NFER | Educational Research

Routledge
Taylor & Francis Group

Educational Research

Special issue: Randomised controlled trials (RCTs) in education research – methodological debates, questions, challenges
Guest Editors: Ben Styles and Carole Torgerson

ISSN: 0013-1881 (Print) 1469-5847 (Online) Journal homepage: http://www.tandfonline.com/loi/rere20

Randomised trials in education in the USA

Larry V. Hedges & Jacob Schauer

To cite this article: Larry V. Hedges & Jacob Schauer (2018) Randomised trials in education in the USA, Educational Research, 60:3, 265-275, DOI: 10.1080/00131881.2018.1493350

To link to this article: https://doi.org/10.1080/00131881.2018.1493350

Published online: 04 Jul 2018.

EDUCATIONAL RESEARCH
2018, VOL. 60, NO. 3, 265–275
https://doi.org/10.1080/00131881.2018.1493350

Randomised trials in education in the USA

Larry V. Hedges and Jacob Schauer

Institute for Policy Research, Northwestern University, Evanston, IL, USA

ABSTRACT

Background and purpose: Studies of education and learning that were described as experiments have been carried out in the USA by educational psychologists since about 1900. In this paper, we discuss the history of randomised trials in education in the USA in terms of five historical periods. In each period, the use of randomised trials was motivated by the research interests and conditions of the era. We have characterised these periods in terms of decades with sharp boundaries as a convenience.

Sources of evidence and main arguments: Although some of the early studies used random allocation (and even random allocation of clusters such as schools), early researchers did not clearly understand the role of randomisation or clearly distinguish it from methods such as alternation. In 1940, E. F. Lindquist published an important book whose goal was to translate R. A. Fisher's ideas into language congenial to education researchers, but this had little impact on education research outside of psychology. There was a substantial increase in the number of randomised trials during the period from 1960 to 1980, as the US government enacted and evaluated a variety of social programmes. This was followed by a dramatic decrease during the period from 1980 to 2000, amid debates about the relevance of randomised trials in education research. The creation of the US Institute of Education Sciences in 2002 provided major financial and administrative support for randomised trials, which has led to a large number of trials being conducted since that time.

Conclusions: These developments suggest that there is a promising future for randomised trials in the USA. American education scientists must remain committed to explaining why evidence from randomised field trials has an indispensable role to play in making wise decisions about education policy and advancing our capacity to improve education for a productive workforce and a successful society.

ARTICLE HISTORY
Received 19 July 2017
Accepted 17 June 2018

KEYWORDS
Randomised trial; education; history; statistics; programme evaluation; USA

Introduction

Randomised trials have been a part of American education research since the beginning of the early part of the twentieth century. Both individually randomised trials (including laboratory experiments) and cluster randomised field trials have been used throughout this period. Surprisingly, what appears to be the first randomised trial in American education was a cluster randomised trial that assigned schools to different treatments

CONTACT Larry V. Hedges l-hedges@northwestern.edu

to test the effect on student learning of time allocation for practice (see Cummins, 1919). Cummins appears to have analysed the trial incorrectly, ignoring the clustering by school in the statistical analysis, but the allocation was explicitly random (p. 51).

In this paper, we discuss the history of randomised trials in education in the United States in terms of five historical periods. In each period, the use of randomised trials was motivated by the research interests and conditions of the era. We have characterised these periods in terms of decades with sharp boundaries as a convenience, however the periods overlap and whatever precision there may be in this division should be conceived on the scale of decades, not specific years.

The early years: 1900–1940

The earliest experimental research in American education arose in educational psychology. American educational psychology was strongly influenced by the German psychophysical tradition of Wilhelm Wundt, C. E. Muller, and Hermann Ebbinghaus. The idea of experimental control had been prominent in psychophysics for at least half a century by 1900 and had a significant influence on the designs of early education researchers (see Boring, 1954). E. L. Thorndike of Teachers College, Columbia University, was a pioneer in the use of controlled laboratory studies of learning in America (see Thorndike and Woodworth, 1901 for an early example of a study design that compared treated and control groups). What constituted valid controls, however, was understood quite differently during this period than it was after 1940.

