History of the Control Group

TRUDY DEHUE

Volume 2, pp. 829–836

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

History of the Control Group

A Special Kind of Group

Currently, anyone with a qualification in social science (and also medicine) takes the notion of the 'control group' for granted. Yet, when one comes to think of it, a control group is an exceptional kind of group. Usually, the notion of a 'group' refers to people sharing a particular identity or aim, having a sense of oneness, and most likely also a group leader. Members of control groups, however, do not even need to know one another. One may run into groups of tourists, teenagers, hikers, or hooligans, but never a group of controls. Different from groups in the regular sense, control groups are merely a number of people. They do not exist outside the realm of human science experimentation, and *qua* group, they exist only in the minds of the experimenters who compose and study them. Members of control groups are not supposed to develop group cohesion because that would entail a contaminating factor in the experiment. Researchers are interested in the average response of individual members rather than in their behavior as a group (*see* **Clinical Trials and Intervention Studies**).

Control groups serve to check the mean effectiveness of an intervention. In order to explain their need, a methods teacher will first suggest that if a particular person shows improvement this does not yet guarantee effectiveness in other people too. Therefore, a number of people must be studied and their average response calculated. This, however (the teacher will proceed), would not be enough. If we know the average result for one group of people, it remains unclear whether the intervention caused the outcome or a simultaneous other factor did. Only if the mean result of the group differs significantly from that of an untreated control group is the conclusion justified that the treatment caused the effect (or, more precisely, that the effect produced was not accidental).

Apart from the idea of comparing group averages, the notion of the true control group entails one pivotal extra criterion. The experimental and control groups must be comparable, that is, the former should not differ from the latter except through the action of the independent variable that is to be tested. One option is to use the technique of 'matching', that is, to pretest people on factors suspected of causing bias and then create groups with similar test results. Most contemporary methodologists, however, agree that random assignment of the participants to the groups offers a better guarantee of comparability. Assigning people on the basis of chance ensures that both known and unknown invalidating factors are cancelled out, and that this occurs automatically rather than being dependent on individual judgment and trustworthiness. In behavioral and social (as well as clinical) research, the ideal scientific experiment is a so-called *randomized controlled trial*, briefly an RCT.

In view of its self-evident value, and in view of the extensive nineteenth-century interest in human science experimentation, the introduction of the control group may strike us as being remarkably late. Until the early 1900s, the word control was not used in the context of comparative experiments with people, whereas ensuring comparability by matching dates back to the 1910s, and composing experimental and control groups at random was first suggested as late as the 1920s.

The present history connects the seemingly late emergence of the control group to its special nature as a group without a shared extra-individual identity. Moreover, it explains that such groups were inconceivable before considerable changes occurred in society at large. First, however, we need to briefly explain why famous examples of comparison such as that of doctor Ignaz Semmelweis, who fought childbed fever by comparing two maternity clinics in mid-nineteenth-century Vienna [23], are not included in the present account.

The Past and the Present

Comparison is 'a natural thing to do' to anyone curious about the effects of a particular action. Therefore, it should not come as a surprise that instances of comparison can also be found in the long history of interventions into human life. In a scholarly article on the history of experimentation with medical treatments, Ted Kaptchuck discussed various eighteenth-century procedures of comparison (although not to similar groups) such as the deliberately deceptive provision of bread and sugar pills

to check the claims of homeopathy [24]. And several examples of group comparison in the treatment of illnesses (although without randomization) are also presented in the electronic *James Lind Library* (`www.jameslindlibrary.org`).

Entertaining, however, as such examples of comparison may be, they are hardly surprising, since checking the effects of ones actions by sometimes withholding them is a matter of everyday logic. Moreover, it would be quite artificial to depict these examples as early, if still incomplete, steps toward the present-day methodological rule of employing control groups. Historians of science use derogatory labels such as 'presentist history', 'finalist history', 'justificationary history', and 'feel good history', for histories applying present-day criteria in selecting 'predecessors' who took 'early steps' toward our own viewpoints, whilst also excusing these 'pioneers' for the understandable shortcomings still present in their ideas. Arranging the examples in chronological order, as such histories do, suggests a progressive trajectory from the past to the present, whereas they actually drew their own line from the present back into the past. Historian and philosopher of science Thomas Kuhn discussed the genre under the name of 'preface history', referring to the typical historical introduction in textbooks. Apart from worshipping the present, Kuhn argued, preface histories convey a deeply misleading view of scientific development as a matter of slow, but accumulative, discovery by a range of mutually unrelated great men [25, pp. 1–10; 136–144].

