Julian C. Jamison*

The Entry of Randomized Assignment into the Social Sciences

https://doi.org/10.1515/jci-2017-0025
Received November 4, 2017; revised January 20, 2019; accepted February 11, 2019

Abstract: Although the concept of randomized assignment in order to control for extraneous confounding factors reaches back hundreds of years, the first empirical use appears to have been in an 1835 trial of homeopathic medicine. Throughout the 19th century there was a growing awareness of the need for comparison groups, albeit often without the realization that randomization could be a clean method to achieve that goal. In the second and more crucial phase of this history, four separate but related disciplines introduced randomized control trials within a few years of one another in the 1920s: agricultural science; clinical medicine; educational psychology; and social policy (specifically political science). This brought increasing rigor to fields that were focusing more on causal relationships. In a third phase, the 1950s through 1970s saw a surge of interest in more applied randomized experiments in economics and elsewhere – both in the lab and especially in the field.

Keywords: randomization, RCT, field experiment, lab experiment, confounding, causality, history of science

> *Development of Western science is based on two great achievements: the invention of the formal logical system (in Euclidean geometry) by the Greek philosophers, and the discovery of the possibility to find out causal relationships by systematic experiment (during the Renaissance).*
> Albert Einstein (1953)

1 Introduction

The quote above appears in Pearl [1], a comprehensive reference on the statistics of causality. In an informative history of the "art and science of cause and effect", Pearl refers to the randomized experiment as "the only scientifically proven method of testing causal relations from data, and to this day, the one and only causal concept permitted in mainstream statistics." Interestingly, although Einstein dates the idea of causal experiments – any relationship to randomization goes unspoken by him – to the Renaissance, Pearl claims that it waited upon Fisher in the 1930s. As we shall see, they were both partially correct: the explicit idea appeared hundreds of years ago, but only in an isolated fashion; it did not become a holistic scientific construct until the 1920s.

Einstein and Pearl agreed upon the central role of rigorous experiments in determining causality, which has long been understood and accepted in the physical and biological sciences but has undergone a more recent rise in the social sciences. The basic idea is straightforward: Suppose you wish to test the relative effect of treatment (or intervention, broadly construed) A vs treatment B – one of which could be a null treatment or

Article note: I thank Art Boylston, Jesse Bump, Austin Frakt, Markus Goldstein, Don Green, Judy Gueron, Glenn Harrison, Dean Jamison, Dean Karlan, Chris Lysy, Jack Molyneaux, Andreas Ortmann, Lior Pachter, Charlie Plott, Tasmia Rahman, Al Roth, and an editor and two reviewers for helpful discussions and input. My interest was first piqued when I read an article by Druin Burch in *Natural History* (June 2013) that mentioned van Helmont's early contribution to the topic. Thanks also go the James Lind Library, which is a wonderful resource for the relevant medical history, including translations of some of the early documents referenced below.

***Corresponding author: Julian C. Jamison,** Department of Economics, University of Exeter Business School, Exeter, United Kingdom; and J-PAL; and The World Bank eMBeD unit, Washington, United States, e-mail: j.jamison@exeter.ac.uk, ORCID: https://orcid.org/0000-0003-2671-1153

status quo. Take a large number of subjects (individuals, schools, firms, villages, etc.) and divide them randomly into two groups. The first group gets A and the second group gets B; other than that their experiences are identical. Since the division was random and the sample size was large, we can be highly confident that the two groups started out with the same average levels of all relevant characteristics, both observable and unobservable. Therefore any aggregate differences between the groups measured after the experiment can be causally identified with the corresponding treatments.

Naturally there are assumptions to be made, and there are many complications that arise in specific instantiations of this approach. The goal of this paper is not to enter into the debate about the relative merits of randomization, although certain elements of that debate will make appearances throughout. However, it is abundantly clear that they are not always the right tool for the job. Imagine the idea of testing the efficacy of parachutes versus a control group on mortality rates when disembarking an airplane at a height of 10000 feet (3000 meters) above ground level. Not only would a randomized experiment be unethical, it would be completely unnecessary. That experiment has never been undertaken (see Smith and Pell [2] for a review of the literature), and yet we are convinced from theory and from analogy and from common sense that we know the actual relative efficacy of the two approaches.[1]

Let us begin with a definition for our central concept of *randomized assignment*. An empirical research study consists of one or more *observations*, where an observation consists of measured *conditions* (what was done, when, to whom or what, etc.) and measured *outcomes* (what resulted). For instance, an observation might be that a 29-year-old male was made to sit for four minutes alone in a dark room in the morning (the conditions), after which he coughed and ran away (the outcomes). Randomized assignment occurs when the value of at least one condition is assigned randomly across observations. In the example above, the researcher might randomly assign some subjects to sit for four minutes and others for ten minutes. Or the researcher might assign the same subject, arriving on various occasions at different times of day and wearing different clothes, to sometimes sit in a dark room and sometimes in a brightly lit one.

The value of randomized assignment is that it implies that the measured status of the condition which is randomized is unconfounded with any of the other conditions, and hence that any variation in outcomes as a function of that status must (in expectation) be due to the influence of the randomized condition – at least within the set of potential conditions examined. In the second example above, the researcher can causally determine the impact of light on any measurable outcomes of interest, but only for that subject and only when sitting alone in a room at times of day when the experiment is carried out. Randomization yields 'internal validity' (comparing "like with like" in the phrase of Chalmers [3]) but not 'external validity'. Of course the larger the relevant population or domain of observed conditions, the more widespread and robust is the conclusion.

This is distinct from the random sampling of subjects from a larger population in order to draw conclusions that are representative of the entire population. That too is randomization, and it has an important place in social science, but the rationale and history are not the same; see Fienberg and Tanur [4].[2] Although the two are different in purpose and application, there has been some confusion over the years in the philosophy of science literature.[3] The current paper is focused only on the concept of randomization for causal inference.

A second alternative usage of randomization, also outside the scope of this paper, is fairness. This covers everything from access to limited medical resources, to prize lotteries for allocating valuable social assets.

1 Note also that sometimes causality is irrelevant in science, e. g. when estimating the speed of light or how quickly a feather falls in a vacuum.

2 They mention a creative and early use of randomization carried out by Mahalanobis [5] in India while surveying factory workers. Instead of assigning one enumerator to each area, he divided the areas into five independent random samples and had each enumerator work in every area. This is a nice example of embedding experimental design (involving randomized assignment) into survey design.

3 Urbach [6] claims that from a Bayesian perspective, randomization can be of no use in testing statistical hypotheses; Papineau [7] rightly responds that that only holds true for random sampling and not for randomized experimentation of the type considered here.

There are fewer examples of this type of randomization within the history of science,[4] although fairness has certainly been used as an argument to sway policy-makers who are uncertain about the ethics of randomizing the assignment of interventions.

Finally, randomization is relevant for the validity of certain specific statistical tests. Many formal inference analyses are predicated on assumptions regarding the data-generating process which can only be satisfied, or are more easily satisfied, when there has been random allocation into treatments. In this sense it is closely related to the notion of randomized assignment for causal inference which is our focus, although they are not identical. Whilst the broad history of statistical inference is also not the focus here, it would be remiss not to mention the ground-breaking work of Ronald Fisher.

