, *op. cit.*, p. 187.

⁷²Brown and Worthington, *op. cit.*, p. 605.

⁷³Breslich, *op. cit.*, p. 508.

⁷⁴Beauchamp, *op. cit.*, p. 50.

⁷⁵Douglass, *op. cit.*, p. 175.

⁷⁶Douglass, *op. cit.*, p. 175.

⁷⁷*Ibid.*, p. 175.

school work that would be different from those of the typical high-school pupil.

The evidence presented in the first of the quotations of Douglass is the most complete that has been given in any of the experiments relative to the degree of representativeness of the pupils. Most of the experimenters have ignored this important aspect of experimentation altogether.

Douglass and Reeder sought to allow for non-representativeness of the groups of pupils by the use of formulae. It is probable that they were partially justified in using these formulae in that recognition was thus given, unknowingly, to variable errors of measurement.⁷⁸ They do not seem to have been justified in using these formulae as a means of allowing for non-representativeness, because their groups were not selected in a random fashion. In each case they were ordinary school classes. Douglass presents arguments to prove that his groups were representative. If they were representative, then the use of the formulae was futile, since the means of representative groups are the same as the populations from which they are drawn. If the groups were not representative, and we have reason to believe they were not, Douglass should have estimated the effect of non-representativeness on his gains and restricted his conclusions accordingly. Reeder and all of the other experimenters should have done the same thing.

The conclusions quoted in the foregoing descriptions of the experiments in supervised study are, for the most part, stated as generalizations. The previous discussion of the shortcomings of these experiments indicates how undependable the data are on which these generalizations are based. None of the experimenters made use of random or truly representative groups, and yet none of the conclusions have been more than slightly restricted because of this.

The dependability of generalizations found in the reports of the controlled experiments to determine the merit of supervised study. An examination of the achievement differences reported in the controlled experiments to determine the merit of supervised study reveals wide variation from positive results in its favor to negative results in opposition. The majority of differences in achievement seem to favor the superiority of supervised study. Some of these positive differences are so large that it would seem probable that if allowances had been made for variable errors of measurement, validity, and sampling, the

⁷⁸It has been shown that the error of difference formulae in which $\frac{\sigma_{\text{dist.}}}{\sqrt{N}}$ is used to determine the values inserted under the radical recognizes both errors of sampling and errors of measurement. See p. 72.

chances would yet be strongly in favor of the significance of the difference. However, many of these highly positive results in favor of supervised study seem questionable. It is all too evident that the majority of the experimenters, or the teachers used by them, were zealous for the experimental factor. In many cases, it is obvious that the pupils of the supervised groups received not only instruction in how to study, but much more instruction in the subject-matter than the unsupervised pupils. The previous discussion of the techniques employed in the experiments described substantiates the contention that the reported differences are decidedly questionable. In certain experiments the negative differences reported after rotation would also seem to favor supervised study. The former supervised pupils retained some of the habits acquired when supervised study was applied to them, and this retention was effective in reducing the difference in achievement after rotation.

The differences reported in the experiments of Breed⁷⁹ and Brown and Worthington⁸⁰ are for the most part small and are approximately equally divided for and against supervised study. The negative differences are difficult to explain. An examination of the report of the experiment of Brown and Worthington showed that the negative differences were reported only for the groups in which intelligence tests were not used to secure equivalence. Where intelligence tests were used, the difference was in favor of supervised study. The groups used in the experiment of Breed were equated merely on the basis of marks on informal preliminary tests prepared by the cooperating teachers or on the basis of their previous semester grades in the school subject. The negative differences may have been due to failure to secure equivalence.

The generalizations based on such differences cannot be more dependable than the differences themselves. Conclusions favorable to supervised study are not applicable to other schools or classes, because one cannot be sure that the difference in achievement was due to supervised study and not something else. Conclusions unfavorable to supervised study do not seem necessarily a condemnation of it, since it is evident that faulty equivalence may have been responsible for the negative differences. Since the conclusions and generalizations of the individual experiments are of this nature, it is impossible to synthesize a general conclusion to all of the experiments. Such a synthesis might well be a summation of errors rather than an approach to truth.

⁷⁹Breed, *op. cit.*, p. 281-83.

⁸⁰Brown and Worthington, *op. cit.*, p. 605-9.