Rather than lining up unconnected look-alikes through the ages, the present account asks when, why, and how employing control groups became a methodological condition. The many reputed nineteenth-century scholars who explicitly *rejected* experimental comparison are neither scorned nor excused for their 'deficiency'. Rather, their views are analyzed as contributions to debates in their own time. Likewise, the ideas of early twentieth-century scholars who advanced group comparison are discussed as part of debates with their own contemporaries.

Nineteenth-century Qualms

If control groups were not recommended before the early twentieth century, the expression of "social experimentation" did appear in much earlier methodological texts. Eighteenth-century scholars had already discussed the issue of experimentation as a suitable method for investigating human life [7]. David Hume's *Treatise of Human Nature*, first published in 1739, is subtitled: *Being an Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects*. Hume and his Enlightment contemporaries, however, borrowed the terminology of experimentation from natural science as a metaphor for major events happening without the intervention of researchers. Observing disturbances of regular life, they argued, is the human science substitute of natural science experimentation.

Nineteenth-century views on social experimentation were largely, but not entirely, the same as those of the eighteenth century. Distinguished scholars such as **Adolphe Quetelet** (1796–1874) in Belgium, Auguste Comte (1798–1857) in France, and George Cornwall Lewis (1806–1863) as well as John Stuart Mill (1806–1873) in Britain used the terminology of experimentation for incidents such as natural disasters, famines, economic crises, and also government interventions. Different, however, from eighteenth-century scholars and in accordance with later twentieth-century views, they preserved the epithet of *scientific* experimentation for experiments with active manipulation by researchers. As scientific experimentation entails intentional manipulation by researchers, they maintained, research with human beings cannot be scientific.

Roughly speaking, there were two reasons why they excluded deliberate manipulation from the useable methods of research with human beings. One reason was of a moral nature. When George Cornwall Lewis in 1852 published his two-volume *Treatise on the Methods of Observation and Reasoning in Politics*, he deliberately omitted the method of experimentation from the title. Experimentation, Lewis maintained, is 'inapplicable to man as a sentient, and also as an intellectual and moral being'. This is not 'because man lies beyond the reach of our powers', but because experiments 'could not be applied to him without destroying his life, or wounding his sensibility, or at least subjecting him to annoyance and restraint' [26, pp. 160–161].

The second reason was of an epistemological nature. In 1843, the prominent British philosopher, economist, and methodologist John Stuart Mill published his S*ystem of Logic* that was to become very influential in the social sciences. This work extensively discussed Mill's 'method of difference', which

entailed comparing cases in which an effect does and does not occur. According to Mill, this 'most perfect of the methods of experimental inquiry' was not suitable for research with people. He illustrated this view with the 'frequent topic of debate in the present century', that is, whether or not government intervention into free enterprise impedes national wealth. The method of difference is unhelpful in a case like this, he explained, because comparability is not achievable: '[I]f the two nations differ in this portion of their institutions, it is from some differences in their position, and thence in their apparent interests, or in some portion or the other of their opinions, habits and tendencies; which opens a view of further differences without any assignable limit, capable of operating on their industrial prosperity, as well as on every other feature of their condition, in more ways than can be enumerated or imagined' [31, pp. 881–882].

Mill raised the objection of incomparability not only in complex issues such as national economic policies but in relation to all research with people. Even a comparatively simple question such as, whether or not mercury cures a particular disease, was 'quite chimerical' as it was impossible in medical research to isolate a single factor from all other factors that might constitute an effect. Although the efficacy of 'quinine, colchicum, lime juice, and cod liver oil' was shown in so many cases 'that their tendency to restore health... may be regarded as an experimental truth', real experimentation was out of the question, and '[S]till less is this method applicable to a class of phenomena more complicated than those of physiology, the phenomena of politics and history' [31, pp. 451–452].