By 1923, a textbook on the design of education 'experiments' had appeared (McCall, 1923). Although this textbook was chronologically contemporary with R. A. Fisher's early writings on experimental design and statistical analysis methods, it is not clear that American educational psychologists were aware of Fisher's work at this time. In fact, a close reading of early studies in psychology suggests that they used the word 'experiment' in a more general way than we do today. While they appreciated the need for control of variation and comparing like with like, they did not yet appreciate the special role of random assignment, which both helps ensure comparable experimental groups and forms the basis for most statistical analyses of experiments. For example, McCall's book mentions equating groups 'by chance' (i.e., random assignment) but described random assignment and alternation both as examples of how to do this, apparently oblivious to the difference between the two. With the exception of a brief discussion of lottery systems, McCall does not really describe how to carry out random assignment. Moreover, equating experimental arms 'by chance' is only one of the methods mentioned for assuring equality of experimental groups. McCall devotes several pages to methods for deliberately matching participants in different groups on one or more baseline measures or on a preliminary measure of growth. These typically involve the researcher handpicking similar individuals for the treatment and control groups, ironically often using a system of alternation.

McCall and other education researchers of this period do not distinguish a special role for random assignment and largely considered matching a superior method for equating groups. The educational studies of this era reflect this preference. Many do not discuss the exact mechanism of treatment assignment but do describe the similarity between groups (e.g., Frost, 1921; Hurlock, 1925). In his examination of oversight

programmes for rural schools, Pittman (1921) 'arbitrarily' assigned geographic clusters of schools to treatment (p. 12). Throughout his account, he attempts to demonstrate the groups' equivalence on the variables of teacher experience and salary, community and individual characteristics, and school conditions (see pp. 12–18). Moreover, statistical analysis methods such as significance testing tend to be poorly described.

There is little evidence that Fisher's ideas including randomisation and experimental design had much influence on either American education or psychology before 1940. A survey of uses of the analysis of variance in psychology before 1940 revealed only 17 published papers. Nine of these were on educational topics, but four of them involved the use of analysis of variance to study the psychometric properties of tests, two clearly did not use random allocation, and the remaining two might have used random allocation, but did not explain their allocation strategy.

The introduction of Fisherian experimental design: 1940–1960

The publication of the first edition of Fisher's *The Design of Experiments* in 1935 did much to expose his ideas to many fields, including psychology (e.g., Rucci and Tweney, 1980). Experimental design and analysis in the tradition of Fisher were introduced into American educational research in 1940 by a remarkable book by E. F. Lindquist. Lindquist was an educational psychologist at the University of Iowa, which was quite near Iowa State University, one of the earliest centres of agricultural statistics in the USA. He is perhaps best known for his work in psychometrics and large-scale assessment, having founded the American College Testing Program (one of the largest commercial testing programmes in the USA) and the Iowa Testing Program. In the preface to his 1940 book, *Statistical Analysis in Educational Research*, Lindquist announced that his primary purpose was to 'translate Fisher's expositions into a language and notation familiar to the student of education'. He describes how to carry out random sampling and random allocation. He illustrates a variety of statistical methods and experimental designs that are useful in educational research. Of particular interest is his exposition of cluster randomised trials (his Design III), which he describes as follows:

> Design III: Experiment conducted in 10 schools. Five schools, selected at random, use Method A, the other five use method B. (p. 82)

He understands that clustering of students within schools renders an analysis assuming individual randomisation invalid, leading to underestimation of significance levels (p-values that are too small), and advises an analysis using school means (cluster means) as the unit of analysis. This is remarkable, given that this book predates by almost four decades the famous paper by Cornfield (1978) on cluster randomisation that is credited as bringing these issues to the attention of the researchers in the health sciences (see Donner and Klar, 2000).

Educational researchers did not immediately embrace Lindquist's ideas. For example, Quin McNemar, a Stanford psychologist with interests in education, criticised the exposition of the material and was sceptical about Lindquist's (correct) analysis of the cluster randomised design, saying that

> The reviewer suspects that something is wrong with a test of significance which does not involve the variation of the individuals upon which the means are based. (McNemar, 1940, p. 747)

The second edition of Linquist's book in 1953 improved the exposition but offered few real examples of the use of random assignment studies in education. His books, and those that followed during this period, did change the training psychologists and educational psychologists received in statistics and experimental design, helping to set the stage for the next period in the history of randomised trials in the USA.