Fisher took a position in 1919 as statistician at Rothamsted agricultural research station, where his main job was to analyze the piles (literally) of existing data from previous 'experiments'. He started to develop theories of his own about how to optimally run experiments, culminating in the publication of his classic book [10]. Although he had advocated randomization (in the sense of e. g. randomly allocating different seeds or fertilizers to different plots of land) as a theoretical concept in 1925, his first empirical publication that used randomization as a technique was two years later (Eden and Fisher [11]). Although Fisher was not the first to apply randomized assignment, his statistical tools and tireless advocacy of this approach played a major role in its later widespread adoption.

The goal of this paper is to chart the initial introduction of randomized assignment for inference, both in actual practice and as a conceptual construct, into various intersecting and intertwining branches of medical and social science, especially psychology, economics, and policy. One motivation for doing so is the growing success of this approach (at least in terms of relative popularity and claimed standing as a 'gold standard'). However the focus is on the narratives and conditions surrounding the entry (and occasional re-entry) in and across disciplines, including especially the intellectual environment of the early adoptions, rather than on the factors that did or did not lead to later success and their relative merits. The main contributions to the existing literature are: corrections of various misstatements regarding the original appearances of randomization; earliest-known examples of randomized assignment in a variety of disciplines; bringing together for the first time the discussion across medicine and multiple social sciences; and given all these elements, being able to draw conclusions about patterns regarding the viability and acceptance of randomization when it was a novel scientific research construct.

The remainder of the paper proceeds as follows. Section 2 provides the early background, tracing various isolated instances of both randomized assignment and not-quite-randomized assignment. Section 3 briefly discusses the history of randomized assignment in clinical medicine, the field with which it is most closely associated. Then we turn to social science proper, beginning with psychology in section 4; economics in Section 5; and finally social policy (including public health) in Section 6. Section 7 provides concluding remarks.

2 Prelude

In the Book of Daniel (1:8-16), Daniel does not wish to consume the royal fare and suggests a test: he and his friends will eat only pulse and drink only water for ten days, after which the official can compare their health to those of the young men consuming the royal fare.[5] Although this episode nicely captures the idea of a comparison group, there is an obvious problem with endogeneity and selection bias. Hence not only is randomization in any form missing, but there is no sense of a controlled or fair experiment. Furthermore, in terms of chronology, although Charles Darwin did not advance his theory of natural selection until the

4 See Silverman and Chalmers [8] for general discussion of the latter, and many quotations below as a sample of the former. They find few examples with explicit arguments favoring fairness, and e. g. I do not agree with their interpretation of van Helmont [9].
5 A similar story is reported in the Book of Numbers.

mid-19th century (Darwin [12]), Nature had conveniently begun to experiment via randomization in the context of allopatric speciation after vicariance to test his theory some millions of years earlier.[6]

It is likely that scholars in antiquity understood the basic idea of comparing two similar groups in order to reliably test interventions. However, the first written documentation of which I am aware is by the poet Petrarch [13] in a letter to Boccaccio:

> *I solemnly affirm and believe, if a hundred or a thousand men of the same age, same temperament and habits, together with the same surroundings, were attacked at the same time by the same disease, that if one half followed the prescriptions of the doctors of the variety of those practicing at the present day, and that the other half took no medicine but relied on Nature's instincts, I have no doubt as to which half would escape.*

Although there is no mention of randomization and no concrete suggestion to collect data, it is clear that the goal was to devise two groups that were as similar as possible. It is also clear what Petrarch thought of doctors. However, this example serves to illustrate that the general idea was in circulation and yet simultaneously that it was not part of regular practice in terms of implementation, implying that it held no special place in convincing physicians or governments of efficacy.

Thus we arrive at the generally accepted first surviving mention of randomized assignment, due to Flemish chemist and physician Jan Baptist van Helmont. Everyone at the time, including van Helmont, believed that bloodletting was a fantastic cure for most ailments. However he believed that evacuation (i. e. inducing vomiting and defecation) was an even better approach, and he proposed a simple way to settle the argument once and for all:

> *Let us take out of the Hospitals, out of the Camps, or from elsewhere, 200 or 500 poor People, that have Fevers, Pleurisies, etc. Let us divide them in halfes, let us cast lots, that one half of them may fall to my share, and the other to yours; I will cure them without bloodletting... we shall see how many Funerals both of us shall have.*

For better or worse, there is no evidence that this test was ever put into practice, but the idea is up to modern standards.[7] When was this written? Nobody knows precisely. Many articles cite van Helmont [14], but that is the first English translation (from which the above quote is taken) of the original Latin publication (van Helmont [9]). Even that year is clearly too late, since van Helmont died in 1644; some of his writings were controversial, so the corpus did not see the light of day until his son brought them out posthumously.

Despite van Helmont's mistaken (but typical) views on clinical practice, he was an inquisitive and thoughtful researcher, a Renaissance man befitting Einstein's quote above. This will be a theme for many of those who intersect the origins of randomization, suggesting that each successive development was not nearly as simple as it appears in retrospect. Along those lines, we proceed by mentioning two more notable occurrences in the history of clinical trials, albeit unrandomized.

James Lind was a Scottish naval surgeon (a position that did not require extensive medical training, although he later earned an MD) who was an early believer in the theory that citrus fruits could help cure scurvy, which we now know is indeed caused by a deficiency of vitamin C. He provided a partial test of this claim on a voyage in 1747 (published in Lind [15]), when he divided 12 afflicted sailors into six pairs and gave each pair a different treatment – one of which was two oranges and one lemon daily.[8] He made a point of the fact that the men were similar to begin with and were treated identically in all ways apart from the experimental variation:

---

6 Lior Pachter brought this prehistoric example to my attention.

7 One slight drawback is that he appears to suggest that the patients first be divided into two groups, without specifying how that is to be done, after which randomization occurs. Presumably he has in mind some method for division that is approximately symmetric, but even if not it constitutes randomizing at what is now referred to as the cluster level – admittedly low powered in this case, but still more rigorous than many modern papers.

8 Other treatment arms included seawater, sulfuric acid, and spicy paste plus barley water. These did not prove efficacious.

Their cases were as similar as I could have them. They all in general had putrid gums, the spots and lassitude, with weakness of their knees. They lay together in one place, being a proper apartment of the sick in the fore-hold; and had one diet common to all.

While Lind did not include an untreated control group, Watson [16] did exactly that in a study of smallpox variolation: as he put it, "it was proper also to be informed of what nature unassisted, not to say undisturbed, would do for herself." Although both men explicitly attempted to perform their tests on a homogeneous population, as well as to maintain parity apart from the treatments of interest, neither of them suggests randomization or any other method to objectively achieve such a goal.

Finally, it is worth mentioning a somewhat flamboyant experiment in the arena of animal husbandry performed by famous microbiologist Louis Pasteur in 1881. He was attempting to publicly prove that he had developed an animal anthrax vaccine, so he asked for 60 sheep and split them into three groups: 10 would be left entirely alone; 25 would be given his vaccine and then exposed to a deadly strain of the disease; and 25 would be untreated but also exposed to the virus. It is unclear whether sheep have more or less natural variation than Fisher's plots of land, but there is no mention of randomization or selection bias in the paper (Pasteur [17]). Perhaps this was not a major issue given the stark results: all of the exposed but untreated sheep died, while all of the vaccinated sheep survived healthily.