Organicism and Determinism

How to explain the difference between these nineteenth-century objections and the commonness of experimentation with experimental and control groups in our own time? How could Lewis be compunctious about individual integrity even to the level of not 'annoying' people, whereas, in our time, large group experiments hardly raised an eyebrow? And why did a distinguished methodologist like Mill not promote the solution, so self-evident to present-day researchers, of simply *creating* comparable groups if natural ones did not exist?

The answer is that their qualms were inspired by the general holism and determinism of their time. Nineteenth-century scholars regarded communities as well as individuals as organic systems in which every element is closely related to all others, and in which every characteristic is part of an entire pattern of interwoven strands rather than caused by one or more meticulously isolated factors. In addition, they ascribed the facts of life to established laws of God or Nature rather than to human purposes and plans. According to nineteenth-century determinism, the possibilities of engineering human life were very limited. Rather than initiating permanent social change, the role of responsible authorities was to preserve public stability. Even Mill, for whom the disadvantages of a *laissez-faire* economy posed a significant problem, nevertheless, held that government interference should be limited to a small range of issues and should largely aim at the preservation of regular social order.

In this context, the common expression of 'social experimentation' could not be more than a metaphor to express the view that careful observation of severe disturbances offers an understanding of the right and balanced state of affairs. The same holistic and determinist philosophy expressed itself in nineteenth-century statistics, where indeterminism or chance had the negative connotation of lack of knowledge and whimsicality rather than the present-day association of something to 'take' and as an instrument to make good use of [33, 22]. Nineteenth-century survey researchers, for instance, did not draw representative population samples. This was not because of the inherent complexity of the idea, nor because of sluggishness on the researchers' part, but because they investigated groups of people as organic entities and prototypical communities [17]. To nineteenth-century researchers, the idea of using chance for deriving population values, or, for that matter, allocating people to groups, was literally unimaginable.

Even the occasional proponent of active experimentation in clinical research rejected chance as an instrument of scientific research. In 1865, the illustrious French physiologist Claude Bernard (1813–1878) published a book with the deliberately provocative title of *Introduction à L'étude de la Médecine Expérimentale* [1] translated into English as *An Introduction to the Study of Experimental Medicine.* Staunchly, Bernard stated that 'philosophic obstacles'

to experimental medicine 'arise from vicious methods, bad mental habits, and certain false ideas' [2, p. 196]. For the sake of valid knowledge, he maintained, 'comparative experiments have to be made at the same time and on as comparable patients as possible' [2, p. 194].

Yet, one searches Bernard's *Introduction* in vain for comparison of experimental to control groups. As ardently as he defended experimentation, he rejected statistical averages. He sneered about the 'startling instance' of a physiologist who collected urine 'from a railroad station urinal where people of all nations passed' as if it were possible to analyze 'the *average* European urine!' (italics and exclamation mark in original). And he scorned surgeons who published the success rates of their operations because average success does not give any certainty on the next operation to come. Bernard's expression of 'comparative experimentation' did refer to manipulating animals and humans for the sake of research. Instead of comparing group averages, however, he recommended that one should present 'our most perfect experiment as a type' [2, pp. 134–135]. To Bernard, the rise of probabilistic statistics meant 'literally nothing scientifically' [2, p. 137].

Impending Changes

The British statistician, biometrician, and eugenicist **Sir Francis Galton** (1822–1911) was a crucial figure in the gradual establishment of probabilism as an instrument of social and scientific progress. Galton was inspired by Adolphe Quetelet's notion of the statistical mean and the normal curve as a substitute for the ideal of absolute laws. In Quetelet's own writings, however, this novelty was not at odds with determinism. His well-known *L'homme moyen* (average man) represented normalcy and dispersion from the mean signified abnormality. It was Galton who gave Quetelet's mean a progressive twist.

Combining the evolution theory of his cousin Charles Darwin with eugenic ideals of human improvement, Galton held that 'an average man is morally and intellectually a very uninteresting being. The class to which he belongs is bulky, and no doubt serves to keep the course of social life in action... But the average man is of no direct help towards evolution, which appears to our dim vision to be the primary purpose, so to speak, of all living existence', whereas '[E]volution is an unresting progression,' Galton added, 'the nature of the average individual is essentially unprogressive' [20, p. 406].