The first flowering of randomised trials: 1960–1980

Between 1960 and about 1980, two distinct influences in American education led to a dramatic increase in the number of randomised trials in education. One was the systematic efforts of the federal government to alleviate poverty by improving education through the so-called Great Society programmes. Coupled with this expanding federal effort in education and social welfare was an emphasis on systematic evaluation of these programmes. The two decades from the mid-1960s to about 1980 has been called the golden era of evaluation research in which substantial numbers of randomised field trials were conducted to inform education policy (see Haveman, 1987). A bibliography by Boruch, McSweeny, and Soderstrom (1978) lists over 300 field experiments in 10 different areas, including over 90 trials in education. The aspects of education studied included teacher training programmes, curricular innovations, the use of instructional objectives, educational uses of technology, compensatory education and job training, school-based programmes for social adjustment, the interaction of programme and student characteristics (aptitude–treatment interaction studies), career education and the effects of testing and assessment.

Many of these trials were large and costly evaluations carried out by private research firms under contract to the US government. In fact, many of the most prominent US social research firms had their origins in this period. Some trials were conducted in universities and many were remarkably innovative, despite relatively small budgets. For example, the evaluation of the Harvard Project Physics curriculum used a cluster randomised design and recruited a nationally representative sample of schools, yet managed to carry out the evaluation for a budget of only about $10,000 (Welch and Walberg, 1972).

Trials conducted during this era continue to have a significant policy impact today. The Perry Preschool Project began in 1962 and assigned 123 low-income children to a high-quality preschool programme or a control group. The study has followed participants through to age 40, finding that participants had higher earnings, committed fewer crimes and were more likely to be employed than those control group (see Schweinhart, Barnes, and Weikart, 1993 or Schweinhart et al., 2004). The Abecedarian Project randomly assigned 111 infants born in 1972 and 1977 to a high-quality infant care or a control group. The study followed them through to age 35, finding that participants were more likely to hold a bachelor's degree and be employed (Campbell and Ramey, 1994). Moreover, intervention group participants were more likely to be in good health than the control group participants (Campbell et al., 2002). These two trials have provided the most persuasive evidence in current US policy debates that early childhood education programmes are cost effective.

Several education researchers in this period were motivated by scientific questions about instruction rather than evaluation of federal policies, and they conducted many trials during this period. The problem of improving instruction is fundamental to

education, but two distinct new approaches to this problem animated education researchers in this period. One was based on the development of new systematic methods to measure teacher behaviour. One compendium of such instruments eventually ran to six volumes (Simon and Boyer, 1967). Armed with new behavioural observation methods, researchers began experimental studies of the relation between systematically observed and meticulously coded teacher behaviour and variables like classroom average student achievement. A compendium and review of such studies during part of this period are given by Dunkin and Biddle (1974).

It seemed natural to such researchers that associations discovered in correlational studies like these should be confirmed in randomised trials (e.g., Gage, 1978a on the correlational-experimental loop). Researchers conducted many trials of varying quality (Dunkin and Biddle cite 28 experiments by 1974). Some of these experiments were quite sophisticated cluster randomised trials. Some used quite simple interventions based on results of correlational studies (e.g., Coladarci and Gage 1984). Others introduced interventions encouraging parent involvement (Crawford, et al., 1978). The most sophisticated of all was a cluster randomised $2 \times 2 \times 2$ factorial design (which allow researchers to investigate the effects of multiple factors in a single experiment) that involved manipulating three teaching variables simultaneously: highly versus minimally structured lessons, frequently versus infrequently asking questions and thorough versus neutral responses to student answers, for a total of eight treatment conditions (Gage, 1976).

Although the federal government invested substantial resources in field trials in research on teaching, this research tradition that began with great enthusiasm did not lead to the discovery of large or replicable effects of instructional practices on student outcomes. For example, the sophisticated (and very expensive) cluster randomised factorial design mentioned above found precisely no effects: no significant main effects and no significant interactions. Most of the other large-scale trials also failed to find statistically significant effects of the treatments they investigated. This was not always because of insufficient sample sizes; many of these trials had reasonable statistical power to detect modest effects. The consensus that emerged was that the research tradition had been a failure.