3 Medicine

Many people associate the RCT (randomized control trial, which involves randomization into a control group and one or more 'treatment' groups for comparison) with medicine, where it has come to be viewed as the 'gold standard'.[9] Partly for this reason; partly because – as described above and below – it was primarily clinicians who took the first steps along this path; and partly because the timing of randomized assignment entering the establishment in medicine so closely coincided with that in other fields; it makes sense to include some discussion of medicine in this context even if it is not properly a social science. Of course, many of the same factors around human behavior are at play.

After the early empirical approaches of Lind and Watson, the next big step was taken by "a society of truth-loving men" in Nuremberg in 1835 (see Löhner [19] and discussion in Stolberg [20]). In order to evaluate the effect of a salt-based homeopathic treatment, 100 local citizens were recruited to volunteer. 100 vials were numbered consecutively, mixed together, and then separated into two groups of 50. All vials were filled with pure snow water, and the salt potentiation was added to one of the groups. The experimenters noted which numbered vials these corresponded to, and the resulting list was sealed and kept secret until the end of the trial. After the vials were once again well mixed with one another, the participants each ingested the contents of one vial, reporting their symptoms two weeks later "in order to compare the effect with the cause". The results suggested no effect of the homeopathic remedy, although since outcomes were self-reported it is possible that there was bias introduced at that stage – namely reporting nothing so as to match the control.

On the one hand this is a remarkable event: it clearly constituted randomized assignment (the first instantiation of which I am aware) to treatment and control, as well as being double blind and remarkably transparent about procedures (prospectively!) and about attrition. On the other hand, it does not seem to have made much impact on the general practice of medical trials, and even now it is neither widely known nor appreciated. That being said, it was not an entirely isolated incident: dating to van Helmont in the early 17th century and Mesmer in the late 18th century,[10] much of the drive for rigorous testing was due to the high-

9 This term was apparently borrowed from monetary economics, where it refers to the actual metal gold. The analogy is that both approaches describe the best measure or method of comparison available at a given time. See Claassen [18] for discussion of the history of the term and the respect toward RCTs in medicine.

10 Mesmer (1781) proposed but did not carry out a challenge to his colleagues regarding his theory of 'animal magnetism', in which he writes: "In order to avoid any later argument and all the questions that could be raised about differences in age, in

stakes battle between homeopathy and allopathy. In the Nuremberg case, perhaps one of the reasons it had less impact was that the participant subjects were not in need of a cure; they were simply being tested to see if they noticed any effects.

Although randomization did not become common practice for another century, the idea of demanding a proper comparison group was gaining adherents. For instance, later in the 19th century, we find examples of doctors using alternation "to avoid the imputation of selection" (Balfour [22]) or to induce "an equally large number of randomly selected patients treated as usual" (Fibiger [23]).[11] Note that although Fibiger obviously believed that what he did was equivalent to random allocation, which was indeed his goal, what he actually did was to alternate treatment based on the day the patient arrived at the hospital. From a modern perspective this looks importantly distinct, but at the time these were all simply methods to produce a valid control group (and in practice alternation likely worked quite well in most instances).

The modern era of RCTs in medicine begins with Colebrook [24], in which "drawing lots" was used to decide which kids would be irradiated (it's not as bad as it sounds) – but if the parents refused consent then those children were added to the control group, which undoes much of the point of randomization but still [re]introduces the concept. The rigorous version appears two years later in Doull [25], a study of the effect of ultra-violet light on the common cold. Doull worked at the Johns Hopkins School of Public Health and needed to figure out how to allocate his subjects into three groups in a manner that would allow for valid comparisons and analysis. According to Marks [26], he consulted with a local biostatistician with a doctorate in mathematics, who suggested using colored dice to randomly allocate patients. Note the similar timing for these randomizations in clinical medicine as in agriculture (Eden and Fisher [11]).

The final piece of the medical puzzle falls into place with the famous streptomycin trial for tuberculosis (Medical Research Council [27]).[12] This is probably the most famous RCT in history, and many people have erroneously claimed that it was in fact the first RCT in history. The design was the brainchild of Austin Bradford Hill, whose degree was in economics (earned while recovering from tuberculosis himself) but who worked as a biostatistician and epidemiologist.[13] In addition to the important step of highlighting the need for randomization and of promoting it – he later wrote down influential formal criteria for imputing causality – Bradford Hill also promulgated another key aspect in the 1948 paper: the explicit idea of using randomization to consciously conceal foreknowledge, i. e. to "blind" the experimenter to treatment status whenever possible.

4 Psychology

Human sensation, like psychical phenomena and mesmerism, was for most of history not considered a domain susceptible to quantitative scientific analysis. That began to change with the work of Gustav Fechner in the mid-19th century, who initiated the field of psychophysics (Fechner [29]) along with Ernst Weber. In particular Fechner studied sensitivity of physical perception: e. g. how finely can a subject distinguish two masses, as a function of the base weight and the marginal difference between them? Although he deserves much credit for introducing concepts such as empirical experimentation and mathematical data analysis to this entire field, his methods were far from perfect. In particular Fechner experimented on himself; for ex-

temperament, in diseases, in their symptoms etc. the assignment of the patients shall be made by the method of lots." His ideas were later tested, though without explicitly randomized assignment, by a 1784 commission led by Benjamin Franklin for the king of France (see Kaptchuk [21]).

11 This was a study of diphtheria; Fibiger later won the Nobel Prize for his work on cancer.

12 The experiments for whooping-cough reported in Medical Research Council [28], although published three years later, used an exactly analogous experimental approach and were actually begun several months earlier than those in the better-known streptomycin paper.

13 Bradford Hill later earned fame for his [non-randomized] work exhibiting a link between smoking and lung cancer; he was knighted in 1961.

ample in the perception experiments he knew all the relative weights in advance. He believed that he could consciously control for any resulting bias.

Müller [30] took the next step, splitting the roles of subject and experimenter. He concurrently emphasized the notion of presenting stimuli in an irregular order (*in buntem Wechsel*; see Dehue [31]), but neither he nor Fechner employed randomization – although Müller did eventually start to promote the use of explicit randomization around the turn of the century. Meanwhile randomization was used by Richet [32] but only as an inherent component of the stimulus itself. This is because he was testing telepathy, a topic that was all the rage in Europe at the time and which was eminently suitable for rigorous evaluation.[14] Randomly chosen playing cards were studied intently by one person, who tried to mentally pass the information to another. Thus, the randomization was not carried out in order to compare different treatments.

We turn now to one of the more well-known protagonists in this arc, and indeed the proponent of what is likely the first instance of randomized assignment in social science, namely Charles S. Peirce. According to Stigler [34], Peirce was educated at home by his father, a mathematics professor at Harvard. He was ambidextrous and had the habit of writing questions with his left hand while writing the answers with his right hand. By December 1883, when he began the series of experiments described below, he was on the faculty at Johns Hopkins, where he was primarily known as a philosopher but also worked in physics, mathematics, cartography... and psychology.

