The Royal Society of Medicine

Journal of the Royal Society of Medicine; 2021, Vol. 114(4) 212–217
DOI: 10.1177/01410768211003771

Richard Cabot, pair-matched random allocation, and the attempt to compare like with like in the social sciences and medicine. Part 1: the context of the social sciences

Brandon C Welsh[1], Scott H Podolsky[2] and Steven N Zane[3]
[1]School of Criminology and Criminal Justice, Northeastern University, Boston, MA 02115, USA
[2]Department of Global Health and Social Medicine, Harvard Medical School, Boston, MA 02115, USA
[3]College of Criminology and Criminal Justice, Florida State University, Tallahassee, FL 32304, USA
Corresponding author: Brandon C Welsh. Email: b.welsh@northeastern.edu

Introduction

In 1935, Richard Cabot began designing the Cambridge-Somerville Youth Study to evaluate the impact of a social intervention of 'directed friendship' on youth delinquency. By the commencement of the study in 1937, hundreds of boys had been 'matched' into pairs based on 142 separate variables in an attempt to ensure that like were being compared with like in the intervention and control groups. The matching procedure was elaborate and adhered to the principle that 'each personality would be studied both statistically and configurationally',[1] a nod to educational research.[2] But Cabot introduced a final methodological manoeuvre – a coin flip to determine which boy within each pair would be assigned to the intervention group.

Today, the Cambridge-Somerville Youth Study is recognised as the first randomised trial in criminology,[3] one of the earliest randomised clinical trials of a social intervention,[4] and seemingly the first in the social and behavioural sciences to use alternate or random allocation after matching study participants into pairs.[5]

Historical research has previously investigated the personal, professional and institutional influences that inspired Cabot's vision for the study and its research design.[5–8]

Yet Cabot – both a physician and a social interventionist – lived at the interface of medicine and the social sciences. His study marks a telling moment in the history of attempts to compare like with like – of the evolving articulation among pre-allocation stratification, matching, alternate allocation, random allocation and other innovations intended to ensure fair comparisons of the effects of interventions. It is also a useful starting point from which to explore the migration of innovations across seemingly siloed disciplines.

This article uses Cabot's iconic study as a focal point from which to explore the tension between pre-allocation stratification and matching on the one hand, and alternate allocation and random allocation on the other, as a means of ensuring comparisons of like with like. It also examines the comparative histories of this tension in the social sciences (with a focus on criminology) and medicine as a means of further contextualising Cabot's study.

Richard Clarke Cabot and the Cambridge-Somerville Youth Study

Richard Clarke Cabot (1868–1939) was a physician and professor of clinical medicine and social ethics at Harvard University.[9–11] He made a number of important contributions to medicine and public health, including differential diagnosis using blood,[12] the establishment of the clinical pathological conference,[13] and, in 1905, starting the first medical social work programme in the country at Massachusetts General Hospital.[14,15] In the social sciences, he is best known for advancing the field of social ethics,[16] advocating for social work practice and research, serving as president of the National Conference of Social Work,[17] and developing and directing the Cambridge-Somerville Youth Study.

The Cambridge-Somerville Youth Study would be Cabot's final project, consuming the last five years of his life. In 1935, he incorporated a charity named after his late wife, the Ella Lyman Cabot Foundation, with the express purpose of funding an experimental intervention of young boys judged to be at increased risk of becoming delinquent.[18] In the same year, he created a selection committee, comprising three prominent practitioners in juvenile and criminal justice, to identify and recruit boys for the study.[1] The committee was charged with

recruiting boys who were between the ages of 5 and 13 years, attended public and parochial schools and who lived in the working-class areas of Cambridge and Somerville (Massachusetts), and were deemed to be 'pre-delinquent'. Characteristics of pre-delinquency included 'persistent truancy, persistent breaking of the rules, sex difficulties, petty pilfering and stealing, failing to return home after school, and, among the kindergarteners, temper tantrums'.[19]