Galton was interested in finding more ways of employing science for the sake of human progress. In an 1872 article 'Statistical Inquiries into the Efficacy of Prayer', he questioned the common belief that 'sick persons who pray, or are prayed for, recover on the average more rapidly than others.' This article opened with the statement that there were two methods of studying an issue like the profits of piety. The first one was 'to deal with isolated instances'. Anyone, however, using that method should suspect 'his own judgments' or otherwise would 'certainly run the risk of being suspected by others in choosing one-sided examples'. Galton vigorously broke a lance for substituting the study of representative types with statistical comparison. The most reliable method was 'to examine large classes of cases, and to be guided by broad averages' [19, p. 126].

Galton elaborately explained how the latter method could be applied in finding out the revenues of praying: 'We must gather cases for statistical comparison, in which the same object is keenly pursued by two classes similar in their physical but opposite in their spiritual state; the one class being spiritual, the other materialistic. Prudent pious people must be compared with prudent materialistic people and not with the imprudent nor the vicious. We simply look for the final result - whether those who pray attain their objects more frequently than those who do not pray, but who live in all other respects under similar conditions' [19, p. 126].

As it seems, Galton was the first to advocate comparison of group averages. Yet, his was not an example of treating one group while withholding the treatment from a comparison group. The emergence of the control group in the present-day sense occurred when his fears of 'being suspected by others in choosing one-sided examples' began to outgrow anxieties on doing injustice to organic wholes. This transition took place with the general changeover from determinism to progressivism in a philosophical as well as social sense.

Progressivism and Distrust

By the end of the nineteenth century, extreme destitution among the working classes led to social

movements for mitigation of *laissez faire* capitalism. Enlightened members of the upper middle class pleaded for some State protection of laborers via minimum wage bills, child labor bills, and unemployment insurances. Their appeals for the extension of government responsibility met with strong fears that help would deprive people of their own responsibility and that administrations would squander public funds. It was progressivism combined with distrust that constituted a new definition of social experimentation as statistical comparison of experimental and control groups. Three interrelated maxims of twentieth-century economic liberalism were crucial to the gradual emergence of the present-day ideal experiment.

The first maxim was that of *individual responsibility*. Social success and failure remained an individual affair. This implied that ameliorative attempts were to be directed first and foremost at problematic individuals rather than on further structural social change. Helping people implied trying to turn them into independent citizens by educating, training, punishing, and rewarding them. The second maxim was that of *efficiency*. Ameliorative actions financed with public money had to produce instant results with simple economical means. The fear that public funds would be squandered created a strong urge to attribute misery and backwardness to well-delineated causes rather than complex patterns of individual and social relations. And the third maxim was that of *impersonal procedures*. Fears of abuse of social services evoked distrust of people's own claims of needs, and the consequent search for impersonal techniques to establish the truth 'behind' their stories [38]. In addition, not only was the self-assessment of the interested recipients of help to be distrusted but also that of the politicians and administrators providing help. Measurement also had to control administrators' claims of efficiency [34].

Academic experts on psychological, sociological, political, and economical matters adapted their questions and approaches to the new demands. They began to produce technically useful data collected according to standardized methodological rules. Moreover, they established a partnership with statisticians who now began to focus on population varieties rather than communalities. In this context, the interpretation of chance as something one must make good use of replaced the traditional one of chance as something to defeat [17, 22, 33].

The new social scientists measured people's abilities, motives, and attitudes, as well as social phenomena such as crime, alcoholism, and illiteracy. Soon, they arrived at the idea that these instruments could also be used for establishing the results of ameliorative interventions. In 1917, the well-reputed sociologist F. Stuart Chapin lengthily discussed the issue. Simple, before and after measurement of one group, he stated, would not suffice for excluding personal judgement. Yet, Chapin rejected comparison of treated and untreated groups. Like Mill before him, he maintained that fundamental differences between groups would always invalidate the conclusions of social experiments. Adding a twentieth-century version to Lewis' moral objections, he argued that it would be immoral to withhold help from needy people just for the sake of research [9, 10]. It was psychologists who introduced the key idea to create equal groups rather than search for them in natural life, and they did so in a context with few ethical barriers.