Fechner had postulated that for any given base weight, there was a minimum additional weight below which it was impossible to perceive any difference, i. e. where the two felt exactly the same. Peirce disagreed, believing that even for very small differences, if subjects were forced to choose which one they thought was heavier,[15] they would be correct slightly more often than they were wrong. Along with a student of his named Jastrow, he proceeded to test his hypothesis in a series of experiments from December 1883 to April 1884. They took turns as experimenter and subject, which Fechner was naturally unable to do while working alone, with the experimenter drawing playing cards whose color (red or black) determined whether weight was first added and then taken away; or vice-versa. As Peirce and Jastrow [35] note in their paper:

> *A slight disadvantage in this mode of proceeding arises from the long runs of one particular kind of change, which would occasionally be produced by chance and would tend to confuse the mind of the subject. But it seems clear that this disadvantage was less than that which would have been occasioned by his knowing that there would be no such long runs if any means had been taken to prevent them.*

This is precisely the type of concern that Fisher and Gosset would argue about almost 50 years later in a different context: trading off the reduction of noise via regularity where possible, versus using randomization to equalize everything but only in expectation. We still argue about such things today.

Was this an example of randomized allocation into treatment and control groups of subjects? Clearly not. Forsetlund et al. [36] argue that Peirce's randomization served only to blind the subject and not to assess the effect of an intervention on an outcome. However this seems like a false dichotomy: Peirce was randomizing not merely to blind the subject (as Richet had) but also to allow for comparisons of "like with like". Because of the structure of the experiment, there were two possible conditions (base weight first or supplemented weight first), and Peirce wanted to ensure that the two corresponding sets of observations differed only in this respect. This certainly required the subject not to know in advance which one came first; but even if the subject didn't know, it could have been the case that one condition was systematically different from the other (e. g. they in fact found that it is easier to perceive increasing than decreasing weights). Randomization solves this problem neatly in a way that no deterministic ordering, however carefully balanced and thought out, can

14 Hacking [33] provides illuminating historical details on this development. As far as results go, Richet was the first of many authors not to find evidence for supernatural powers.

15 Forced choice was an innovation along with randomization, albeit not as momentous. Additionally, subjects were asked to express confidence in their choice on a scale of 0–3, which was a further innovation that is still underutilized today.

do.[16] Furthermore, the paper notes that randomization implies "any possible psychological guessing of what change the operator was likely to select was avoided."

For practical reasons Peirce randomized over stimuli rather than over subjects (the analogy is that one group of individuals would always receive the base weight first, while the other group would receive the supplemented weight first), but the purpose and the implications are the same. This is why we focus here on "randomized assignment" rather than "randomized allocation", and it is clear that Peirce understood the importance of this approach – although it did not immediately catch on with others. Fortunately for him, as it happens, Peirce's substantive hypothesis was at least confirmed in the data.

Early efforts to apply experimental techniques in controlled settings outside the lab also lay with psychologists, although in this case it was educational psychology at the forefront. Starting around the turn of the century there were many studies of learning in classrooms, and a book by McCall [37] on experimental design in education highlights randomization as a particularly efficient approach for avoiding selection bias and other spurious influences. However, no empirical studies cited by McCall involving actual randomization have been found; all extant sources are either silent on the matter or use some form of matching to create a control group for comparison.

By the early 1920s, the importance of a rigorously equivalent comparison group had become clear. Dearborn and Lincoln [38] divided pupils "arbitrarily according to the seating arrangements" but not explicitly randomly; indeed the seating was unlikely to have been random. The earliest definitive examples that I have located (predating what has been found in the existing literature on this topic) appear in the *Journal of Educational Psychology*: Shaffer [39] writes that "five experimental groups were made up by random selection" and Clark [40] writes that "subjects were placed at random in four groups of eight each." There is no particular reason to believe that these were the absolute first such use of the technique in this field, but it is at least highly suggestive that the first conscious use was between 1923 (given that there are no examples in McCall's book that year, despite the author being particularly interested in the method) and 1927. Perhaps more importantly, we observe that by the time of its casual mention in these two publications, randomization was methodologically unremarkable within that field.

5 Economics

Unlike their non-laboratory brethren, experimental economists took to randomization very quickly, as had their counterparts in psychology. Although somewhat late to the game in the grand scheme of things, these researchers tended to be deeply careful about their hypotheses and assumptions, which led to multiple distinct uses of randomization – some but not all of which fall under the category of randomized assignment. In addition, like some of the early agricultural experimenters, they tended to focus on the role of theory in their models and analysis; sometimes there was no need for a control group because theoretical predictions provided the point of comparison.

The first laboratory experiment in economics (Thurstone [41])[17] was concerned with within-subject consistency of choices and did not use randomized assignment. However, we find a creative and early use of randomization in consumer choice in Davidson et al. [43]: in the context of measuring utilities and subjective probabilities, the authors made their own dice with nonsense syllables (such as "ZEJ") on which subjects were asked to bet. In order to be absolutely certain that the results weren't driven by people choosing on the basis of e. g. innate preference or familiarity for a particular sequence of letters, "...the choice of winning nonsense syllable was randomized." This is precisely the idea of randomization in order to control for unobservable confounding factors, despite not being an evaluation of an intervention.

16 One could argue that Richet's experiment had similar features, but intentions matter. Richet used randomization within the stimuli because there was quite literally no alternative, whereas Peirce was breaking with the standard protocol of Fechner and even Müller. He consciously introduced randomization in order to be certain that everything other than what he was interested in studying would be controlled for, just as Fisher did many years later.

17 See Roth [42].

Meanwhile Chamberlin [44] reported on a market experiment with demand and supply curves induced by assigning separate values to individuals who served as either buyers or sellers. Implicit in his procedure was that this was done randomly; Smith [45] reports on a series of market experiments from the late 1950s in which the separation is explicit: "The group of subjects is divided at random into two subgroups, a group of buyers and a group of sellers." This certainly constitutes randomized assignment, but note that the purpose was not to compare buyers against sellers or to avoid selection bias. In many ways it is reminiscent of Peirce and Jastrow [35]: randomization is consciously used to control for any potential bias or asymmetry, including on the part of the experimenter, but it is not used to specifically compare treatments or interventions.

The third major topic within early experimental economics, in addition to individual choice and competitive markets, was game theory: models of strategic interaction. Kalisch et al. [46] studied multi-player games of cooperation, comparing the predictive ability of various equilibrium solution concepts. They were interested in one-shot games rather than the effects of repeated coalitions, so they "rotated" the players after each trial; this wasn't quite randomization but it served a related purpose. In terms of disciplinary background, this was a collaboration of mathematicians turned game theorists. A few years later Atkinson and Suppes [47], also not economists by training,[18] analyzed different learning models in two-person zero-sum games, and they explicitly "randomly assigned" pairs of subjects into one of three different treatment groups. This is the earliest instance of random assignment in experimental economics, for purposes of comparing treatments, that has been found to date.