A large number of boys were referred to the committee, mostly by local schools (approximately 77%), local welfare agencies, churches and the police. Information on the boys was collected from a wide range of sources, including elementary school teachers, juvenile courts, physicians and the parents of eligible boys. Case files of eligible boys were turned over to the 'matchers', a group of psychologists employed by the study. The process of matching involved two steps. A group of older boys (n = 80) were first observed on overnight camping trips to assess relevant social and personality traits to operationalise matching parameters. The psychologists then matched all boys (n = 650) using 142 variables (rated on an 11-point scale) covering a wide range of characteristics, including physical health, emotional and social adjustment, father's occupation, teacher ratings of 'average' or 'difficult', mental health, aggressiveness, acceptance of authority, discipline, and delinquency or disruption at home.[1] This resulted in 325 matched pairs, whom the researchers referred to as 'diagnostic twins'.[20]

Following the matching process, one member of each matched pair was randomly allocated – based on a coin toss – to the treatment group. Overseen by the study director, the process of random allocation was staggered, beginning on 1 November 1937, and ending on 13 May 1939, with the intervention officially starting on 1 June 1939.

In the 1930s, the preventive intervention was described as character development through positive role models, also termed 'directed friendship'.[21] The intervention was similar to today's mentoring programmes.[22] Boys in the treatment group received individual counselling and home visits by paid professional counsellors, known as 'case workers' at the time. Counselling activities included taking the boys on trips and to recreational activities, tutoring them in reading and arithmetic, encouraging them to participate in the YMCA and in summer camps, playing games with them at the project's centre, encouraging them to attend places of worship, and giving advice and general support to the boys' families. Participants were enrolled in the intervention for a mean of 5.5 years, with case workers visiting the treatment boys on average twice a month. The control group received no special services.

In 1942, resource shortages – owing to the country's involvement in World War II – resulted in the study being scaled back to 253 matched pairs and the intervention ending in 1945 (instead of running for 10 years as planned by Cabot). When a boy was dropped from the treatment group, his diagnostic twin in the control group was also dropped.[23] A comparison of all of the remaining pairs – using group differences in arithmetic means – indicated that all differences at baseline between the treatment and control groups on a wide range of variables (e.g. age, IQ, referral to the study as 'average' or 'difficult', mental health) could be ascribed to chance.[1]

There have been four Cambridge-Somerville Youth Study post-intervention assessments of criminal behaviour and other outcomes covering major periods of the life-course, including transition from adolescence to early adulthood, early adulthood, middle-age and old age. Results of the first two assessments were not associated with any statistically significant differences in criminal behaviour.[1,24,25] At 30-years post-intervention (mean age = 47 years), it was found that, compared with boys in the control group, boys in the treatment group were statistically significantly *more* likely to have committed two or more crimes; to have symptoms or signs of alcoholism, mental illness, high blood pressure and heart trouble; to have occupations with lower prestige; and to have died before their 35th birthday.[26,27] In the most recent follow-up assessment, 72 years after intervention (and likely the most prolonged follow-up of randomised cohorts in history), no statistically significant differences between the treatment and control groups were found in age at or cause of death.[28]

Pre-allocation stratification, matched pairs and randomisation: the view from the present

The key benefit of random allocation in experiments is that it 'ensures that any known and unknown differences between groups at baseline are due to chance and not to any systematic bias'.[4] In short, it establishes the basis for like-with-like comparisons. This represents an improvement over traditional matching designs, where observations of known characteristics are used to create 'balanced' groups – but *unknown* differences between groups may remain and lead to allocation bias with consequent lower internally valid comparisons. Random allocation (and alternation) turns 'all systematic sources of bias into random ones'[29]; it is thus expected to eliminate the problem of confounding and is generally seen today as

'necessary to produce the most valid and unbiased estimates of the effects' of social interventions.[30]

It is possible, however, to combine these two approaches. One way has been to first stratify participants (by age, sex or other characteristics) prior to random allocation. A closely related manner, employed by Cabot, has been to create pairs of individuals matched according to particular characteristics, followed by random allocation within each pair either to treatment or to control conditions.

