University of Groningen

Deception, Efficiency, and Random Groups

Dehue, Trudy

Published in:
Isis

DOI:
10.1086/383850

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
1997

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
Dehue, T. (1997). Deception, Efficiency, and Random Groups: Psychology and the Gradual Origination of the Random Group Design. *Isis*, *88*(4), 653-673. https://doi.org/10.1086/383850

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 26-10-2026

History of Science Society

THE UNIVERSITY OF CHICAGO PRESS JOURNALS

Deception, Efficiency, and Random Groups: Psychology and the Gradual Origination of the Random Group Design

Author(s): Trudy Dehue

Source: *Isis*, Vol. 88, No. 4 (Dec., 1997), pp. 653-673
Published by: The University of Chicago Press on behalf of The History of Science Society
Stable URL: https://www.jstor.org/stable/237831
Accessed: 10-12-2018 14:54 UTC

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at https://about.jstor.org/terms

JSTOR

The History of Science Society, The University of Chicago Press are collaborating with JSTOR to digitize, preserve and extend access to *Isis*

This content downloaded from 129.125.148.19 on Mon, 10 Dec 2018 14:54:55 UTC
All use subject to https://about.jstor.org/terms

Deception, Efficiency, and Random Groups

Psychology and the Gradual Origination of the Random Group Design

*By Trudy Dehue**

ABSTRACT

In the life sciences, psychology, and large parts of the other social sciences, the ideal experiment is a comparative experiment with randomly composed experimental and control groups. Historians and practitioners of these sciences generally attribute the invention of this "random group design" to the statistician R. A. Fisher, who developed it in the 1930s for agricultural research. This essay argues that the random group design was advanced in psychology before Fisher introduced it in agriculture and that in this context it was the unplanned outcome of a lengthy historical process rather than the instantaneous creation of a single genius. The article analyzes how the random group design came about bit by bit when methodological practices from nineteenth-century psychophysical laboratories were gradually adapted, extended, and codified by twentieth-century educational psychologists to support procedural objectivity in educational administration. In passing, the article also amends the received historiography of the separate elements of randomization and control groups.

TO CONTEMPORARY STUDENTS OF THE LIFE SCIENCES, psychology, and large parts of the other social sciences, comparative experimentation with randomly composed control and experimental groups is the standard against which all other research designs are measured. For studies without a control group, or without chance allocation of subjects to groups, the authoritative methodologist Donald T. Campbell introduced the

* Theory and History of Psychology, Heymans Institute, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, Netherlands.

An early version of this essay was discussed at the Department of Science and Technology Studies, University of Amsterdam, and at the Section of Theory and History of Psychology, University of Groningen. I thank all colleagues in both groups, and particularly Douwe Draaisma and Peter van Drunen, for their very useful remarks. Special thanks also go to Kurt Danziger, Theodore Porter, Margaret Rossiter, and the anonymous referees for *Isis,* who were very helpful. Furthermore, acknowledgment is due to Lemma Publishers for permission to use parts of an article in Dutch: Trudy Dehue, "Stanford, Columbia en het Ideale Experiment," *Psychologie en Maatschappij,* 1995, *19:*113–127.

Isis, 1997, 88:653–673
© 1997 by The History of Science Society. All rights reserved.
0021-1753/97/8804-0003$02.00

This content downloaded from 129.125.148.19 on Mon, 10 Dec 2018 14:54:55 UTC
All use subject to https://about.jstor.org/terms

qualification *quasi experiments.*[1] Methodologically trained reviewers of such experiments habitually seek alternative variables to explain the reported results.

The random group experiment achieved the status of an exemplar largely through Ronald A. Fisher's 1935 methodological handbook *The Design of Experiments.* In the 1920s and 1930s Fisher, a statistician and biometrician, was employed in an agricultural research station at Rothamstead in England. There he designed research strategies for discovering the causes of observed crop variations or studying the effects of particular treatments. He proposed that experimental groups of plants or plots of land should be subjected to a treatment (say, supply of nitrogen) and subsequently compared to an untreated control group. In order to balance out other—possibly contaminating—differences between the groups (e.g., humidity of the soil, quantity of light), Fisher argued, the experimental and control groups should be randomly composed.[2]

Although working researchers gradually came to regard comparing randomly composed groups as the ideal experiment, the design did not remain undisputed among specialists in statistics. The value of randomization, in particular, was already challenged in the 1930s. Opponents of Fisher, like W. S. Gosset, alias "Student," contended that when causes of variation or effects of treatments are studied, there is more profit in considering beforehand what factors might contaminate the research results. The items to be investigated should be paired or "matched," as it is still called today. Pairs equal on each possible contaminating factor are formed; then one item from each pair is assigned to the experimental group and the other to the control group. Thus every relevant factor apart from the condition under experimental consideration is presumed to occur equally in both groups.[3] Proponents of matching held that this strategy of equalizing the groups is both more reliable and more informative than that of randomization. Fisher, however, stressed that it is impossible to make groups comparable by matching: researchers, if pressed, can usually think up many more contaminating factors than those for which they can create equal pairs; moreover, more factors than they can possibly imagine may play a part in even the simplest experiment—and any one of those factors might be responsible for an experimental effect. Fisher believed that careful randomization offers a far better guarantee that contaminating differences between groups will be excluded.

I will not venture into such specialists' debates on the methodological merits of ran-

[1] Donald T. Campbell and Julian C. Stanley, *Experimental and Quasi-experimental Designs for Research* (Chicago: Rand McNally, 1966). Many comparable expressions have also been in circulation. As early as 1945, the sociologist Ernest Greenwood listed adjectives like *substitute, semi, tentative, partial, uncontrolled,* and *indirect* for experiments without control groups; see Ernest Greenwood, *Experimental Sociology: A Study in Method* (New York: King's Crown, 1945), pp. 11–12, 14.