Concluding statement. The preceding critical evaluation of experiments relating to supervised study has revealed the very meager contributions that these studies have produced. In every case a critical examination of the experimental procedure has revealed faults that make the interpretation of the obtained difference in gains uncertain. If we assume that it is desirable for pupils to conform to certain study procedures and that the controls of conduct which will insure the desired conformity can be acquired only as the result of instruction, there arises the problem of determining the most effective plan for giving this instruction. There are several possible plans: supervised study, special course in how to study, distribution of printed or mimeographed rules and directions for study, suggestions and directions given orally in connection with assignments, and incidental instruction. The problem to be attacked is that of determining the relative effectiveness of supervised study in comparison with other promising plans. Unless we assume that supervised study may not be pedagogically sound, an experiment to determine the relative effectiveness of supervised study and no instruction relative to study procedures is an attempt to prove the obvious. To make this assumption seems absurd. Hence, with the exception of the work done by Douglass the supervised-study experiments evaluated in the preceding pages have been attempts to prove the obvious. Their principle value has been the training of the participating teachers in certain procedures for engendering study habits and the stimulation of interest in supervised study. If the experiments had been planned so as to compare two plans of instruction designed to engender study habits, they might have added to our knowledge about how to instruct pupils in regard to methods of study.

CHAPTER V

EXPERIMENTATION AS A PROCEDURE IN EDUCATIONAL RESEARCH

Is educational research in a plateau period? In Chapter I the development of experimentation in the field of education was traced briefly, and statements by a number of writers were quoted as evidence of a rather general belief that this type of educational research has not yielded much in the form of dependable conclusions about the effectiveness of educational procedures. In commenting on these quotations, the suggestion was made that, perhaps, educational experimentation has reached a plateau period of development. The exposition of the procedure of experimentation in Chapters II and III and the evaluation of a group of controlled experiments in Chapter IV provide a partial basis for an answer to the implied question. Before setting forth the judgment of the present writers, statements from a few recent writers will be noted.

Whipple states in regard to experimental techniques “. . . . that it is only recently that research in education has arrived at the employment of some of the most obvious principles of scientific procedure.”¹ Gates and Barr in the following quotations seem to believe that progress is beginning again because of more perfect techniques:

Three years ago Dean Henmon and others deplored what then appeared to be a serious neglect of experimental studies of the learning and teaching process. It is gratifying to say that since that time scientific workers have shown a renewed activity in these fields of research. . . . Within the last two years there have been gratifying advances in the study of the principles underlying efficiency in learning.²

The experimental study of education is passing out of the play, manipulative, or exploratory state. Better acquaintance with experimental methods should bring better research. It seems that there are a number of practices found in reports of careful research workers that should enjoy more general acceptance.³

It is also the opinion of the writer that the quality and the accuracy of statistical writing have improved very greatly in the last few years. One has only to examine the numbers of an educational journal five years ago and compare them with those of the present year to be convinced of this.⁴

¹Whipple, G. M. “The Improvement of Educational Research,” *School and Society*, 26:252, August 27, 1927.

²Gates, A. J. “Recent Advances in Educational Psychology,” *School and Society*, 29:2, 3, January 5, 1929.

An address before the American Association for the Advancement of Science, Section of Education, December 29, 1928.

³Barr, A. S. “Research,” *Journal of Educational Research*, 19:56, January, 1929. (An editorial.) (The editorial goes on to point out “the more obvious practices of careful research workers.”)

⁴Holzinger, K. J. “Accuracy in Calculation,” *The Elementary School Journal*, 29:516, March, 1929.

The number of such statements is less than that of the critical statements quoted on pages 13-15. This suggests that there is no general agreement in regard to the future of experimentation in the field of education. To the present writers it appears reasonably certain that numerous refinements of technique will be introduced and that, consequently, the quality of experimental research will be greatly improved. The ultimate status is another question. Before venturing a prediction, it is desirable to point out certain inherent limitations of the experimental method and certain crucial difficulties.

Inherent limitations of the experimental method. Experimentation in common with other methods of research can never tell us what should be. The object of all research is to test ideas or hypotheses. Experimentation may tell us, perhaps, which of two methods is the better if certain criteria are assumed. Philosophy rather than experimentation must be used to tell whether or not a method should be used at all. The appropriateness of a method depends on the character of the individuals desired as a result of the educative process. For example, research cannot tell us whether or not gifted children should be educated differently than other children until it is decided whether or not it is desirable to have them different. Such questions are matters of value and belong to the field of philosophy rather than to science.