The mix of disciplines in the early years of experimental economics was broad and clearly invigorating. In addition to mathematicians and philosophers (both Davidson and Suppes fit the latter camp) bringing experience in mathematical decision theory, there were importantly the psychologists such as Atkinson and especially Sidney Siegel, a coauthor in the Davidson et al. [43] paper. Economics was more often interested in testing the implications and predictions of specific theories, which does not necessarily require any comparison at all, or in comparing and contrasting the fidelity of various theories to data. In order to optimally organize all these experiments, there were a large number of methodological procedures borrowed from psychology. Siegel was a proponent of many of them, although with no special focus on randomization, and he worked hard to make these new techniques available to the world of economics, including a fruitful collaboration with economist Lawrence Fouraker on studies of bargaining and cooperation (Siegel and Fouraker [49]).

Although Siegel and others were publishing in psychology journals, most of the economics papers discussed here ended up as unpublished manuscripts or book chapters. Chamberlin [44] appeared in an economics journal, but does not explicitly mention randomization. On the other hand, Smith [45] is in an economics journal, discusses randomized assignment, and became highly influential in the development of the field.[19] Although Smith always gave much general methodological credit to Siegel (see Smith [51]), who unfortunately died prematurely, it is not clear whether the notion of randomized assignment was directly borrowed from psychology or was instituted independently as a natural reaction to the environment. What is clear is that he and the rest of the first generation of economists who were full-time experimentalists, such as Charles Plott, continued to use randomization not only for basic division into treatment groups but also (as many others mentioned in this survey) to control for anything unexpected that may have caused different outcomes in different trials.[20]

18 Remarkably, Patrick Suppes was also a coauthor in Searle et al. [48], the first RCT in development economics, which is discussed below; he was also a coauthor in the Davidson et al. [43] article mentioned just above. Suppes was an analytic philosopher who worked in fields as diverse as quantum mechanics, decision theory, and psychology.

19 Suppes and Carlsmith [50] came out slightly earlier that year, albeit in a less widely-read economics journal, and also explicitly randomized subjects into one of two experimental groups. Partly because the topic of that paper and related ones above did not flourish to the same extent, and partly because the authors proceeded to other work, it has not had the same impact within experimental economics as the oeuvre of Smith, who went on to win a Nobel Prize for his contributions in this area.

20 "My use of randomness was often to protect me from myself or maybe a grad student [...] to make sure that the results were not a consequence of some subtle experimental procedure." (Charles Plott, personal communication).

6 Social policy

Many attempts have been made to analyze the development of rigorous experimentation in social policy,[21] and some of this work points to randomized evaluations going back well into the first half of the 20th century. Unfortunately, as we saw regarding the field of educational psychology, most such claims turn out to be incorrect (typically involving instead careful but nonrandomized choice of the control group) or simply unverifiable. A perhaps surprising candidate for the position of first RCT in social science comes from the field of political science.

Leading up to the US presidential election of 1924, Harold Gosnell worked on a project whose goal was to increase voting rates in Chicago. The primary intervention was a mailed post-card (sent not just in English but also in Polish, Czech, and Italian) describing the necessity of registration prior to voting, and the results were encouraging. However, there have been conflicting opinions in the scholarly literature as to whether he used randomization to achieve those results. In the full report (Gosnell [59]), he himself writes:

> *The second step in the process of sampling was the division of the citizens in each of the districts canvassed into two groups, one of which was to be experimented upon while the other was not. It was assumed that the non-experimental groups could be used as a sort of control.* [...] *In order to avoid possible contacts between the experimental and the control groups, the dividing lines between the two groups were as sharply drawn as possible.*

This strongly suggests that each of the 12 districts where the study was carried out was divided into two parts, one of which was somehow chosen as treatment and one as control. Forsetlund et al. [36] acknowledge that Gosnell mentioned using "random sampling" as a method to control for non-experimental variables, but they conclude from the description above (and from the lack of any explicit affirmative discussion of how randomization was introduced) that "random allocation is very unlikely to have been used to create the comparison groups." Indeed there is no irrefutable proof, but that conclusion seems overly pessimistic. In particular, they and others may have been unaware of Gosnell's original short report on the project (Gosnell [60]), in which he states:

> *In order to set up this experiment it was necessary to keep constant, within reasonable limits, all the factors that enter into the electoral process except the particular stimuli which were to be tested.* [...] *The method of random sampling was used to control these factors during the testing of the particular stimuli used in the experiment.*

Although the phrase "random sampling" refers in modern parlance to choosing a representative subset of a population, which is as we have seen distinct from randomized allocation, this was not true at the time. There are multiple examples of random sampling being used in the context of randomly choosing between subsets of subjects, including Walters [61]. Indeed it is clear from Gosnell's own description that he was not referring to sampling in the modern sense, since he had no need for that: "Special efforts were made to list all the eligible voters in these areas." The most likely conclusion is that Gosnell did indeed randomize but at the "cluster" level, i. e. in order to determine which of the previously determined halves of the district (which had themselves been matched across treatment and control on baseline demographics and other observables) would receive the intervention and which would not.

Whether Gosnell's study was randomized or not, two things are clear: First is that he did not immediately influence others to randomize, within political science or social experimentation broadly. Second, however, is that like more and more others at that time he clearly understood the need for a rigorous control group in order to isolate causal factors, which is what kept driving scholars toward randomization. The timing is not coincidental here: social policy, educational psychology, clinical medicine, and agriculture all used randomized assignment within a few years of each other (seemingly for the first time in each case, excepting the medical homeopathy trial of 1835) in the mid to late 1920s.

---

21 See for example Logan [52], Boruch et al. [53], Farrington [54], Oakley [55], Greenberg and Shroder [56], Levitt and List [57], and Gueron and Rolston [58]. I followed up original sources listed in these resources and the various papers and books therein cited as far as possible, as well as performing my own database searches for "random[ize]" more broadly.

Turning back to political science, Gosnell himself did not pursue this methodological approach. Eldersveld [62] explicitly randomizes in a similar get-out-the-vote experiment, but it did not really become popular or mainstream in political science until the turn of the 21st century (see Green and Gerber [63]). However, this strand of literature does provide another example of the close interaction between social science fields. The second RCT on this topic was conducted by a social psychologist in Pennsylvania in 1935: Hartmann [64] randomly divided city wards into two treatment arms and a control group. Considerably later, lab experimentalists in formal political theory (e. g. Fiorina and Plott [65]) studied issues such as majority rule, using random assignment across conditions and even within positions on a committee.

A major and fascinating early experiment in industrial psychology took place at the Hawthorne factory of the Western Electric Company, near Chicago. From the mid-1920s to the early 1930s, various environmental factors (such as lighting level) were varied – though apparently not randomly. Early results were interpreted as improvements arising merely from being studied, which is now referred to as the Hawthorne (or observer) effect, although re-analysis of the original data (see List and Rasul [66]) casts doubt on whether that conclusion was accurate for the original experiment.[22]

The first clearly and individually randomized social experiment was the Cambridge-Somerville youth study. This was devised by Richard Clarke Cabot, a physician and pioneer in advancing the field of social work. Running from 1942–45, the study randomized approximately 500 young boys who were at risk for delinquency into either a control group or a treatment group, the latter receiving counseling, medical treatment, and tutoring. Results (Powers and Witmer [67]) were highly disappointing, with no differences reported; this may have been due to substantial attrition. Despite that difficulty, sociology and criminology continued to be early adopters in the use of random experimentation, with studies by Reimer and Warren [68] on parole caseload levels, by Hanson and Marks [69] on interviewer accuracy in the 1950 US Census, and by Ares et al. [70] on the large-scale Manhattan bail project.