[2] R. A. Fisher, *The Design of Experiments* (Edinburgh/London: Oliver & Boyd, 1935). Fisher was employed at the Rothamstead agricultural research station from 1919 to 1933, but his interest in statistics was inspired mainly by his earlier involvement with eugenics and biometrics. In 1933 he succeeded Karl Pearson as the Galton Professor of Eugenics at the University of London. On Fisher see Donald A. MacKenzie, *Statistics in Britain: The Social Construction of Scientific Knowledge, 1865–1930* (Edinburgh: Edinburgh Univ. Press, 1981); and Stephen M. Stigler, *The History of Statistics* (Cambridge, Mass.: Harvard Univ. Press, 1986).

[3] An early and very clear presentation of the idea of matching was given by Francis Galton in 1872. Doubting the positive effects of piety, Galton proposed to settle the issue empirically. He explained that "the principles are broad and simple. We must gather cases for statistical comparison, in which the same object is keenly pursued by two classes similar in their physical but opposite in their spiritual state; the one class being spiritual, the other materialistic. Prudent pious people must be compared with prudent materialistic people and not with the imprudent nor the vicious. . . . We simply look for the final result—whether those who pray attain their objects more frequently than those who do not pray, but who live in all other respects under similar conditions": Francis Galton, "The Efficacy of Prayer," *Fortnightly Review,* 1872, *12:*124–135, on p. 126. An abridged version of this research proposal was printed in Galton, *Inquiry into Human Faculty and Its Development* (New York: Macmillan, 1883), pp. 277–294, on p. 278.

This content downloaded from 129.125.148.19 on Mon, 10 Dec 2018 14:54:55 UTC
All use subject to https://about.jstor.org/terms

domization.[4] This essay discusses the historical background of the random group design. Many credit Fisher not just with its dissemination, but also with its invention.[5] I argue, however, that the random group design was advanced in psychology before Fisher introduced it in agriculture and that in this context it was the unplanned outcome of a lengthy historical process rather than the instantaneous creation of a single genius. The essay considers how the random group design came about bit by bit, as methodological practices from nineteenth-century psychophysical laboratories were gradually adapted, extended, and codified by twentieth-century educational psychologists to support procedural objectivity in educational administration. In passing, it also amends the received historiography of the separate elements of randomization and control groups.

RANDOMIZATION IN PSYCHOPHYSICS

Between 1845 and 1849, the German scholar Gustav Fechner lifted his arms with a weight in each hand over sixty-seven thousand times. After each lifting, Fechner meticulously wrote down whether he had been able to tell the heavier weight from the lighter one. Thus he tried to establish the relationship between the weights as measured and the weights as perceived—that is, between physical stimuli and mental sensations. Using the "difference threshold" or "just noticeable difference" as an accounting unit, Fechner arrived at a formula expressing a logarithmic relation of increase in stimulus and increase in sensation.[6]

In Fechner's time, a claim to measure anything as subjective as human sensation was an impressive novelty. Soon the "psychophysical research" he initiated became a craze. Innumerable psychophysical experiments were conducted in biology, in physiology, and, particularly, in the newly established psychological laboratories at European and American universities. Countless weights were lifted, and Fechner's so-called method of right and wrong cases was employed for testing other senses as well. A popular variation was touching the skin with two pins at a particular distance and trying to establish the relationship between the actual distance and the distance as perceived. The differential sensitivities of foreheads and chins, shoulders and fingers, hips, bellies, legs, knees, and ankles were extensively discussed. Experimenters would complain about sensitive and scaly skins and would apologetically argue that 5,710 pricks should suffice to prove a point.[7]

[4] See E. S. Pearson on the issue: "Few practicing statisticians have the ability and experience of a Gosset, Fisher, or Yates, and therefore to safeguard against blunders it is no doubt important to teach would-be applied statisticians the value of randomization where it is possible in design of experiments. But that does not prove that Fisher was 'right' and Gosset 'wrong.' " E. S. Pearson, *"Student": A Statistical Biography of William Sealy Gosset,* ed. R. L. Plackett and G. A. Barnard (Oxford: Clarendon, 1990), p. 109. On the debate see Ian Hacking, "Telepathy: Origins of Randomization in Experimental Design," *Isis,* 1988, *7:*427–451, on pp. 428–429.

[5] A sample of books ascribing the design to Fisher: David S. Moore, *Statistics: Concepts and Controversies* (San Francisco: Freeman, 1985), p. 62; Ann B. Blalock and Hubert M. Blalock, eds., *Introduction to Social Research* (1970; rpt., Englewood Cliffs, N.J.: Prentice-Hall, 1982), pp. 65–66; John Neter, William Wasserman, and Michael H. Kutner, *Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs* (Homewood, Ill.: Irwin, 1985), p. 899; P. W. Holland, "Statistics and Causal Inference," *Journal of the American Statistical Association,* 1986, *81:*945–970, on p. 953; Roger Mead, *The Design of Experiments: Statistical Principles for Practical Applications* (Cambridge: Cambridge Univ. Press, 1988), pp. 4–5; and S. E. Maxwell and H. D. Delaney, *Designing Experiments and Analyzing Data: A Model Comparison Perspective* (Belmont, Calif.: Wadsworth, 1990), pp. 38–40. I am indebted to Tom Snijders for providing most of these references.

[6] Gustav Fechner, *Elemente der Psychophysik* (Leipzig: Von Breitkopf & Haertel, 1860).

[7] W. Camerer, "Versuche über den Raumsinn der Haut nach der Methode der richtigen und falschen Fälle," *Zeitschrift für Biologie,* N.S., 1883, *1:*280–300, on p. 281; and A. Paulus, "Versuche über den Raumsinn der Haut der unteren Extremität," *ibid.,* O.S., 1871, *7:*240–241. For a history of these experiments see Edwin G.