Another limitation of experimentation is that final answers to educational problems may not be desirable. It is evident to anyone who has given the matter critical thought that even the most perfect of learning experimentation, giving the most conclusive results, and considered as having solved the problem, may yet be detrimental when applied to practice. Teaching is a dynamic process, and if the teacher is satisfied to apply an experimentally determined "best" method, year after year and without change, it appears reasonable that in time its effectiveness will be lowered below that of other methods experimentally proven inferior, but which would be used with greater enthusiasm.

Another weakness of experimentation, though perhaps of less inherent nature, is that the results obtained apply to the typical rather than to the atypical child. It may be said that group experimentation tells us what to do for the child who needs help least. The emphasis on averages causes us to tend to neglect the problems of the individual. The feeling that conclusive results may be obtained only with large groups have led us to neglect the study of the individuals who make up these groups. The careful observation of a single individual

may yield more valuable knowledge in regard to the phenomena of learning than hundreds of carelessly performed experiments on large groups. A hundred individuals may be just as unrepresentative a sample of all individuals as a single one.

Crucial difficulties in educational experimentation. Each item in the procedure of experimentation in the field of education involves a difficulty to be overcome, but certain difficulties appear to be much more serious than others. The present writers believe that four may appropriately be designated as crucial. In giving this designation, they recognize that the seriousness of these difficulties varies, but in general the designation appears to be justified.

1. *Definition of experimental factor and adjustment of other educative factors to it.* Precise definition of the experimental factor is essential in order to give definite meaning to the findings. It cannot be very meaningful to prove that Method A is superior to Method B if the investigator can define these methods only by saying that they are the methods carried out in the experiment. Precise definition is not always easy, but reasonable satisfactory statements may be secured by specifying in writing the details of the experimental procedure prior to the beginning of the experiment or by keeping a detailed record of the procedure actually carried out.

There is, however, another aspect of this difficulty. In order to secure results that have a maximum of practical value, it is not sufficient merely to make a precise specification of the experimental factor. The procedure specified must be one that is adjusted to other educative factors in a way that is compatible with sound educational practice. This means that the combination of all educative procedures, both experimental and non-experimental, must be one that is effective. For example, consider supervised study as an experimental factor. It is reasonable to assume that the maximum effectiveness of this instructional procedure will be attained only when it is combined with certain types of assignments, recitational activities, and perhaps textbooks. In other words, the effectiveness of supervised study depends upon the other instructional procedures with which it is combined and not merely upon the techniques of which it consists.

A more striking illustration of this aspect of the difficulty is afforded by an experiment to determine the effect of class size upon pupil achievement. It appears reasonable to say that the teaching of a class of fifty pupils should involve procedures that differ in some respects from those that are most effective with classes of fifteen to twenty pupils. Hence when size of class is made the experimental fac-

tor, it is necessary to have the accompanying instructional procedures adjusted to the size of the particular classes included in the experiment. If this is not done, the experimental factor will not be tested out under optimum conditions, and the findings will have only limited significance. The important problem relating to the effect of class size is, "What will be the effect of organizing a school into large classes rather than into small classes?" rather than, "What is the effect of a few large classes that are balanced by small classes?" In order to have typical large-class conditions, it is necessary to have a school organized into large classes, and this in turn will mean a large pupil-teacher ratio unless the number of classes per teacher is reduced in proportion. Similarly, typical small-class conditions will involve a small pupil-teacher ratio unless the number of classes per teacher is increased in proportion. As usually thought of, large classes are understood to mean a large pupil-teacher ratio and small classes, a small pupil-teacher ratio. Hence in setting up a class-size experiment, it is not sufficient merely to organize a few large classes and a corresponding number of small classes. The pupil-teacher ratio must be comparable to the size of class.

2. *Control of non-experimental factors.* The more important educative factors were identified in Chapter II, and the difficulty of controlling them, especially the less tangible ones such as teacher zeal, is so apparent that an extended discussion is unnecessary here. As pointed out on pages 52-56, control of certain factors has been attempted by employing the rotation method. When employing this method, as well as other procedures for securing control, it is important to make certain that the total instructional situation is compatible with sound educational practice. This requirement is not always easy to satisfy. For example, it is doubtful whether the rotation method is appropriate when the total experimental period is less than two terms or semesters. Certainly, it is not compatible with typical educational practice to rotate the teachers when the total period is only a few weeks. The control of instructional techniques by detailed directions, which are to be followed rigidly, may lead to teaching that is not compatible with sound educational practice because in good teaching there must be adaptation of techniques to the needs and purposes of pupils as they become apparent.