In many ways public health acts as medicine on a social scale, and we find a similar trend for attempting to use randomization when possible even in large-scale interventions – starting a couple of decades later. A noteworthy early example involved testing the effectiveness of Jonas Salk's polio vaccine in the early 1950s, when there was a debate about whether to implement comprehensive vaccination. In order to evaluate its efficacy (and its potential risks), given that the disease had a relatively low incidence rate in the US (especially for so-called paralytic polio), a large sample was needed: in this case over a million children. Some local health departments were hesitant to randomize and preferred an approach in which all second-graders would be vaccinated, with the first- and third-graders serving as controls. Other health departments felt that this would not be sufficiently rigorous and therefore not sufficiently compelling, thereby wasting all the money and effort spent, so they preferred a randomized (double-blind) placebo approach. In the end about half of the participants ended up using each method, which illustrates some common difficulties of using RCTs in the field. Results (see Francis et al. [71], but also Meier [72] for a broader perspective) were highly encouraging, and polio vaccination has been standard ever since.[23]

Another impressive early randomized experiment in the realm of public health involved family planning in Taiwan (see Population Council [73] for the experimental design and Takeshita [74] for results). The city of Taichung was divided into three roughly matched sectors, each of which consisted of hundreds of neighborhoods (of 25–30 families). Individual neighborhoods were randomized into either a control treatment; a treatment involving information by mail only; or one of two more exhaustive treatments (which included group meetings and personalized home visits), either with the wife only or with both spouses. The relevance of the sectors is that the percentage randomized into the exhaustive treatments differed across sectors, from 20 % up to 50 %. This allowed the researchers to look at intensity of treatment and to examine what they called "circulation effects" (i. e. spillovers), an extraordinarily sophisticated protocol for the time. A slightly later family planning experiment (Chang et al. [75]), also in Taiwan, randomized ten experimental counties

22 The potential existence of the Hawthorne effect may itself argue against some forms of randomized assignment and in favor of less formally rigorous but less disruptive techniques.

23 Modern vaccines, however, use live virus as opposed to Salk's killed virus.

in which field workers received a monetary bonus for every woman who accepted birth control,[24] versus ten control counties. Testing marginal financial incentives for health or similar workers has returned as an active (and supposedly cutting-edge) area of research.

Although psychologists had been doing randomized applied work since the 1920s in studies of learning, and in the laboratory even longer, they tended to do less directly policy-relevant research. However, Campbell [76] gives an overview of social experimentation from a psychological perspective, with a hierarchy that lists "true" experiments involving a randomized control group at its top. Bridging the lab and the field, Deci [77] studied whether external rewards (pecuniary or otherwise) 'crowd out' intrinsic motivation. The lab studies were randomized, while the field study used two pre-existing groups as treatment and control. Later, of course, marketing research (the home of so-called 'A/B testing') applied psychological principles to advertising, consumer interfaces, and more.

Meanwhile Heather Ross, an MIT graduate student at the time, initiated in the 1960s what is considered the first field experiment in economics – after but not long after the corresponding research in public health, psychology, and sociology. She proposed to study the effects of a negative income tax (i. e. phased income supplementation by the government for very low incomes) in what became the New Jersey Income Maintenance Experiment. The experiment randomized, at the household level, both the level of guaranteed minimum income and the (negative) tax rate. Ross [78] finds little evidence of a concomitant reduction in labor supply, although later analysis of the data suggested that it does in fact exist.

This project was followed by several other major randomized social experiments in economics, for instance Brook et al. [79] comparing outcomes of free as opposed to merely low cost health care in the RAND Health Insurance Experiment. Around the same time as Ross, Peter Bohm conducted a field experiment to test theoretical principles rather than direct policy questions. Bohm [80] reports the results of an experiment involving willingness-to-pay for a Swedish closed-circuit television program, in which subjects were randomized into one of six possible treatment groups. The link to policy was made particularly effectively over the succeeding decades in the area of welfare, e. g. in the case of the Supported Work program evaluation (Hollister, Kemper, and Maynard [81]), as described fully in Gueron and Rolston [58].

One arena in which economists were perhaps at the forefront of randomization, and which continues to be one of the most fruitful areas of application, is in the field of international development. The Radio Mathematics Project in Nicaragua began in 1974 as an effort to study the efficacy of teaching math skills via radio, initiated by education economists. The first publication from this study, comparing test scores and finding generally positive effects, was Searle et al. [48]. A later paper vividly highlighted the importance of randomized evaluation, this time in the context of students repeating school years, by showing that the full (rigorous) results ran contrary to early results reported using only the original pilot data which had not been randomized; see Jamison [82].

7 Conclusion

This paper has argued that a single notion of randomized assignment captures not only the usual application of random allocation into treatment and control groups, but also more broadly any randomization that controls for observable and unobservable factors. This allows for the legitimate direct comparison of empirical observations across conditions in a broad range of environments, and hence for ascriptions of causality. Such an approach appears to be novel in the literature on this topic, and it allows a more holistic vision of the development of the concept over time. In particular we expand beyond a focus on clinical medicine, or indeed any single discipline (e. g. limited previous work on experiment economics, psychology, and social policy), and look at some of the interconnections between them. Several specific examples that have not been highlighted in the existing literature are also resurrected en route.

24 $0.50 if pills or condoms; $2.50 if loop (IUD).

In what can be called the first phase of the introduction of randomization to empirical social science, a scattering of 19th-century research studies consciously employed the technique to good effect. Notable examples include the 1835 Nuremberg salt trial and the 1884 Peirce psychophysics experiment. Although these were completely rigorous even by modern standards, they did not immediately spawn imitators or enjoy influence. One possible reason is that they were simply unknown and unlucky, but there are two more plausible explanations. A purely practical possibility is that these experiments did not formally involve the *allocation* of subjects into groups, but rather randomization across treatments. It may not originally have occurred to practitioners that randomization could just as easily be used for allocation, which was the most typical need. But the most likely explanation is that the main goal was to provide a valid comparison, and no particular distinction was made at the time between randomization and other methods for doing so, such as matching and alternation. We see support this in the work of Fibiger [23], who equates alternation with randomization, and in the multiple educational psychology studies of the early 20th century.

What had been a purely practical concern for learning (and debating one's colleagues, as in homeopathy versus allopathy) became a more conceptual or theoretical concern in the 20th century. Issues surrounding causality and epistemology raised the bar for social and clinical science. It became increasingly necessary to be able to claim that certain discrete factors led to certain outcomes with a high degree of confidence in that knowledge, in order to convince one's peers as well as policy-makers. Although these issues were explicitly articulated in the late 1940s and 1950s, the more explanatory transition period that gave rise to them displays a remarkable convergence across fields. In particular, one of the main contributions of the paper is the argument that four related but distinct disciplines experienced the introduction of randomized assignment within just a few years of one another around the late 1920s: political science (which was still closely tied to economics and political philosophy at the time); agricultural research; educational psychology; and of course medicine. This constituted the second phase of randomized assignment entering the social and related sciences.