This content downloaded from 129.125.148.19 on Mon, 10 Dec 2018 14:54:55 UTC
All use subject to https://about.jstor.org/terms

Possible sources of experimental bias were also considered. Thus, in the early 1870s, researchers at the Physiological Institute in Tübingen reasoned that knowing the pins' distance in previous trials might give rise to particular expectations about the next trials and that such expectations might bias the results. Some psychophysical researchers held that skilled subjects were able to avoid forming expectations, but the Tübingen experimenters decided not to rely on this ability. They concluded that it takes two to experiment: stimulations were to be administered by an assistant (*Experimentator*) to an observer (*Beobachter*) kept ignorant as to the actual distances of the pins. The Tübingen group also reasoned that the observer's expectations should be thwarted by secretly inserting control or "deception" trials (*Vexirversuche*) in which only one pin was used. And, most significant for the purposes of this essay, they decided to administer the trials with different distances—including the control or deception trials with only one pin (i.e., zero distance)—in irregular order (*in buntem Wechsel*).[8] (See Figure 1.)

In 1879 the psychophysicist Georg E. Müller, at the time *Privatdozent* at Göttingen, published a well-received methodological treatise entitled "Über die Maassbestimmungen des Ortsinnes der Haut mittels der Methode der richtigen un falschen Fälle [On establishing the space-sensibility of the skin employing the method of right and wrong cases]." Whereas the Tübingen researchers had mentioned their method of irregular order only in passing, Müller, expressly praising them for the idea, emphasized that stimuli should always be administered *in buntem Wechsel.* In several other writings Müller and his students elaborated upon the methodological condition of irregularity.[9]

With the separation of the roles of the experimenter and the observer, Fechner's original one-man job was distributed between two parties. The experimenter, who used to pick the stimuli, administer them, and observe the sensations, had been discharged from the third

Boring, *Sensation and Perception in the History of Experimental Psychology* (New York: Appleton-Century-Crofts, 1942), pp. 479–487. On the novelty of claims to measure sensation see Gail A. Hornstein, "Quantifying Psychological Phenomena," in *The Rise of Experimentation in American Psychology,* ed. Jill G. Morawski (New Haven, Conn.: Yale Univ. Press, 1988), pp. 1–34; and Kurt Danziger, *Constructing the Subject: Historical Origins of Psychological Research* (Cambridge: Cambridge Univ. Press, 1990).

[8] R. Kottenkampf and H. Ulrich, "Versuche über den Raumsinn der Haut der obern Extremität," *Z. Biol.,* O.S., 1870, *6:*37–53, on p. 40; Paulus, "Versuche über den Raumsinn," p. 240; and Adolf Riecker, "Versuche über den Raumsinn der Haut des Unterschenkels," *Z. Biol.,* O.S., 1873, *9:*95–104, on p. 96. Deception trials soon became part of the standard methodology of psychophysics. They were discussed in, among other works, Wilhelm Wundt, *Grundzüge der physiologischen Psychologie* (Leipzig: Wilhelm Engelmann, 1874), p. 472; Camerer, "Versuche über den Raumsinn"; Gustav Fechner, *Ueber die Methode der richtigen und falschen Fälle in Anwendung auf die Maassbestimmungen der Feinheit oder extensiven Empfindlichkeit des Raumsinnes* (Leipzig: Hirzel, 1884), pp. 129–141; Victor Henri and Guy A. Tawney, "Ueber die Trugwahrnehmung zweier Punkte bei der Berührung eines Punktes der Haut," *Philosophische Studien,* 1895, *11:*394–406; and Henri, *Die Raumwahrnehmungen des Tastsinnes: Ein Beitrag zur experimentellen Psychologie* (Berlin: Reuther & Reichard, 1898).

[9] Georg E. Müller, "Über die Maassbestimmungen des Ortsinnes der Haut mittels der Methode der richtigen und falschen Fälle," *Archiv für die Gesammte Physiologie des Menschen und der Thiere,* 1879, *19:*191–235, on p. 221. There was a controversy with Wilhelm Wundt on the topic of randomization. Wundt held that expectation and habituation are precisely the phenomena to be investigated and that therefore serial stimulation is to be preferred. Müller, however, took the position that the expectations were errors that should be ruled out. On this debate see Edward B. Titchener, *Experimental Psychology: A Manual of Laboratory Practice,* Vol. 2: *Quantitative Experiments,* Pt. 2: *Instructor's Manual* (New York: Macmillan, 1905), pp. 99–102. A clear explanation of the importance of randomization is given in Lillie J. Martin and Müller, *Zur Analyse der Unterschiedsempfindlichkeit* (Leipzig: Ambrosius Barth, 1899), pp. 175–178; see also Müller and Friedrich K. Schumann, "Experimentelle Beiträge zur Untersuchung des Gedächtnisses," *Zeitschrift für Psychologie und Physiologie der Sinnesorgane,* 1894, *6:*81–190, 257–339. On Müller's methods see Edward J. Haupt, "G. E. Müller as a Source of American Psychology," in *Psychology: Theoretical and Historical Perspectives,* ed. Robert W. Rieber and Kurt Salzinger (Washington, D.C.: American Psychological Association, in press).

This content downloaded from 129.125.148.19 on Mon, 10 Dec 2018 14:54:55 UTC
All use subject to https://about.jstor.org/terms