Creating Groups

Psychologists had a tradition of psychophysiological experimentation with small groups of volunteers in laboratory settings for studying the law-like relationships between physical stimuli and mental sensations. During the administrative turn of both government and human science, many of them adapted their psychophysiological methods to the new demands of measuring progress rather than just discovering laws [14, 15]. One of these psychologists was John Edgar Coover, who studied at Stanford University in Palo Alto (California) with the psychophysical experimenter Frank Angell. As a former school principal, Coover gave Angell's academic interests an instrumental twist. He engaged in a debate among school administrators on the utility of teaching subjects such as Latin and formal mathematics. Opponents wanted to abolish such redundant subjects from the school curriculum, but proponents argued that 'formal discipline' strengthens general mental capacities. Coover took part in this debate with laboratory experiments testing whether or not the training of one skill improves performance in another ability. In a 1907 article, published together with Angell, he explained that in the context of this kind of research a one-group design does not do. Instead, he compared the achievements of experimental 'reagents'

who received training with those of control 'reagents' who did not [13]. Coover and Angell's article seems to be the first report of an experiment in which one group of people is treated and tested, while another one is only tested.

From the 1910s, a vigorous movement started in American schools for efficiency and scientific (social) engineering [6]. In the school setting, it was morally warrantable and practically doable to compare groups. Like the earlier volunteers in laboratories, school children and teachers were comparatively easy to handle. Whereas historian Edwin Boring found no control groups in the 1916 volume of the *American Journal of Psychology* [3, page 587], historian Kurt Danziger found 14 to 18% in the 1914–1916 volumes of the *Journal of Educational Psychology* [14, pp. 113–115].

Psychological researchers experimented in real classrooms where they tested the effects of classroom circumstances such as fresh versus ventilated air, the sex of the teacher, memorizing methods, and educational measures such as punishing and praising. They also sought ways of excluding the possibility that their effects are due to some other difference between the groups than the variable that is tested. During the 1920s, it became customary to handle the problem by matching. Matching, however, violated the guiding maxims of efficiency and impersonality. It was quite time- and money-consuming to test each child on every factor suspected of creating bias. And, even worse, determining these factors depended on the imaginative power and reliability of the researchers involved. Matching only covered possibly contaminating factors that the designers of an experiment were aware of, did not wish to neglect, and were able to pretest the participants on.

In 1923, William A. McCall at Columbia University in New York, published the methodological manual *How to Experiment in Education* in which he emphasized the need of comparing similar groups [30]. In the introduction to this volume, McCall predicted that enhancing the efficiency of education could save billions of dollars. Further on, he proposed to equate the groups on the basis of chance as 'an economical substitute' for matching. McCall did not take randomization lightly. For example, he rejected the method of writing numbers on pieces of paper because papers with larger numbers contain more ink and are therefore likely to sink further to the bottom of a container. But, he stated, 'any device which will make the selection truly random is satisfactory' [30, pp.41–42].

Fisher's Support

In the meantime, educational psychologists were testing various factors simultaneously, which made the resulting data hard to handle. The methodological handbook *The Design of Experiments* published in 1935 by the British biometrician and agricultural statistician **Ronald A. Fisher** provided the solution of **analysis of variance (ANOVA)**. As Fisher repeatedly stressed, random allocation to groups was a central condition to the validity of this technique. When working as a visiting professor at the agricultural station of Iowa State College, he met the American statistician **George W. Snedecor**. Snedecor published a book based on Fisher's statistical methods [37] that was easier to comprehend than Fisher's own, rather intricate, writings and that was widely received by methodologists in biology as well as psychology [28, 35]. Subsequently, Snedecor's Iowa colleague, the educational psychologist Everett Lindquist, followed with the book *Statistical Analysis in Educational Research* which became a much-cited source in the international educational community [27].