The third and final phase, from the 1950s to the 1970s, was the application of randomization to larger-scale and policy-oriented problems. We find occurrences in public health as a natural analogue of clinical medicine, as well as in criminology and sociology. Contemporaneously, lab psychology moved to the field with randomized applications in marketing science and industrial psychology. Psychologists also started collaborating with economists to undertake lab studies that were motivated by and informed 'real-world' topics such as bargaining, consumer demand, and market efficiency. Eventually, but still within this formative period, economists moved to the field themselves: studying tax systems and economic development. Although there has been some pushback,[25] RCTs have continued to be increasingly popular in social science field research and policy evaluation.

Of course, there are many environments where randomized assignment is simply infeasible – imagine nation-wide health systems – although even in such settings creative experimental designs can begin to nibble around the edges. Even when feasible, in addition to the Hawthorne effect mentioned earlier, there are occasions when randomization *per se* can make research more difficult. Subjects may be unwilling to be 'experimented upon', and in fact Kramer and Shapiro [86] claim that it is much harder to recruit subjects for randomized than for nonrandomized drug trials. By definition this potential effect is difficult to test, although one can try to compare characteristics of subjects who respond to varied recruitment approaches.[26] Sometimes purely qualitative work will be less disruptive and allow greater fidelity to subjects' intrinsic behaviors; other times quantitative randomized evaluations in the field will leave subjects entirely noncognizant that there is even an experiment taking place. What remains clear is that over time the intellectual community has assigned value to specific attributes that obtain when randomization is employed. The particular attributes – blinding, equivalent comparisons, causal attributions, practicality and expediency, fairness, rigor as a social

25 Within economics for instance Harrison and List [83], Deaton [84], List and Rasul [66], and Harrison [85] have argued against the primacy of randomized assignment, especially concerning the apparent desire of some RCT proponents to occupy, by act or definition, the entire space of rigorous research.

26 See Harrison et al. [87] and Gazzale et al. [88] in the context of laboratory experiments.

construct, and more – have varied over time and across subfields of social science. But whenever value is assigned, researchers have stepped in to fill the gap and will continue to do so.

References

1. Pearl J. Causality. Cambridge: Cambridge University Press; 2000.
2. Smith GCS, Pell JP. Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. Br Med J. 2003;327:1459–61.
3. Chalmers I. Comparing like with like: some historical milestones in the evolution of methods to create unbiased comparison groups in therapeutic experiments. Int J Epidemiol. 2001;30:1156–64.
4. Fienberg SE, Tanur JM. Experimental and sampling structures: parallels diverging and meeting. Int Stat Rev. 1987;55(1):75–96.
5. Mahalanobis PC. Recent experiments in statistical sampling in the Indian Statistical Institute. J R Stat Soc. 1946;109:325–78.
6. Urbach P. Randomization and the design of experiments. Philos Sci. 1985;52:256–73.
7. Papineau D. The virtues of randomization. Br J Philos Sci. 1994;45(2):437–50.
8. Silverman WA, Chalmers I. Casting and drawing lots: a time honoured way of dealing with uncertainty and ensuring fairness. Br Med J. 2001;323:1467–8.
9. van Helmont JB. Ortus Medicinæ. Id Est, Initia Physicæ Inaudita. Amsterdam: Elsevier; 1648.
10. Fisher RA. The design of experiments. London: Oliver and Boyd; 1935.
11. Eden T, Fisher RA. Studies in crop variation, IV. The experimental determination of the value of top dressings with cereals. J Agric Sci. 1927;17:548–62.
12. Darwin C. On the origin of species by means of natural selection. London: John Murray; 1859.
13. Petrarch J. Letter to Boccaccio (V.3). Rerum Senilium Libri. Liber XIV: Epistola 1. 1364.
14. van Helmont JB. Oriatrike, or physick refined: the common errors therein refuted and the whole art reformed and rectified London: Lodowick-Loyd; 1662.
15. Lind J. A treatise of the scurvy. In three parts. Containing an inquiry into the nature, causes and cure, of that disease. Together with a critical and chronological view of what has been published on the Subject Edinburgh: Kincaid and Donaldson; 1753.
16. Watson W. An account of a series of experiments, instituted with a view of ascertaining the most successful method of inoculating the smallpox. London: J Nourse; 1768.
17. Pasteur L. Compte-rendu Sommaire des Expériences Faites à Pouilly-le-Fort près Melun, sur la Vaccination Charbonneuse. C R Acad Sci. 1881;92:1378–83.
18. Claassen JAHR. The gold standard: not a golden standard. Br Med J. 2005;330:1121.
19. Löhner G. Die Homöopathischen Kochsalzversuche zu Nürnberg. Nuremberg. 1835.
20. Stolberg M. Inventing the randomized double-blind trial. J R Soc Med. 2006;99:642–3.
21. Kaptchuk TJ. Intentional ignorance: a history of blind assessment and placebo controls in medicine. Bull Hist Med. 1998;72(3):389–433.
22. Balfour TG. as quoted in West C. Lectures on the diseases of infancy and childhood. London: Longman Brown Green and Longmans; 1854.
23. Fibiger J. Om Serumbehandling af Difteri. Hospitalstidende. 1898;6:309–25.
24. Colebrook D. Irradiation and health. Med Res Coun Spec Rep. 1929;131:4–13.
25. Doull JA, Hardy M, Clark JH, Herman MB. The effect of irradiation with ultra-violet light on the frequency of attacks of upper respiratory disease. Am J Hyg. 1931;13:460–77.
26. Marks HM. James angus doull and the well-controlled common cold. J R Soc Med. 2008;101(10):117–9.
27. Medical Research Council. Streptomycin treatment of pulmonary tuberculosis: a medical research council investigation. Br Med J. 1948;2:769–82.
28. Medical Research Council. Prevention of whooping-cough by vaccination: a medical research council investigation. Br Med J. 1951;1:1463–71.
29. Fechner G. Elemente der Psychophysik. Leipzig: von Breitkopf & Haertel; 1860.
30. Müller GE. Über die Maassbestimmungen des Ortsinnes der Haut Mittels der Methode der Richtigen un Falschen Fälle. Arch Gesammte Physiologie Menschen Thiere. 1879;19:191–235.
31. Dehue T. Deception, efficiency, and random groups: psychology and the gradual origination of the random group design. Isis. 1997;88(4):653–73.
32. Richet C. La Suggestion Mentale et le Calcul des Probabilités. Rev Philos Fr étrang. 1884;18:609–74.
33. Hacking I. Telepathy: origins of randomization in experimental design. Isis. 1988;79(3):427–51.
34. Stigler SM. A historical view of statistical concepts in psychology and educational research. Am J Educ. 1992;101:60–70.