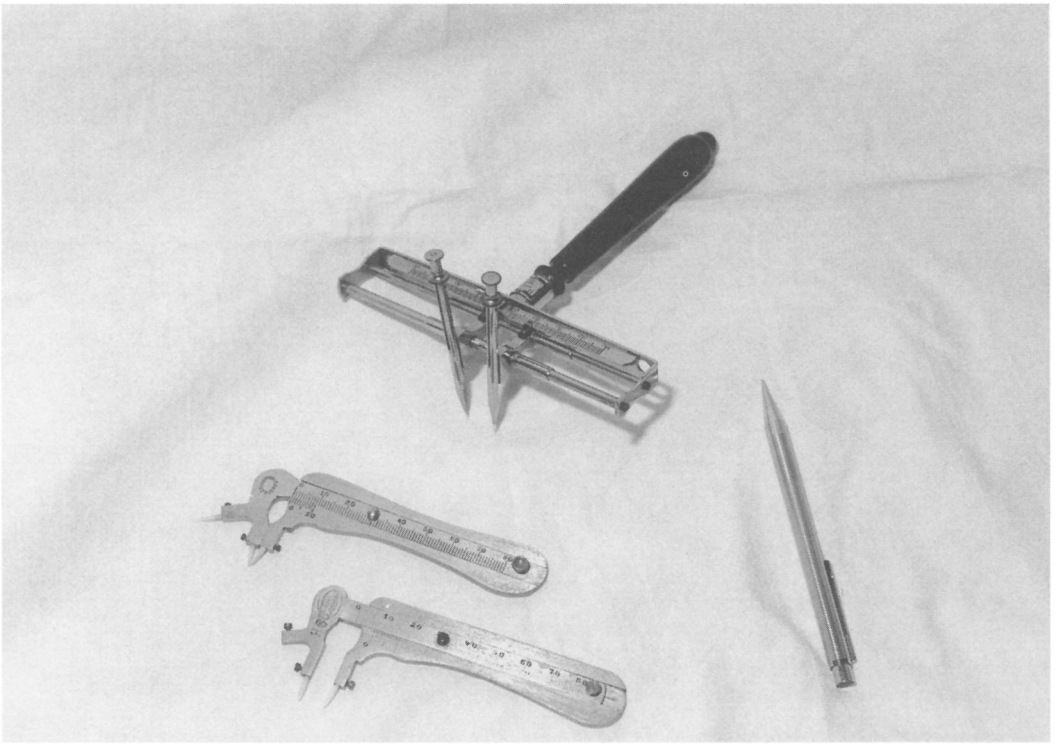

Figure 1. *The Tübingen experimenters in the 1870s used pins fixed on wooden blocks, but by the turn of the century standardized "aestesiometers" were available from the Leipzig firm E. Zimmerman. The instrument with three pins was specially designed for experiments with "deception trials" in which the "observer" was secretly touched at only one point every now and then. (Courtesy Archives of Dutch Psychology, Groningen.)*

task. (See Figure 2.)[10] In an 1884 presentation, "On Small Differences of Sensation," the American scholar Charles Sanders Peirce introduced a further division of functions by also relieving the experimenter of the job of picking the stimuli. Whereas the German researchers did not describe any ways of establishing irregularity and apparently relied on the experimenter's independent ability to achieve a truly arbitrary selection, in the experiments conducted by Peirce and his student Joseph Jastrow the sequence of the trials was carefully randomized with the help of two packs of playing cards. In "A Critique of Psychophysic Methods," published in 1887, Jastrow suggested that one could also throw dice.[11]

In 1903 Müller, who had become the distinguished head of the psychophysical labo-

[10] For an instructive discussion of the social implications of this change see Danziger, *Constructing the Subject* (cit. n. 7), pp. 27–32. Children were sometimes used as observers because they were believed to have a stronger tactile sense than adults. See, e.g., W. Camerer, "Versuche über den Raumsinn er Haut bei Kindern angestellt an der obern Extremität nach der Methode der richtigen und falschen Fälle," *Z. Biol.,* 1881, *17:*1–22; and F. B. Dresslar, "Studies in the Psychology of Touch," *American Journal of Psychology,* 1894, *6:*313–368.

[11] Charles Sanders Peirce and Joseph Jastrow, "On Small Differences of Sensation," in *Memoirs of the National Academy of Sciences, 1884, Vol. III, Part I* (Washington, D.C.: Government Printing Office, 1885), pp. 75–83; and Jastrow, "A Critique of Psycho-physic Methods," *Amer. J. Psychol.,* 1887, *1:*271–309, on p. 289. The historian of statistics Stephen Stigler praised Peirce and Jastrow's 1884 experiment as the first known experiment with a precise, mathematically sound randomization scheme: Stephen M. Stigler, "Mathematical Statistics in the Early States," *Annals of Statistics,* 1978, *6:*239–265, on p. 248.

This content downloaded from 129.125.148.19 on Mon, 10 Dec 2018 14:54:55 UTC
All use subject to https://about.jstor.org/terms

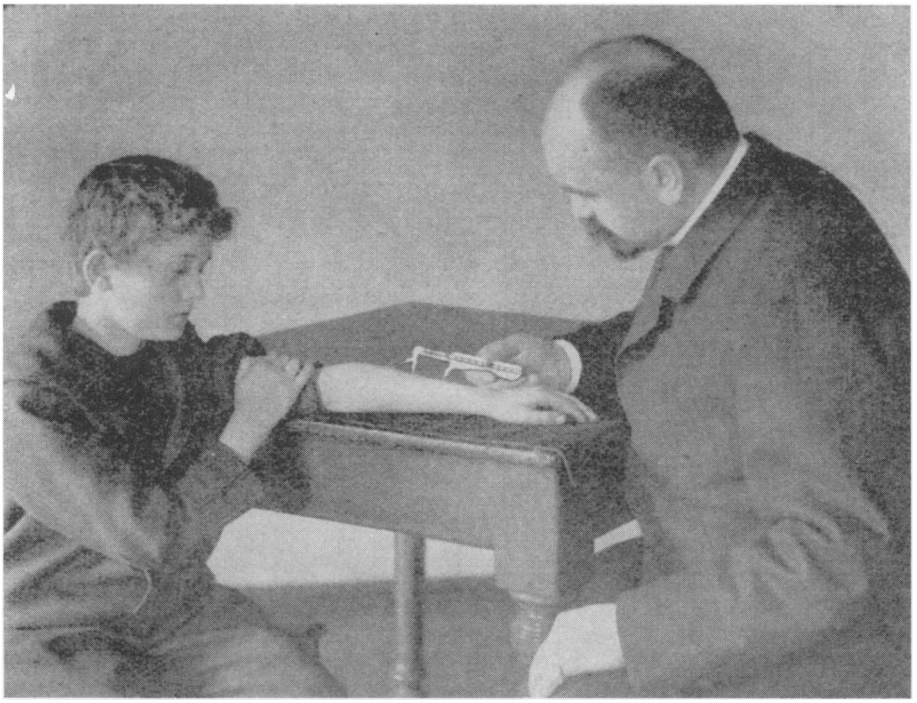

Figure 2. *Pinprick experiments with an experimenter, an observer, and one-pin deception trials. (From R. Schulze,* Aus der Werkstatt der experimentellen Psychologie und Pädagogik *[1909; rpt., Leipzig: Vogtländer, 1913], p. 68.)*