Another approach to improving instruction was the idea that any particular instructional method (treatment) might not be equally effective for all students but that its effectiveness might depend on the characteristics (aptitudes) of the student. This idea, proposed in elegant detail in Lee Cronbach's (1957) presidential address to the American Psychological Association, motivated a generation of research on aptitude–treatment interactions in American educational psychology. These trials were largely (but not exclusively) carried out in laboratories using individually randomised designs. For much of this period, they represented the pinnacle of theoretical and methodological sophistication in American education psychology. In 1977, Cronbach and Snow published a handbook based on two decades of research experience that provided excellent methodological advice on how to design aptitude–treatment interaction studies. The book also offered a thorough review of much of the extensive body of research that emerged. However, it concluded that few aptitude–treatment interactions could be replicated; interactions detected in one experiment were not found on others. This made it difficult to generalise scientific theories about

individual responses to treatment in ways that could genuinely and reliably improve instruction in US schools. Oddly enough, it was Cronbach himself who helped bring an end to research on aptitude–treatment interactions, in part because of these inconsistencies. In his 1975 paper entitled 'Beyond the two disciplines of scientific psychology', he concluded that

> higher order interactions make it unlikely that social scientists will be able to establish generalizations applicable beyond the laboratory or that generalizations established in the field work will be maintained. Social research should be less concerned with hypothesis testing and more concerned with interpreting findings in local contexts. (abstract)

A low point for education trials in the USA: 1980–2000

The end of the Great Society programmes and other large-scale federal anti-poverty programmes sharply reduced, but did not completely eliminate large-scale trials used in evaluation research. A few large-scale trials in education would be conducted in the next two decades, but these were almost exclusively conducted at the behest of government agencies and conducted by research firms outside of academia. The failure to find large main effects on instructional methods in large-scale trials in research on teaching methods and the failure to find replicable aptitude–treatment interactions were a huge blow to the prominence of trials within academic education research. Cronbach's pessimistic conclusion that it was unlikely that social generalisations could be established and that social scientists (and by implication, educational scientists) should focus on interpreting findings in local contexts was particularly influential in American education research. Cronbach was perhaps the leading education scientist of his time and he was a pre-eminent figure at the Stanford School of Education, the leading education school of the era. American education research turned increasingly towards ethnography and other qualitative research methods for inspiration after about 1980, beginning a period of 'paradigm wars', which was a dark era for those interested in randomised trials in education.

One of the important trials that had a huge influence on US education policy was the Tennessee class size experiment, also known as Project STAR (for Student–Teacher Achievement Ratio), which assessed the effect of smaller class sizes on student achievement. This randomised trial was commissioned in 1985 by the Tennessee state legislature and implemented by a consortium of Tennessee universities and the Tennessee State Department of Education. The total cost of the experiment, including the cost of hiring new teachers and classroom assistants, was approximately 12 million dollars.

Initially, all Tennessee school districts were asked to participate in Project STAR, and about 180 schools in about 50 of the 141 school systems in the state expressed interest in participating. Only about 100 schools had sufficient students in each grade to meet the size criteria (at least 57 students per grade necessary to form one small- and two regular-sized classes) for participation. This size criterion, which was necessary to permit assignment to class types within schools, excluded very small schools from the study. Ultimately, 79 elementary schools in 42 school districts became sites in the STAR experiment. Districts had to agree to participate for four years and allow site visitations for verification of class sizes, interviewing and data collection, including extra student testing. They also had to allow random assignment of pupils and teachers to class types from kindergarten through grade 3.

The state paid for the additional teachers and classroom assistants, and only class size conditions changed within schools. School districts and buildings followed their own policies, curricula, etc. The experiment randomly assigned kindergarten students into small classes (with 13–17 students), larger classes (with 22–26 students) or larger classes with a full-time classroom assistant. Teachers were also randomly assigned to classes of different types. These assignments of students and teachers to class type were maintained through the third grade. Some students entered the study in the first grade and subsequent grades, but were randomly assigned to classes at that time.