According to its supporters,[31,32] this hybrid study design has several advantages over simple random allocation. First, although randomisation is designed to help eliminate confounding, covariate imbalance is still possible.[31] That is, the treatment and control groups may still differ *by chance*. This is less likely with large samples, but in small trials it can threaten internal validity.[33] Attempts to balance the treatment and control groups by matching on known, measured covariates prior to random allocation might help reduce this possibility. However, 'depending on the choice of variables used to make the statistical adjustments for imbalances, the likelihood of bias may increase rather than decrease'.[34,35]

Second, matching prior to random allocation can improve study power when the matching is effective, meaning that there is a positive within-pair correlation on relevant variables.[36] By decreasing variation within matched pairs on known covariates, matching can improve the precision of estimates of treatment effects – as has been demonstrated using over 5000 simulated datasets.[33] There is some debate, however, about whether these advantages necessarily hold in cases of matched-pair cluster randomisation, whereby units such as households or schools rather than individuals are matched and randomised.[37–40]

Third, random allocation within matched pairs provides a straightforward way of dealing with differential attrition, which can present serious challenges for longer follow-up assessments of controlled trials. Here, the researcher can drop both members of the pair in the event one member is missing. The downside is that this can result in considerable attrition and small final samples. These considerations, of course, represent our present perspective.

Back to history: the context of the social sciences

There was a growing interest in experimentation in the social sciences during the 1920s.[2,41,42] Prior to the advent of alternate or random allocation, comparison groups for experiments were generated within a number of trials using ''systematically balanced designs'',[43] which involved 'matching on prognostic variables'.[4] Early experimenters actually regarded matching designs as preferable to random allocation; this is because chance selection was thought to produce larger errors than matching on measured variables.[2]

Educationalists started to use comparison groups in educational experiments as early as 1908,[44] and indeed used matching rather than random allocation to create comparison groups. For example, an early textbook's treatment of experimental design for education research promoted 'measurement' (i.e. matching) over 'chance selection' (i.e. alternation, rotation or random allocation) for generating equivalent comparison groups.[2] While random *sampling* was agreed to be the best means of obtaining a representative study population, there was no agreement that random *allocation* to comparison groups was necessarily the best means of achieving group comparability. Thus, McCall wrote: 'Measurement, if adequate and accurate, is the best basis for selecting subjects irrespective of their number. Chance selection is merely an economical substitute for measurement, and is practicable only where the number of experimental subjects is sufficiently large'.[2] As Forsetlund et al. observe,

> McCall clearly regarded reliance on chance as inferior to active matching of groups using measures of general ability, and we have been unable to find any account of him having used chance (random allocation) to generate comparison groups in intervention studies.[4]

This can be observed in early experiments in education that used matching rather than random allocation for generating equivalent comparison groups. For example, Winch[45] reported on a series of experiments on memory in 10-year-old children. In each experiment, a class of students was first divided into a treatment group that performed rote memory exercises and a control group that did not. To generate equivalence between groups, the whole class first took a memory test and was then divided into two groups with roughly equal total scores. Following this intervention, the two groups were given substance memory tests (i.e. stories) to evaluate whether the rote memory exercises improved substance memory, as assessed by comparing mean group scores.

In another early example of matching, Thorndike and Ruger[46] reported on an experiment comparing the test outcomes of students exposed to recirculated air compared with those exposed to fresh air. Two groups of students were matched based on the results of a series of practice exercises involving addition,

number and letter checking, and finding and copying addresses. The two groups were then given a series of tests to measure their ability across a number of academic topics. The average initial scores in each group were roughly the same, and the groups were deemed equivalent. From January to April, the treatment group was exposed to recirculated air in their classroom, while the comparison group was only exposed to fresh air (i.e. opening windows). At the end of the semester, the groups were compared in terms of improvement from initial scores across a number of examinations.