35. Peirce CS, Jastrow J. On small differences of sensation. Mem Natl Acad Sci 1884. 1885;3:75–83.
36. Forsetlund L, Chalmers I, Bjørndal A. When was random allocation first used to generate comparison groups in experiments to assess the effects of social interventions? Econ Innov New Technol. 2007;16(5):371–84.
37. McCall WA. How to experiment in education. New York: Macmillan; 1923.
38. Dearborn WF, Lincoln EA. A class experiment in learning. J Educ Psychol. 1922;13(6):330–40.
39. Shaffer LF. A learning experiment in the social studies. J Educ Psychol. 1927;18(9):577–91.
40. Clark BE. The effect upon retention of varying lengths of study periods and rest intervals in distributed learning time. J Educ Psychol. 1928;19(8):552–9.
41. Thurstone LL. The indifference function. J Soc Psychol. 1931;2:139–67.
42. Roth AE. On the early history of experimental economics. J Hist Econ Thought. 1993;15:184–209.
43. Davidson D, Siegel S, Suppes P. Some experiments and related theory on the measurement of utility and subjective probability. Office of Naval Research Contract NR 171-034 Technical Report 1. 15 August 1955.
44. Chamberlin EH. An experimental imperfect market. J Polit Econ. 1948;56(2):95–108.
45. Smith VL. An experimental study of competitive market behavior. J Polit Econ. 1962;70(2):111–37.
46. Kalisch G, Milnor JW, Nash J, Nering ED. Some experimental n-person games. RAND Research Memorandum 948, Aug 25 1952.
47. Atkinson RC, Suppes P. An analysis of two-person game situations in terms of statistical learning theory. Office of Naval Research Contract NR 171-034 Technical Report 8. 25 April 1957.
48. Searle B, Matthews P, Suppes P, Friend J. Formal evaluation of the 1976 first-grade instructional program. In: Suppes P, Searle B, Friend J, editors. The radio mathematics project: Nicaragua 1976–77. Stanford CA: Institute for Mathematical Studies in the Social Sciences; 1978. p. 97–124.
49. Siegel S, Fouraker LE. Bargaining and group decision-making: experiments in bilateral monopoly. New York: McGraw-Hill; 1960.
50. Suppes P, Carlsmith JM. Experimental analysis of a duopoly situation from the standpoint of mathematical learning theory. Int Econ Rev. 1962;3(1):60–78.
51. Smith VL. Discovery – a memoir Bloomington: AuthorHouse; 2008.
52. Logan CH. Evaluation research in crime and delinquency: a reappraisal. J Crim Law Criminol. 1973;63(3):378–87.
53. Boruch RF, McSweeny AJ, Soderstrom EJ. Randomized field experiments for program planning, development, and evaluation. Eval Q. 1978;2(4):655–95.
54. Farrington DP. Randomized experiments on crime and justice. Crime Justice. 1983;4:257–308.
55. Oakley A. A historical perspective on the use of randomized trials in social science settings. Crim Delinq. 2000;46(3):315–29.
56. Greenberg D, Shroder M. The digest of social experiments. 3rd ed. Washington DC: Urban Institute Press; 2004.
57. Levitt SD, List JA. Field experiments in economics: the past, the present, and the future. Eur Econ Rev. 2009;53:1–18.
58. Gueron JM, Rolston H. Fighting for reliable evidence. New York: Russell Sage Foundation; 2013.
59. Gosnell HF. Getting out the vote: an experiment in the stimulation of voting. Chicago: University of Chicago Press; 1927.
60. Gosnell HF. An experiment in the stimulation of voting. Am Polit Sci Rev. 1926;20(4):869–74.
61. Walters JE. Seniors as counselors. J High Educ. 1931;2:446–8.
62. Eldersveld SJ. Experimental propaganda techniques and voting behavior. Am Polit Sci Rev. 1956;50:154–65.
63. Green DP, Gerber AS. The underprovision of experiments in political science. Ann Am Acad Polit Soc Sci. 2003;589:94–112.
64. Hartmann GW. A field experiment on the comparative effectiveness of 'emotional' and 'rational' political leaflets in determining election results. J Abnorm Soc Psychol. 1936;31(1):99–114.
65. Fiorina MP, Plott CR. Committee decisions under majority rule: an experimental study. Am Polit Sci Rev. 1978;72(2):575–98.
66. List JA, Rasul I. Field experiments in labor economics. In: Ashenfelter O, Card D, editors. Handbook of labor economics Vol 4a. Amsterdam: North Holland; 2011. p. 103–228.
67. Powers E, Witmer H. An experiment in the prevention of juvenile delinquency: the Cambridge-Somerville Youth Study. New York: Columbia University Press; 1951.
68. Reimer E, Warren M. Special intensive parole unit. Natl Probat Parole Assoc J. 1957;3:222–9.
69. Hanson EH, Marks ES. Influence of the interviewer on the accuracy of survey results. J Am Stat Assoc. 1958;53:283. 635–55.
70. Ares CE, Rankin A, Sturz H. The Manhattan Bail Project: an interim report on the use of pre-trial parole. NY Univ Law Rev. 1963;38:67–95.
71. Francis T Jr, et al.. An evaluation of the 1954 poliomyelitis vaccine trials. Am J Publ Health. 1955;45(5(pt2):1–63.
72. Meier P. The biggest public health experiment ever: the 1954 field trial of the salk poliomyelitis vaccine. In: Statistics: a guide to the unknown San Francisco: Holden-Day; 1972. p. 2–13.
73. Council P. The taichung program of pre-pregnancy health. Stud Fam Plann. 1963;1(1):10–2.
74. Takeshita J. The taichung program of pre-pregnancy health. Stud Fam Plann. 1964;1(4):10–2.
75. Chang MC, Cernada GP, Sun TH. A field-worker incentive experimental study. Stud Fam Plann. 1972;3(11):270–2.
76. Campbell DT. Reforms as experiments. Am Psychol. 1969;24:409–29.
77. Deci EL. Effects of externally mediated rewards on intrinsic motivation. J Pers Soc Psychol. 1971;18(1):105–15.
78. Ross H. An experimental study of the negative income tax. Child Welf. 1970;49(10):562–9.

79. Brook RH, et al.. Does free care improve adults' health? – Results from a randomized controlled trial. N Engl J Med. 1983;309:1426–34.
80. Bohm P. Estimating demand for public goods: an experiment. Eur Econ Rev. 1972;3:111–30.
81. Hollister GH, Kemper P, Maynard RA. The national supported work demonstration. Madison, WI: Univ of Wisconsin Press; 1984.
82. Jamison DT. Radio education and student failure in Nicaragua: a further note. In: Friend J, Searle B, Suppes P, editors. Radio mathematics in Nicaragua. Institute for Mathematical Studies in the Social Sciences: Stanford CA; 1980. p. 225–36.
83. Harrison GW, List JA. Field experiments. J Econ Lit. 2004;42:1009–55.
84. Deaton A. Instruments, randomization, and learning about development. J Econ Lit. 2010;48:424–55.
85. Harrison GW. Field experiments and methodological intolerance. J Econ Methodol. 2013;20(2):103–17.
86. Kramer M, Shapiro S. Scientific challenges in the application of randomized trials. J Am Med Assoc. 1984;252(19):2739–45.
87. Harrison GW, Lau MI, Rutström. Rick attitudes, randomization to treatment, and self-selection into experiments. J Econ Behav Organ. 2009;70:498–507.
88. Gazzale R, Jamison JC, Karlan A, Karlan D. Ambiguous solicitation: ambiguous prescription. Econ Inq. 2013;51(1):1002–11.