ratory at Göttingen, also recommended this division of the experimental tasks. In order to make sure that stimuli really are administered "*auf ganz zufälliger Weise*" (in a completely accidental way), Müller suggested, the researchers should write all the stimulus values—including those for the controlling deception trials—on pieces of paper and draw one of them before each trial. Müller's procedure was adopted by Edward B. Titchener, of Cornell, for his 1905 student's manual for psychophysical research.[12]

RANDOMIZATION IN PSYCHICAL RESEARCH

Randomization of stimulus order can also be found in early parapsychological, or psychical, research. In the last decades of the nineteenth century, several prominent psychophysicists engaged in psychical experimentation, mostly in hope of unmasking claims of telepathy and the like or of proving physical explanations for alleged psychical phenomena.

[12] Georg E. Müller, "Die Gesichtspunkte und die Tatsachen der psychophysischen Methodik," *Ergebnisse der Physiologie,* 1903, 2:267–517, on pp. 297–299; and Edward B. Titchener, *Experimental Psychology: A Manual of Laboratory Practice,* Vol. 2: *Quantitative Experiments,* Pt. 1: *Student's Manual* (New York: Macmillan, 1905), p. 11. Boring noted that Titchener's books served as encyclopedias and wrote in the 1950s that "even now, half a century later, it is hard to name a more erudite set of volumes or single book in English, in psychology, by a single author": Edwin G. Boring, *History of Experimental Psychology* (1950; 2nd ed., Englewood-Cliffs, N.J.: Prentice Hall, 1957), p. 413.

This content downloaded from 129.125.148.19 on Mon, 10 Dec 2018 14:54:55 UTC
All use subject to https://about.jstor.org/terms

Ever since, randomization has been a central issue in this kind of research, as well as in debates about the credibility of its positive results.[13]

The origins of randomization in psychical research have been analyzed by Ian Hacking. In his fascinating and authoritative "Telepathy: Origins of Randomization in Experimental Design," Hacking argued that psychical research was the context for the very first "faltering use" of randomization. Hence, he also reasoned, psychical research is the seemingly implausible origin of the present-day "ideal design" with randomly composed experimental and control groups.[14]

Hacking's historical reconstruction began with a meticulous account of Peirce and Jastrow's experiments. He emphasized that Peirce was not a mainstream psychophysicist, for he wanted to refute rather than confirm the common psychophysical thesis that stimulus differences can no longer be perceived once they drop below an alleged "just noticeable difference." Peirce suspected that in fact sensorial accuracy decreases gradually and that very small sensorial impressions are still perceived in a subliminal way. At the beginning of "On Small Differences of Sensation," Peirce and Jastrow classified their experiments as belonging to "physiological psychology," but Hacking drew attention to the end of the text. In the final paragraph the authors suggested that their evidence of subliminal perception may explain "the insight of females as well as certain 'telepathic' phenomena."[15] Apart from disproving the notion of an absolute difference threshold, Hacking concluded, Peirce wanted to unmask pretentious or mysterious claims of female intuition or thought transference.

Hacking also discussed an 1885 publication by the Parisian neurophysiologist Charles Richet, who tested the ability of subjects to improve on chance in guessing the color of playing cards that they had not seen, at least not with their bodily eyes. Richet took meticulous care that the cards were randomly chosen. Hacking's main document, however, was an extensive 1917 research report by John Edgar Coover of the psychology department of Stanford University. In Coover's experiments on thought transference, an "agent" picked a playing card and tried to "transfer" it to a "reagent" who had to tell what card was selected. The cards were randomly drawn; and, most important, misleading control trials were randomly inserted in which the reagent was asked to guess the card though in fact the agent had neither looked at a drawn card nor tried to transfer a card.[16]

[13] A recent discussion in the *Psychological Bulletin* illustrated this extensively: Daryl J. Bem and Charles Honorton, "Does Psi Exist? Replicable Evidence for an Anomalous Process of Information Transfer," *Psychol. Bull.,* 1994, *115:*4–18; Ray Hyman, "Anomaly or Artifact? Comments on Bem and Honorton," *ibid.,* pp. 19–24; and Bem, "Response to Hyman," *ibid.,* pp. 25–27. On psychophysicists' psychical experiments see Seymour M. Mauskopf and Michael R. McVaugh, *The Elusive Science: Origins of Experimental Psychical Research* (Baltimore: Johns Hopkins Univ. Press, 1980); and Deborah J. Coon, "Testing the Limits of Sense and Science: American Experimental Psychologists Combat Spiritualism, 1880–1920," *American Psychologist,* 1992, *47:*143–151.

[14] Hacking, "Telepathy" (cit. n. 4).

[15] The full paragraph reads: "The general fact [of no threshold] has highly important practical bearings, since it gives new reason for believing that we gather what is passing in one another's minds in large measure from sensations so faint that we are not fairly aware of having them, and can give no account of how we reach our conclusions about such matters. The insight of females as well as certain 'telepathic' phenomena may be explained in this way." Peirce and Jastrow, "On Small Differences of Sensation" (cit. n. 11), p. 83; some minor inaccuracies that have crept into Hacking's quotation in "Telepathy," p. 434, are corrected here.