In a more advanced educational setting, Chapin[42] noted that 59 experiments were conducted at the University of Minnesota to investigate the effects of class size on academic achievement. In one such experiment, researchers compared a large class of 59 students with a small class of 21 students. Students in each class were assessed on intelligence scores and grades, and 11 students from the small class (treatment group) were matched with 11 students from the large class (control group). The classes had the same instructor, text and method of instruction, and the mean final exam outcomes of the matched subgroups were compared.

Matching based on initial measurements appears to have remained the dominant approach to experimentation in education into the 1930s. There were, however, a few examples of experiments in education using random allocation prior to 1937, all of which appear to have taken place at Purdue University in the late 1920s and early 1930s.[4] For example, Walters[47,48] investigated whether counselling services improved the performance of freshman students who were deemed to be 'academically delinquent'. In one study, freshman students with failing grades were 'divided into three groups by random sampling': a group with instructor counsellors, a group with student counsellors and a control group with no counsellors.[48]

By 1937, and in keeping with his protestations to social workers and others to evaluate their projects and use rigorous methods,[17] Cabot managed to set the bar even higher for evaluating his own delinquency prevention intervention, a decision that would come to be seen as significant for evaluation science. In reporting on the results of the first evaluation of the Cambridge-Somerville Youth Study, Edwin Powers (who served on the study research staff from 1937 to 1947, and directed the study from 1941 to 1951) and Helen Witmer reported that Cabot deemed matching on its own to be insufficient:

> The next question was to determine whether any given boy should fall into the treatment or the control group. It was evident that an arbitrary decision might give rise to a constant error. The proper method of determining this question was, of course, by chance. Accordingly, a coin was flipped and the cases fell into the treatment or comparison groups in accordance with its fall.[1]

Powers and Witmer went on to reflect in their 1951 assessment: 'It was believed that, even if the measures used in the matching were not perfectly reliable, chance would tend to preserve, in groups as large as 325 each, an even balance of important factors'.[1] Later commentators on the study have supported this view: 'The researchers used this paired or fully blocked design because the experimental treatment was lengthy and complex, so they sought to maximize the equivalence of the comparisons they could make'.[32]

Closely tied to this view was the need to overcome additional uncertainties about testing the effects of an intervention on social behaviour.[49] Writing in the book chapter on the matching process, Powers and Witmer concluded with the following:

> This account of the matching process, unavoidably complex, brings to light the difficulties of achieving adequate experimental controls in investigations of therapeutic methods; indeed, in any investigation that concerns itself with personalities or social behavior. For the purpose of this Study, however, it was vital to attempt as sound an equating as possible.[1]

Recent historical research on the origins of the research design of the Cambridge-Somerville Youth Study has suggested two plausible explanations for Cabot's decision to employ matching with random allocation.[5] First, it represented a natural carry-over from Cabot's background in clinical practice and research. This had long been the prevailing view about the research design in general, held by those conducting research on and writing about the study.[50] Second, the combination of matching and random allocation appealed to Cabot on the grounds of using an even more rigorous method of experimentation than what was, in the late 1930s, the standard of the day. In fact, Cabot's decision to employ random allocation within matched pairs was made in the context of a debate among leading statisticians regarding the best method for creating equivalent groups for purposes of experimentation. It took place against the backdrop of Cabot's own dual identity as both social scientist and physician.

In Part 2, we will provide a deeper exploration of the historical context of attempts to compare like with like in medicine and public health.

Declarations

Competing Interests: None declared.

Funding: None declared.

Ethics approval: Not applicable.

Guarantor: BCW.

Contributorship: Each of the authors participated in all parts of study development and writing of the manuscript.

Acknowledgements: We are especially grateful to the reference staff at Harvard University Archives, Harvard Law School Library's Historical and Special Collections, and the Center for the History of Medicine at the Countway Library. We also wish to thank Iain Chalmers for suggesting the central questions that guided our research, as well as for his sage advice and insightful comments throughout the development of this article.

Provenance: Invited article from the James Lind Library.