Fisher's help was welcomed with open arms by methodologists, not only because it provided a means to handle multi factor research but also because it regulated experimentation from the stage of the experimental design. As Snedecor expressed it in 1936, the designs researchers employed often 'baffled' the statisticians. 'No more than a decade past, the statistician was distinctly on the defence', he revealed, but '[U]nder the leadership of R. A. Fisher, the statistician has become the aggressor. He has found that the key to the problem is the intimate relation between the statistical method and the experimental plan' [36, p. 690]. This quote confirms the thesis of historians that the first and foremost motive to prescribe randomization was not the logic of probabilistic statistics, but the wish to regulate the conduct of practicing researchers [8, 16, 29, 34]. Canceling out personal judgment, together with economical reasons, was the predominant drive to substitute matching by randomization. Like Galton in 1872, who warned against eliciting accusations of having chosen one-sided examples, early twentieth-century

statisticians and methodologists cautioned against the danger of selection bias caused by high hopes on particular outcomes.

Epilogue

It took a while before **randomization** became more than a methodological ideal. Practicing physicians argued that the hopes of a particular outcome are often a substantial part of the treatment itself. They also maintained that it is immoral to let chance determine which patients gets the treatment his doctor believes in and which patient does not, as well as keeping it a secret as to which group a patient has been assigned. Moreover, they put forward the argument that subjecting patients to standardized tests rather than examining them in a truly individual way would harm, rather than enhance, the effectiveness of diagnoses and treatments.

In social research, there were protests too. After he learned about the solution of random allocation, sociologist F. Stuart Chapin unambiguously rejected it. Allocating people randomly to interventions, he maintained, clashes with the humanitarian mores of reform [11, 12]. And the Russian-American anthropologist Alexander Goldenweiser objected that human reality 'resents highhanded manipulation' for which reason it demands true dictators to 'reduce variety by fostering uniformity' [21, p. 631]. An extensive search for the actual use of random allocation in social experiments led to the earliest instance in a 1932 article on educational counseling of university students, whereas the next seven appeared in research reports dating from the 1940s (all but one in the field of educational psychology) [18].

Nevertheless, the more twentieth-century welfare capitalism replaced nineteenth-century *laissez-faire* capitalism, the more administrators and researchers felt that it is both necessary and morally acceptable to experiment with randomized groups of children as well as adults. From about the 1960s onward, therefore, protesting doctors could easily be accused of an unwillingness to give up an outdated elitist position for the truly scientific attitude. Particularly in the United States, the majority of behavioral and social researchers too began to regard experiments with randomly composed groups as the ideal experiment. Since President Johnson's War on Poverty, many such social experiments have been conducted, sometimes with thousands of people. Apart from school children and university students, also soldiers, slum dwellers, spouse beaters, drug abusers, disabled food-stamp recipients, bad parents, and wild teenagers have all participated in experiments testing the effects of special training, social housing programs, marriage courses, safe-sex campaigns, health programs, income maintenance, employment programs, and the like, in an impersonal, efficient, and standardized way [4, 5, 32].

References

[1] Bernard, C. (1865). *Introduction à L'étude de la Médecine Expérimentale*, Ballière, Paris.

[2] Bernard, C. (1957). *An Introduction to the Study of Experimental Medicine*, Dover Publications, New York.

[3] Boring, E.G. (1954). The nature and history of experimental control, *American Journal of Psychology* **67**, 573–589.

[4] Boruch, R. (1997). *Randomised Experiments for Planning and Evaluation*, Sage Publications, London.

[5] Bulmer, M. (1986). Evaluation research and social experimentation, in *Social Science and Social Policy*, M. Bulmer, K.G. Banting, M. Carley & C.H. Weiss, eds, Allen and Unwin, London, pp. 155–179.

[6] Callahan, R.E. (1962). *Education and the Cult of Efficiency*, The University of Chicago Press, Chicago.

[7] Carrithers, D. (1995). The enlightment science of society, in *Inventing Human Science. Eighteenth-Century Domains*, C. Fox, R. Porter & R. Wokler, eds, University of California Press, Berkeley, pp. 232–270.

[8] Chalmers, I. (2001). Comparing like with like. Some historical milestones in the evolution of methods to create unbiased comparison groups in therapeutic experiments, *International Journal of Epidemiology* **30**, 1156–1164.