As pointed out on pages 54-55, attempts to control teacher factors by having each teacher instruct an experimental group and a control group will not always be successful. If the requirement of compatibility with sound educational practice is observed, it is not likely

that this procedure will result in control of zeal and effort, which were shown to be important educative factors.

It should be noted that control may be secured by measuring the differences that exist and allowing for them in interpreting the difference in the gains in achievement. Although this method will seldom secure precise control, because we lack instruments for measuring most of the important educative factors, it is advisable to keep a detailed log of the experiment and to note in this any observed differences. By so doing, a critical investigator will usually be able to avoid overlooking gross failures to control important educative factors.

3. *Measurement of achievement.* The general difficulty of securing reliable and valid measures of achievements in experimental investigations was discussed on pages 61-69, and little more needs to be said here. In order to understand the seriousness of this difficulty, it is necessary to bear in mind that the problem being studied specifies either explicitly or implicitly the achievement to be measured, and frequently the specifications include a number of relatively subtle elements of achievement. For example, if the project method is made the experimental factor, the claims made for this method by its advocates imply that the outcomes include such general patterns of conduct as initiative, resourcefulness, persistence, and interest in school work. In many experiments the quality of permanency of achievement is implied. For example, in reviewing the experiments on supervised study, it was pointed out that permanency of the study habits should be considered by ascertaining if they functioned in future study.

4. *Generalizing.* In order to generalize from the results of an experiment, it is necessary to have some index of the degree to which the pupils included in the experiment are representative of the larger population for which it is desired to state conclusions. If it can be demonstrated that the group of pupils included in the experiment is representative of the larger population, the investigator may state his conclusions as generalizations. If it is known that the group of pupils included in the experiment constitutes a random sample of the larger population, formulae are available for calculating the allowance that must be made for the operation of chance. Unfortunately the group of pupils available for experimental purposes can seldom be selected by a process of random sampling, and usually it is not possible to prove them highly representative. Hence, the experimenter faces the task of generalizing from data secured from a group of pupils whose degree of representativeness is not known in any precise way.

In generalizing, it is necessary to consider also the representativeness of the total instructional situation. For example, if a class-size experiment is organized so that each teacher has a large class and a small class, the experimenter is justified in generalizing only for such teaching situations.

Controlled experimentation versus informal experimentation. In attempting to evaluate educational experimentation as a research procedure, it is important to distinguish between the types commonly designated as controlled and informal experimentation. Controlled experimentation involves careful control of all non-experimental factors and is designed to lead to relatively precise and dependable results. In other words the conclusion is expected to be a definite statement of the relative merits of the educative procedures compared. Since this conclusion is based on objective data it is expected to be dependable and final, at least within the defined limits of the investigation.

Informal experimentation may be thought of as the trying out of an educative procedure to ascertain whether it works. Fundamentally it differs from controlled experimentation in the degree of refinement of the experimental procedure. For example, a teacher who tries out a new textbook controls other educative factors only to a very limited extent. The control group may be a class or classes taught during a previous term, and pupil achievement is measured in terms of teacher estimates or grades made on the final examination. The teacher may conclude that the textbook is unsatisfactory, but this conclusion cannot be regarded as a demonstrated fact.

Informal experimentation is frequently profitable. The teacher and others connected with the experimentation are usually stimulated and the findings may be in the direction of truth. But such investigations can not be expected to contribute to our fund of scientific knowledge relating to education.

The outlook. When we consider the crucial difficulties encountered in experimental investigations, it is difficult to be very optimistic in regard to the improvement of research procedures so that the findings will be highly dependable. As we have indicated, there is evidence that experimental techniques are being improved, and it is possible to present a strong case in support of the statement that we are leaving the plateau period. It is more difficult to predict the future, but it seems doubtful whether we are justified in expecting that in time it will be possible to set up an experiment or a group of experiments that will yield definite and final answers to any question concerning the relative merits of a given educative factor.

Some questions can be answered satisfactorily. A few have been answered. But for many questions, perhaps most questions, it is likely that we are not justified in expecting more than an "indication."

Controlled experimentation, however, is worthwhile. In addition to the dependable information that may be contributed, there are valuable by-products. Experimental investigations are stimulative. Experiments with the project method have stimulated a greater interest in this instructional procedure, and we know that under certain conditions it works.

UNIVERSITY OF ILLINOIS-URBANA



3 0112 065081975