References

1. Powers E and Witmer HL. *An Experiment in the Prevention of Delinquency: The Cambridge-Somerville Youth Study*. New York: Columbia University Press, 1951.
2. McCall WA. *How to Experiment in Education*. New York: Macmillan, 1923.
3. Weisburd D and Petrosino A. Experiments, criminology. In: Kempf-Leonard K, ed., *Encyclopedia of Social Measurement*. San Diego: Academic Press, 2004:877–884.
4. Forsetlund L, Chalmers I and Bjørndal A. When was random allocation first used to generate comparison groups in experiments to assess the effect of social interventions? *Econ Innov New Technol* 2007; 16: 371–384.
5. Welsh BC, Dill NE and Zane SN. The first delinquency prevention experiment: a socio-historical review of the origins of the Cambridge-Somerville Youth Study's research design. *J Exp Criminol* 2019; 15: 441–451.
6. Welsh BC, Zane SN and Rocque M. Delinquency prevention for individual change: Richard Clarke Cabot and the making of the Cambridge-Somerville Youth Study. *J Crim Justice* 2017; 52: 79–89.
7. O'Brien L. 'A bold plunge into the sea of values': the career of Dr. Richard Cabot. *N Engl Q* 1985; 58: 533–553.
8. Evison IS. Pragmatism and idealism in the professions: the case of Richard Clarke Cabot, 1868–1939. Unpublished dissertation, University of Chicago, Chicago, 1995.
9. Anon. Deaths: Richard Clarke Cabot. *J Am Med Assoc* 1939; 112: 2079.
10. White PD. Richard Clarke Cabot. *N Engl J Med* 1939; 220: 1049–1052.
11. Crenner C. *Private Practice: in the Early Twentieth-Century Medical Office of Richard Cabot*. Baltimore: Johns Hopkins University Press, 2005.
12. Cabot RC. *A Guide to the Clinical Examination of the Blood for Diagnostic Purposes*. New York: W. Wood, 1896.
13. Hajar R. The clinicopathologic conference. *Heart Views* 2015; 16: 170–173.
14. Cabot RC. *Social Work: Essays on the Meeting-Ground of Doctor and Social Worker*. Boston: Houghton Mifflin, 1919.
15. Stuart PH. Individualization and prevention: Richard C. Cabot and early medical social work. *Soc Work Mental Health* 2004; 2: 7–20.
16. Cabot RC. *Adventures on the Borderlands of Ethics*. New York: Harper, 1926.
17. Cabot RC. Treatment in social case work and the need of criteria and of tests of its success and failure. *Hosp Soc Serv* 1931; 24: 435–453.
18. Powers E. An experiment in prevention of delinquency. *Ann Am Acad Pol Soc Sci* 1949; 261: 77–88.
19. Cabot RC. Letter to Miss Gertrude Duffy, June 3, 1935. Box 97. Richard Clarke Cabot Papers, HUG 4255, Harvard University Archives.
20. Cabot PS deQ. A long-term study of children: The Cambridge-Somerville Youth Study. *Child Dev* 1940; 11: 143–151.
21. Powers E. Some reflections on juvenile delinquency. *Fed Probat* 1950; 14: 21–26.
22. Eddy JM, Martinez CR Jr, Grossman JB, Cearley JJ, Herrera D, Wheeler AC, et al. A randomized controlled trial of a long-term professional mentoring program for children at risk: outcomes across the first 5 years. *Prev Sci* 2017; 18: 899–910.
23. McCord J. A longitudinal study of personality development. In: Mednick SA, Harway M, Finello KM, eds, *Handbook of Longitudinal Research*, vol. 2. New York: Praeger, 1984: 522–531.
24. McCord J and McCord W. A follow-up report on the Cambridge-Somerville Youth Study. *Ann Am Acad Pol Soc Sci* 1959; 322: 89–96.
25. McCord W and McCord J. *Origins of Crime: A New Evaluation of the Cambridge-Somerville Youth Study*. New York: Columbia University Press, 1959.
26. McCord J. A thirty-year follow-up of treatment effects. *Am Psychol* 1978; 33: 284–289.
27. McCord J. Consideration of some effects of a counseling program. In: Martin SE, Sechrest LB, Redner R, eds, *New Directions in the Rehabilitation of Criminal Offenders*. Washington, DC: National Academy Press, 1981: 394–405.
28. Welsh BC, Zane SN, Zimmerman GM and Yohros A. Association of a crime prevention program for boys with mortality 72 years after the intervention: a follow-up of a randomized clinical trial. *J Am Med Assoc Network Open* 2019; 2: 1–11 (e190782).
29. Rubin D. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Ed Psychol* 1974; 66: 688–701.
30. Farrington DP and Welsh BC. A half century of randomized experiments on crime and justice. *Crime Justice* 2006; 34: 55–132.