[16] Charles Richet, "La suggestion mentale et le calcul des probabilités," *Revue Philosophique de la France et de l'Étranger,* 1884, *18:*609–674; and John Edgar Coover, *Experiments in Psychical Research at Leland Stanford Junior University* (1917; rpt., New York: Arno, 1975). Hacking also discussed similar experiments published in 1895 by experimenters from Copenhagen who asked their subjects to guess unspoken numbers. These researchers reasoned that the "agent" might have an involuntary preference for particular numbers. If such a "number habit"

This content downloaded from 129.125.148.19 on Mon, 10 Dec 2018 14:54:55 UTC
All use subject to https://about.jstor.org/terms

Hacking convincingly demonstrated, then, that psychical research was an important context for the early rise of experimental randomization. Nonetheless, I would point to psychophysics rather than psychical research as the *primary* context. After all, Peirce and Jastrow themselves described their experiments as contributions to psychophysics; Richet is known to have combined psychical with psychophysical interests; and Hacking's prime witness Coover both received his research training in psychophysics and conducted his psychical experiments in a psychology department specializing primarily in psychophysical research.[17] Moreover, as has already been shown, Coover's way of combining randomization with misleading control trials was quite common in early psychophysics.

In themselves, debates on the exact origins of an idea or practice mostly involve trivial hairsplitting. It is not just to set the historical record straight, then, that I begin my account of the history of the random group design with psychophysical experimentation rather than the closely related field of psychical research. In the psychophysical and psychical experiments described so far, randomization was applied to determine the order of experimental conditions: the same subjects were exposed to both the experimental and the control conditions. In the current ideal design, however, randomization is used to compose experimental and control groups: different subjects (people, animals, plants, plots of land, etc.) are randomly assigned to either the experimental or the control groups. As I will show, from a historical point of view it is crucial not to overlook this difference between random ordering and random grouping. The transition from the first kind of randomization to the second involved much more than a single obvious step (which, as Hacking's analysis implies, must still be attributed solely to Fisher). The combination of randomization and grouping in psychology was an outcome of an intricate historical trajectory. Tracing this trajectory makes visible the vital part played by social administration in the early creation of the random group design.

In order to present this extended story, I must first discuss yet another aspect of nineteenth-century psychophysical experimentation that would become part of the methodological tradition in which the design eventually came about. I will subsequently show how methodological elements from psychophysics were integrated in a set of methodological routines that offered the support of procedural objectivity to early twentieth-century management. Attempts to enhance educational efficiency and the trustworthiness of educational data led, first, to the idea of installing a separate control group for the nonstimulus control condition and, later, to use of the older technique of randomization to ensure the comparability of the control and experimental groups. Finally, I will recount how psychologists disseminated the ideal of the random group design—implemented, from the 1930s, with the help of Fisher's statistical methods—in other social sciences as the best tool for testing

corresponded coincidentally with that of the "reagent," a positive result would in fact be produced by this congruity. Therefore it was decided that the numbers to be transferred should be chosen by chance. See Hacking, "Telepathy," pp. 441–444.

[17] In his later autobiography, Jastrow, who made a career in psychology, again discussed his experiments with Peirce as psychophysical experiments: Joseph Jastrow, "Autobiography," in *A History of Psychology in Autobiography,* Vol. 1, ed. Carl Murchison (1930; rpt., New York: Russell & Russell, 1961), pp. 135–163. Michael Heidelberger pointed out that Peirce not only introduced Fechner's psychophysics in the United States but also borrowed much from Fechner in his later philosophical works. Although Peirce can hardly be classified as a psychophysical researcher, let alone as a psychologist, the interest in Fechner evident in the article he coauthored with Jastrow was not just incidental. See Michael Heidelberger, *Die innere Seite der Natur: Gustav Theodor Fechners wissenschaftlich-philosophische Weltauffassung [The inner side of nature: Gustav Theodor Fechner's scientific-philosophical worldview]* (Frankfurt am Main: Klosterman, 1993), pp. 316–320. In 1885 Richet was a cofounder of the French Société de Psychologie Physiologique. See Geneviève Paicheler, *L'invention de la psychologie moderne* (Paris: L'Harmattan, 1992), p. 10.

This content downloaded from 129.125.148.19 on Mon, 10 Dec 2018 14:54:55 UTC
All use subject to https://about.jstor.org/terms

the effectiveness of a wide range of administrative regulations. Research on thought transference will be further integrated as part of this larger account.

THE TRANSFER OF TRAINING PHENOMENON

The psychophysicists' methodological worries did not end with experimental bias due to expectations caused by fixed sequences of the stimuli. A related factor suspected of causing contamination was called "habituation" or "training." As early as the 1850s, there were reports of striking decreases of the threshold of perception that occurred simply because the subjects were trained. In the pinprick experiments, switching stimulus application to the other arm or leg did not solve this problem of the unstable threshold, as the increased sensitivity seemed to transfer automatically to other parts of the body, particularly to symmetrical parts.[18]

By the end of the century, experiments on this so-called transfer of training phenomenon were conducted as intensively as was research on the transfer of thoughts elsewhere. Formerly a suspected source of bias, the transfer effect had itself become a psychophysical issue. As researchers looked beyond the refinement of sensory discrimination, subjects were asked to thrust needles through small holes with one hand and then checked for improvement in the other hand, or trained in tapping with the right big toe and checked for transfer to the left big toe.[19]

One imagines distinguished men, in suits or white coats, sitting on wooden laboratory chairs and concentrating intently on their bare feet. It should be noted, though, that these experiments had a specific aspect in common with the present-day random group experiment that the former experiments had not. Whereas the transfer of training experiments were designed to establish the effects of an intervention, the pinprick and psychical trials had been conducted "just" to prove or disprove the existence of particular phenomena such as the lawlike relationship of stimuli and sensations or the truth of telepathy. To put it another way: whereas in the former experiments the subjects were not tested for improvement of their capacities in thought transference or skin sensibility, the transfer of training experiments were meant for drawing causal conclusions. With a slight anachronism, one could say that in the latter cases the result of a particular treatment was at issue.