The experiment definitively established that students who were assigned to small classes had higher academic achievement during the 4 years of the experiment and that the positive effects of being in a small class were larger for students who experienced a longer duration of exposure to small classes (Nye, Hedges, and Konstantopoulos, 2000). The achievement gains persisted throughout elementary school through grade 8 (Nye, Hedges, and Konstantopoulos, 1999) and into high school (Nye, Hedges, and Konstantopoulos, 2001).

Another influential policy experiment was the national evaluation of the Upward Bound programme. Upward Bound is a key federal programme of assistance to poor children who aspire to obtain a college education. The US Congress mandated a national evaluation of Upward Bound in 1991, which was carried out by Mathematica Policy Research, a well-regarded private research firm with considerable experience with randomised trials of social interventions. The legislation that mandated the evaluation specified that it would be the *only* evaluation of the programme to ever be conducted (a sign that the enthusiasm for large-scale evaluation of social and educational programmes had definitely waned since the 1960s).

The evaluation itself was remarkable in that the trial used a national probability sample of 67 Upward Bound sites that were sufficiently oversubscribed so that participation in the trial would not result in a net reduction in service. Thus, the trial was one of the few large-scale education experiments with a truly representative national sample (of oversubscribed sites). Random assignment was conducted within sites resulting in 1500 individuals assigned to receive Upward Bound services and 1,300 assigned to a control group (Myers and Schirm, 1999).

The evaluation found statistically significant increases in high school mathematics credits and the likelihood of earning a postsecondary certificate or licence from a vocational school, but no detectable effects on other high school outcomes, including graduation and grades, or other postsecondary outcomes, including enrolment, financial aid application or receipt, or the completion of bachelor's or associate's degrees.

Laboratory studies in educational psychology continued in American colleges and universities, but field trials did not. Because relatively few trials were carried out by personnel in colleges and universities, training in trial methodology languished. Even basic training in statistics was uneven in many US schools of education during this period. By the year 2000, the lack of researchers with training and experience in carrying out randomised field trials would have important consequences for the next period in US education research.

The Institute of Education Sciences era: 2002 to the present

Two pieces of legislation in the early 2000s ushered in a sea change in education research in the USA. Congress passed the No Child Left Behind (NCLB) Act in 2001 and the Education Sciences Reform Act in 2002. The latter established the Institute of Education Sciences (IES) as a scientific agency intended to be shielded from political interference with independent authority to fund research and publish reports. NCLB generated instant demand for quality education research by requiring robust scientific evidence to justify certain expenditures of federal funds on education interventions, products and services. To help school personnel identify which of these were backed by rigorous research, IES created the What Works Clearinghouse (WWC). Part of this effort required a clear definition of 'rigorous research', so the WWC established standards of evidence for education studies. The WWC standards rated randomised field trials as the most rigorous research design, enabling randomised trials to meet its highest standards 'without reservations'. Thus, the new legislation created both a demand for rigorous research and the first of what became several dissemination mechanisms for it.

IES also created funding streams dedicated to *producing* high-quality education research through the National Center for Education Research and the National Center for Special Education Research. This was not just a matter of announcing grant competitions, hoping that researchers would apply, and funding the best of the lot. Essential to this effort was reforming the process of reviewing research grant proposals, so that they became more similar to those at the National Institutes of Health and other long-standing scientific agencies. The independent review office of IES established standards for IES products and processes for assuring compliance with those standards. The formation and maintenance of a strong standards and review office is one of the unheralded ingredients of IES's success.

After two decades in which education research training had included little emphasis on randomised field trials, IES was faced with an immediate problem: there were not enough researchers trained and experienced in carrying out trials to meet this rising demand. To increase the capacity of the field to carry out such studies, IES funded pre- and postdoctoral training programmes in universities. The pre-doctoral programmes had the dual purpose of ensuring that training on randomised trials was reincorporated into university curricula, while also drawing PhD students from disciplines outside of education into the field. The postdoctoral training programmes made it possible for those recently awarded doctorates to obtain advanced training in research methods and provided a mechanism for those trained in other disciplines to move to education research with the support of established researchers as mentors.

IES also increased the capacity of the scientific workforce to carry out trials by funding research training for established professionals. They have funded a summer training institute on randomised trials for established researchers since 2005. They have also funded summer institutes on quasi-experimentation and single-case research.