[9] Chapin, F.S. (1917a). The experimental method and sociology. I. The theory and practice of the experimental method, *Scientific Monthly* **4**, 133–144.

[10] Chapin, F.S. (1917b). The experimental method and sociology. II. Social legislation is social experimentation, *Scientific Monthly* **4**, 238–247.

[11] Chapin, F.S. (1938). Design for social experiments, *American Sociological Review* **3**, 786–800.

[12] Chapin, F.S. (1947). *Experimental Designs in Social Research*, Harper & Row, New York.

[13] Coover, J.E. & Angell, F. (1907). General practice effect of special exercise, *American Journal of Psychology* **18**, 328–340.

[14] Danziger, K. (1990). *Constructing the Subject*, Cambridge University Press, Cambridge.

[15] Dehue, T. (2000). From deception trials to control reagents. The introduction of the control group about a century ago, *American Psychologist* **55**, 264–269.

[16] Dehue, T. (2004). Historiography taking issue. Analyzing an experiment with heroin maintenance, *Journal of the History of the Behavioral Sciences* **40**(3), 247–265.

[17] Desrosières, A. (1998). *The Politics of Large Numbers. A History of Statistical Reasoning*, Harvard University Press, Cambridge.

[18] Forsetlund, L., Bjørndal, A. & Chalmers, I. (2004, submitted for publication). Random allocation to assess the effects of social interventions does not appear to have been used until the 1930s.

[19] Galton, F. (1872). Statistical inquiries into the efficacy of prayer, *Fortnightly Review* **XII**, 124–135.

[20] Galton, F. (1889). Human variety, *Journal of the Anthropological Institute* **18**, 401–419.

[21] Goldenweiser, A. (1938). The concept of causality in physical and social science, *American Sociological Review* **3**(5), 624–636.

[22] Hacking, I. (1990). *The Taming of Chance*, Cambridge University Press, New York.

[23] Hempel, C.G. (1966). *Philosophy of Natural Science*, Prentice-Hall, Englewood Cliffs.

[24] Kaptchuck, T.J. (1998). Intentional ignorance: a history of blind assessment and placebo controls in medicine, *Bulletin of the History of Medicine* **72**, 389–433.

[25] Kuhn, T.S. (1962, reprinted 1970). *The Structure of Scientific Revolutions*, Chicago University Press, Chicago.

[26] Lewis, C.G. (1852, reprinted 1974). *A Treatise on the Methods of Observation and Reasoning in Politics*, Vol 1, Arno Press, New York.

[27] Lindquist, E.F. (1940). *Statistical Analysis in Educational Research*, Houghton-Mifflin, Boston.

[28] Lovie, A.D. (1979). The analysis of variance in experimental psychology: 1934–1945, *British Journal of Mathematical and Statistical Psychology* **32**, 151–178.

[29] Marks, H.M. (1997). *The Progress of Experiment. Science and Therapeutic Reform in the United States, 1900–1990*, Cambridge University Press, New York.

[30] McCall, W.A. (1923). *How to Experiment in Education*, McMillan McCall, New York.

[31] Mill, J.S. (1843, reprinted 1973). *A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation*, University of Toronto, Toronto.

[32] Orr, L. (1999). *Social Experiments. Evaluating Public Programs with Experimental Methods*, Sage Publications, London.

[33] Porter, T.M. (1986). *The Rise of Statistical Thinking, 1820–1900*, Princeton University Press, Princeton.

[34] Porter, T.M. (1995). *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*, Princeton University Press, Princeton.

[35] Rucci, A.J. and Ryan D.T. (1980). Analysis of variance and the 'second discipline' of scientific psychology: a historical account, *Psychological Bulletin* **87**, 166–184.

[36] Snedecor, G.W. (1936). The improvement of statistical techniques in biology, *Journal of the American Statistical Association* **31**, 690–701.

[37] Snedecor, G.W. (1937). *Statistical Methods*, Collegiate Press, Ames, Iowa.

[38] Stone, D.A. (1993). Clinical authority in the construction of citizenship, in *Public Policy for Democracy*, H. Ingram & S. Rathgeb Smith, eds, Brookings Institution, Washington, pp. 45–68.

TRUDY DEHUE