The present-day ideal for evaluating treatments is comparison of randomly composed experimental and control groups, but in these early training experiments no control groups were employed and (if only for that reason) subjects were not randomly assigned to groups. Causal inferences were drawn from simple comparison of the subjects' results before and after training. For instance, first the left hand's ability was recorded, then the right hand was trained, and finally the left hand's improvement was established. Likewise, first one

[18] In 1858 Fechner published a treatise entitled "Beobachtungen, welche zu beweisen scheinen, dass durch die Uebung der Glieder der einen Seite die der andern zugleich mitgeübt werden [Observations seeming to prove that by training the limbs at one side those at the other are simultaneously trained]." For a brief history see Charles W. Bray, "Transfer of Learning," *Journal of Experimental Psychology,* 1928, *11:*443–467, on pp. 444–446; and Robert S. Woodworth and Harold Schlosberg, *Experimental Psychology,* rev. ed. (New York: Holt, 1954), pp. 738–743.

[19] G. A. Tawney, "Ueber die Wahrnehmung zweier Punkte mittelst des Tastsinnes, mit Rücksicht auf die Frage der Uebung und die Entstehung der Vexirfehler [On the perception of two points by the cutaneous sense and considering the question of training and the origination of mistakes due to deception]," *Phil. Stud.,* 1897, *13:*163–221; W. W. Davis, "Researches in Cross-education" (1898), discussed in Bray, "Transfer of Learning"; and E. W. Scripture, T. L. Smith, and E. M. Brown, "On the Education of Muscular Control and Power," *Studies of the Yale Psychological Laboratory,* 1894, *2:*115.

This content downloaded from 129.125.148.19 on Mon, 10 Dec 2018 14:54:55 UTC
All use subject to https://about.jstor.org/terms

big toe's pliancy was examined, then the other one was trained, and subsequently the untrained one was tested again.[20] Nevertheless, such experiments would become significant for the earliest applications of random grouping.

TRANSFER AS AN EDUCATIONAL RESOURCE

It is important to note that America was in its much-described Progressive Era, a large-scale movement for social and political reform launched at the end of the nineteenth century.[21] Liberal *laissez-faire* politics was widely rejected; instead, people called for centralized government intervention in social and economic affairs. The state was given responsibility for the welfare of the citizenry, and administrative agencies were installed to plan and steer social progress. Moreover, it was strongly felt that, rather than being based on personal intuitions, administrative measures should be grounded on carefully collected facts and subjected to standardized evaluation of their effectiveness.

This movement for social amelioration through government and technoscientific intervention also made its way into education. Administrators came to look upon learning ability as a natural resource to be discovered and exploited as efficiently and economically as possible; they called for impersonal, standardized assessment of individual capacities and educational approaches. As part of the "cult of efficiency" in education, American psychology found a major area of application. Its research interest in mental phenomena was translated into an interest in mental capacities. The most notorious locus of this interest was the network, amounting to an "Educational Trust," created by psychologists at Stanford University in California with their colleagues at New York's Columbia University.[22] Under the undisputed leadership of the Columbia psychologist Edward L. Thorndike, quantities of data were gathered on behalf of scientific management in education.

From the very beginning, the Stanford-Columbia network attributed practical value to the classical psychophysical transfer of training experiments. Not much time was spent on the transferability of left to right big-toe pliancy, but the psychophysical issue of transfer attracted attention in the context of a debate on "formal discipline"—that is, on the teaching

[20] This pretest–training–posttest scheme amounts to a small but intriguing adaptation of the scheme followed with the Tübingen pinprick experiments. Let the stimuli (double pinpricks) be indicated as *X*, the randomly inserted control trials (deception trials with single pinpricks or zero distances) as Not *X*, and the responses as *O*. Then the Tübingen experiments can be schematically represented as *X–O–X–O*–Not *X–O–X–O–X–O*–Not *X–O*, etc. In the later training experiments, the stimuli became the training and the responses became the tests. It can easily be seen that the test–training–test scheme is a particular "cutout" from the Tübingen trial, namely (Not *X*)–*O–X–O*. The main difference is that the classical *random* alternation of *X*'s and Not *X*'s was not possible in the training experiments, for once a subject has been trained the effect cannot be undone. In this way, the former haphazard alternation of *X*'s and Not *X*'s was almost automatically reduced to the fixed scheme of (Not *X*)–*O–X–O*.

[21] Some historical works on the Progressive Era are Paul S. Boyer, *Urban Masses and Moral Order in America, 1820–1920* (Cambridge, Mass.: Harvard Univ. Press, 1978); Arthur A. Ekirch, *Progressivism in America: A Study of the Era from Theodore Roosevelt to Woodrow Wilson* (New York: New Viewpoints, 1974); and Robert H. Wiebe, *The Search for Order, 1877–1920* (New York: Hill & Wang, 1967).

[22] The expression is borrowed from Dale A. Stout, "Statistics in American Psychology: The Social Construction of Experimental Psychology, 1900–1930" (Ph.D. diss., Univ. Edinburgh, 1987). Historians have amply documented the Progressive movement in the schools. Some works include Raymond E. Callahan, *Education and the Cult of Efficiency* (1962; rpt., Chicago: Univ. Chicago Press, 1964); Lawrence A. Cremin, *The Transformation of the School: Progressivism in American Education, 1876–1957* (New York: Vintage, 1964); David Tyack and Elisabeth Hansot, *Managers of Virtue: Public School Leadership in America, 1820–1980* (New York: Basic, 1982); Danziger, *Constructing the Subject* (cit. n. 7); and JoAnne Brown, *The Definition of a Profession: The Authority of Metaphor in the History of Intelligence Testing, 1890–1930* (Princeton, N.J.: Princeton Univ. Press, 1992).