31. Ariel B and Farrington DP. Randomized block designs. In: Piquero AR, Weisburd D, eds, *Handbook of Quantitative Criminology*. New York: Springer, 2010: 437–454.
32. Weisburd D and Gill CE. Block randomized trials at places: rethinking the limitations of small N experiments. *J Quant Criminol* 2014; 30: 97–112.
33. Balzer LB, Petersen ML and van der Laan MJ; SEARCH Consortium. Adaptive pair-matching in randomized trials with unbiased and efficient effect estimation. Stat Med 2015; 34: 999–1011.
34. Chalmers I. Evaluating the effects of care during pregnancy and childbirth. In: Chalmers I, Enkin M, Keirse MJNC, eds, *Effective Care in Pregnancy and Childbirth*. Oxford: Oxford University Press, 1989:3–38.
35. See also Detre KM, Peduzzi P and Chan Y-K. Clinical judgment and statistics. *Circulation* 1981; 63: 239–240.
36. Wacholder S and Weinberg CR. Paired versus two-sample design for a clinical trial of treatments with dichotomous outcome: power considerations. *Biometrics* 1982; 38: 801–812.
37. Campbell MJ, Donner A and Klar N. Developments in cluster randomized trials and Statistics in Medicine. *Stat Med* 2007; 26: 2–19.
38. Donner A and Klar N. Pitfalls of and controversies in cluster randomization trials. *Am J Public Health* 2004; 94: 416–422.
39. Ivers NM, Halperin IJ, Barnsley J, Grimshaw JM, Shah BR, Tu K, et al. Allocation techniques for balance at baseline in cluster randomized trials: a methodological review. *Trials* 2010; 13: e120.
40. Imai K, King G and Nall C. The essential role of pair matching in cluster-randomized experiments, with application to the Mexican universal health insurance evaluation. *Stat Sci* 2009; 24: 29–53.
41. Brearly HC. Experimental sociology in the United States. *Social Forces* 1931; 10: 196–199.
42. Chapin S. The problem of controls in experimental sociology. *J Ed Sociol* 1931; 4: 541–551.
43. Box JF. R.A. Fisher and the design of experiments, 1922-1926. *Am Stat* 1980; 34: 1–7.
44. Winch WH. The transfer of improvement in memory in school-age children. *Br J Psychol* 1908; 2: 284–293.
45. Winch WH. The transfer of improvement in memory in school-age children, II. *Br J Psychol* 1910; 3: 386–405.
46. Thorndike EL and Ruger GJ. The effects of outside air and recirculated air upon the intellectual achievement and improvement of school pupils. *School Soc* 1916; 4: 260–264.
47. Walters JE. Seniors as counselors. *J Higher Ed* 1931; 2: 446–448.
48. Walters JE. Measuring effectiveness of personnel counseling. *Personnel J* 1932; 11: 227–236.
49. Claghorn KH. The problem of measuring social treatment. *Soc Service Rev* 1927; 1: 181–193.
50. McCord J. The Cambridge-Somerville Study: a pioneering longitudinal experimental study of delinquency prevention. In: McCord J, Tremblay RE, eds, *Preventing Antisocial Behavior: Interventions from Birth Through Adolescence*. New York: Guilford Press, 1992: 196–206.