Finally, although this was not an IES initiative, IES agreed to support a new professional society called the Society for Research on Educational Effectiveness (SREE) to provide a professional home for education researchers interested in experiments and causal inference (which was consistent with IES goals).

American education research has changed dramatically since 2002. Then, virtually no school of education in the USA taught regular courses on experimental design for field trials, while today it is hard to find a serious research-oriented school of education without them. Then, few established education researchers were trained to conduct large-scale randomised experiments. Today, literally hundreds have received training through IES-sponsored summer institutes, and an evaluation (itself a randomised trial) of one such institute has shown that participants are more likely to subsequently conduct randomised trials than comparable individuals. Then, few researchers outside of commercial research firms had experience of doing randomised trials; today, hundreds do. The WWC has lived up to its name, anchoring a scientific knowledge base that has identified dozens of effective education products, interventions and services. After more than a decade, SREE is a professional society that supports the work of hundreds of members who engage in rigorous education research. The SREE journal, the *Journal of Research on Educational Effectiveness*, is prospering. Since its creation, IES has supported over 350 randomised field trials and is continuing to do so.

Conclusions

These developments suggest that there is a promising future for randomised trials in the USA. A rise in the number of studies with rigorous research designs since 2002, especially (but not limited to) randomised trials, has improved the validity of scientific findings in US education. Moreover, a growing focus on interdisciplinary research has provided a richer context to both the implications and the role of education research. This has coincided with expanded efforts to connect scientific findings to sound policy. The results of large-scale studies, such as Project STAR, Upward Bound or research into the benefits of pre-kindergarten education, have informed shifts in how US students are educated. Meanwhile, the WWC has served as an important source of information for policymakers.

However, US researchers are keenly aware that there are threats to the progress that we have made. Science and even evidence itself are under attack in some quarters. American education scientists must remain committed to explaining why evidence from randomised field trials has an indispensable role to play in making wise decisions about education policy and advancing our capacity to improve education for a productive workforce and a successful society.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Institute of Education Sciences [R305B140042].

ORCID

Jacob Schauer http://orcid.org/0000-0002-9041-7082

References

Boring, E. G. 1954. "The Nature and History of Experimental Control." *American Journal of Psychology* 67: 573–589.

Boruch, R. F., A. J. McSweeny, and E. Soderstrom. 1978. "Randomized Field Experiments for Program Planning, Development, and Evaluation: an Illustrative Bibliography." *Evaluation Quarterly* 2 (4): 655-95.

Campbell, F. A., and C. T. Ramey. 1994. "Effects of Early Intervention on Intellectual and Academic Achievement: A Follow-Up Study of Children from Low-Income Families." *Child Development* 65: 684–698. doi:10.2307/1131410.

Campbell, F. A., C. T. Ramey, E. Pungello, J. Sparling, and S. Miller-Johnson. 2002. "Early Childhood Education: Young Adult Outcomes from the Abecedarian Project." *Applied Developmental Science* 6: 42–57. doi:10.1207/S1532480XADS0601_05.

Coladarci, T., and N. L. Gage. 1984. "Effects if a Minimal Intervention on Teacher Behavior and Student Achievement." *American Educational Research Journal* 21: 539–556. doi:10.3102/00028312021003539.

Cornfield, J. 1978. "Randomization by Group: A Formal Analysis." *American Journal of Epidemiology* 108: 100–102.

Crawford, J. Gage, N., Corno, L., Stayrook, N., and Mitman, A. 1978. *An Experiment on Teacher Effectiveness and Parent-Assisted Instruction in the Third Grade* (3 vols). Stanford, CA: Center for Educational Research, Staford University (ERIC Document Reproduction Service No. ED 160648).

Cronbach, L. J. 1957. "The Two Disciplines of Scientific Psychology." *American Psychologist* 12: 671–684. doi:10.1037/h0043943.

Cronbach, L. J. 1975. "Beyond the Two Disciplines of Scientific Psychology." *American Psychologist* 30: 116–127. doi:10.1037/h0076829.

Cronbach, L. J., and R. E. Snow. 1977. *Aptitudes and Instructional Methods: A Handbook for Research on Interactions*. New York: Irvington Publishers.

Cummins, R. A. 1919. *Improvement in the Distribution of Practice*. New York: Teachers College, Columbia University.

Donner, A., and N. Klar. 2000. *Design and Analysis of Cluster Randomized Trials in Health Research*. London: Arnold.

Dunkin, M. J., and B. J. Biddle. 1974. *The Study Of Teaching*. New York: Holt, Rinehart, and Winston, Inc.

Frost, N. 1921. *A Comparative Study of Achievement in Town Schools*. New York: Teachers College, Columbia University.

Gage, N. L. 1976. "A Factorially Designed Experiment on Teacher Structuring, Soliciting, and Responding." *Journal of Teacher Education* 27: 35–38. doi:10.1177/002248717602700109.

Gage, N. L. 1978a. *The Scientific Basis of the Art of Teaching*. New York: Teachers College Press.

Haveman, R. H. 1987. "Policy Analysis and Evaluation Research after Twenty Years." *Policy Studies Journal* 16: 191-218. doi: 10.1111/j.1541-0072.1987.tb00775.x.

Hurlock, E. B. 1925. "An Evaluation of Certain Incentives Used in School Work." *Journal of Educational Psychology* 16 (3): 145–159. doi:10.1037/h0067716.

Lindquist, E. F. 1940. *Statistical Analysis in Educational Research*. Boston: Houghton-Mifflin.

Lindquist, E. F. 1953. *Design and Analysis of Experiments in Psychology and Education*. Boston: Houghton-Mifflin.

McCall, W. A. 1923. *How to Experiment in Education*. New York: MacMillan.

McNemar, Q. 1940. *Book Review of Statistical Analysis in Educational Research, Psychological Bulletin*37: 746–748.

Myers, D., and A. Schirm. 1999. *The Impacts of Upward Bound: Final Report for Phase I of the National Evaluation*. Washington, DC: Mathematica Policy Research.

Nye, B., L. V. Hedges, and S. Konstantopoulos. 1999. "The Long Term Effects of Small Classes: A Five Year Follow-Up of the Tennessee Class Size Experiment." *Educational Evaluation and Policy Analysis* 21: 127–142. doi:10.3102/01623737021002127.

Nye, B., L. V. Hedges, and S. Konstantopoulos. 2000. "The Effects of Small Classes on Achievement: The Results of the Tennessee Class Size Experiment." *American Educational Research Journal* 37: 123–151. doi:10.3102/00028312037001123.

Nye, B., L. V. Hedges, and S. Konstantopoulos. 2001. "The Long Term Effects of Small Classes in Early Grades: Lasting Benefits in Mathematics Achievement in Grade Nine." *The Journal of Experimental Education* 69: 245–257. doi:10.1080/00220970109599487.

Pittman, M. S. 1921. *The Value of School Supervision Demonstrated with the Zone Plan in Rural Schools*. Baltimore: Warwick & York.

Rucci, A. J., and R. D. Tweney. 1980. "Analysis of Variance and the 'Second Disciple' of Scientific Psychology: A Historical Account." *Psychological Bulletin* 87: 166–184. doi:10.1037/0033-2909.87.1.166.

Schweinhart, L. J., H. V. Barnes, and D. P. Weikart. 1993. *Significant Benefits: The HighScope Perry Preschool Study through Age 27*. Ypsilanti, MI: HighScope Press.

Schweinhart, L. J., J. Montie, Z. Xiang, W. S. Barnett, C. R. Belfield, and M. Nores. 2004. *Lifetime Effects: The HighScope Perry Preschool Study through Age 40*. Ypsilanti, MI: HighScope Press.

Simon, A., and G. Boyer. 1967. *Mirrors for Behavior: An Anthology of Classroom Observation Instruments*. Philadelphia: Research for Better Schools.

Thorndike, E., and R. Woodworth. 1901. "The Influence of Improvement in One Mental Function upon the Efficiency of Other Functions (I)." *Psychological Review* 8 (3): 247–261. doi:10.1037/h0074898.

Welch, W. W., and H. J. Walberg. 1972. "A National Experiment in Curriculum Evaluation." *American Educational Research Journal* 9: 373–384. doi:10.3102/00028312009003373.