This content downloaded from 129.125.148.19 on Mon, 10 Dec 2018 14:54:55 UTC
All use subject to https://about.jstor.org/terms

of subjects such as mathematics and Latin. Some educators, arguing that these subjects train the mind, strengthening general mental capacities, argued that they should be taught even to students who would never use the material as such. This viewpoint hearkened back to work on the phenomenon of transfer: the question arose whether, like sensorimotor functions, cognitive abilities acquired in one context would automatically transfer to related areas. Thus the phenomenon of transfer, which had first been investigated in psychophysical laboratories as a source of experimental bias and had then become an interesting issue in its own right, now became a matter of educational management and control. Simultaneously, the responses of experimental subjects, formerly seen as neutral reports of inner sensations or memories, became expressions of personal achievement to be compared to other subjects' performance. Responses were now transformed into "test scores," whereas the stimuli became "trainings."[23]

A quite early example of the new orientation is offered by a series of articles published in 1901 by Thorndike and his colleague Robert S. Woodworth. Thorndike and Woodworth considered the notion of formal discipline under the telling title "Influence of Improvement in One Mental Function upon the Efficiency of Other Functions," which clearly displays both the inheritance from psychophysics and the instrumental turn psychology was taking.[24] Their experimental design was borrowed directly from the psychophysical transfer experiments. For instance, they first tested a given function, like estimating weights or marking misspelled words; trained their subjects on a slightly different function, like estimating other weights and marking other misspelled words; and then repeated the first test. As in traditional psychophysical experiments, the same experimental subjects were tested both before and after the training was given.

In contemporary social science research methodology this design is called the "one-group pretest-posttest design"; it is discussed only to illustrate a series of "threats to validity."[25] A common objection is that the pretest may also act as a form of training: anyone who takes the same test twice is likely to do better the second time. Besides, methodologists point out, in the period between administration of the first and second tests, events other than the treatment may occur that might influence a subject's performance on the second test. It was in order to avoid these problems that—in 1907—the control group was introduced. Interestingly, it seems to have been John Edgar Coover—saved from historical oblivion by Hacking's treatment of him as a clever psychical experimenter of the 1910s—who first introduced the remedy of the nontreated control group in psychology.

MORE ON STANFORD'S COOVER

Before Coover entered academia, he had already gained some experience of life. He had finished vocational school and worked as a telegraphist with the Union Pacific Railway and as a journeyman printer and publisher of a country newspaper. Next, he attended a State Normal School and worked as a school principal. In 1899, at the age of twenty-seven, he registered at the newly founded Stanford University. There the skills he had acquired not only assured him of some income but also guided his academic interests.[26]

[23] On this transition see Danziger, *Constructing the Subject,* pp. 136–155. For a discussion of formal discipline see Walter B. Kolesnik, *Mental Discipline in Modern Education* (Madison: Univ. Wisconsin Press, 1958).

[24] Edward L. Thorndike and Robert S. Woodworth, "Influence of Improvement in One Mental Function upon the Efficiency of Other Functions," *Psychological Review,* 1901, *8:*247–262, 384–395, 553–564.

[25] For an extensive discussion see Campbell and Stanley, *Experimental and Quasi-experimental Designs for Research* (cit. n. 1).

[26] Biographical information is from John Edgar Coover, *Formal Discipline from the Standpoint of Experimental Psychology* (Psychological Monographs, 87) (Princeton, N.J.: Princeton Univ. Press, 1912), p. 308.

This content downloaded from 129.125.148.19 on Mon, 10 Dec 2018 14:54:55 UTC
All use subject to https://about.jstor.org/terms

Though interested in education, Coover did not enter the education department. Instead, he began studying at the one-man Department of Psychology. His teacher was the psychophysical researcher Frank Angell, who had been taken on as a full professor of psychology in 1892, only a year after he gained his Ph.D. with Wilhelm Wundt at Leipzig.[27] Nevertheless, as befitted a man of his background, Coover soon gave Angell's theoretical interests an applied twist. He started experimenting on the issue of formal discipline, work that in 1905 resulted in a master's thesis entitled "Formal Discipline from the Standpoint of Experimental Psychology."

In a 1907 article, coauthored by Angell, Coover advanced the idea of employing a nontreated control group to study the question of formal discipline. He first discussed some experiments by others—including Thorndike and Woodworth, whose 1901 work he described as "rough experiments of very little value." His chief reason for rejecting Thorndike and Woodworth's experiments was their lack of "introspection": that is, the reagents were not asked what they were thinking while taking the tests. Most important in the context of the present history, however, was his remark that "it is to be regretted that the authors do not carry on a 'control' experiment along with their tests to ascertain the training effect of the tests themselves and to throw additional light on the changes taking place in the training intervals."[28] In Coover's own experiments the notion of formal discipline was supported by observed differences between the test results of "reagents" who had been given training and "control reagents" who had not.

With this design, the usual sequence of "test–training–test" was, so to speak, cut in two: one group underwent only the first "test" part, the other the "training–test" half.[29] In other words, the classical misleading control *trial* in which the stimulus was withheld was now applied to special control *subjects.* Yet another actor had been added to the cast of participants in experimentation. First Fechner's dual role as subject and experimenter had been divided between two persons; then the addition of an artificial randomizer relieved the experimenter of the job of randomizing the stimuli; now different types of experimental subjects were also introduced.[30]

[27] Before he went to Stanford, Frank Angell founded the psychological laboratory at Cornell. He should not be confused with his younger cousin James Rowland Angell, who was also involved in psychophysical research and was director of the psychological laboratory at Chicago in the 1890s. On both psychologists see Carl Murchison, ed., *The Psychological Register* (Worcester, Mass.: Clark Univ. Press, 1929–), pp. 5–6; on Frank Angell see also Henry L. Minton, *Lewis M. Terman: Pioneer in Psychological Testing* (New York: New York Univ. Press, 1988).

[28] John Edgar Coover and Frank Angell, "General Practice Effect of Special Exercise," *Amer. J. Psychol.,* 1907, *18:*328–340, on p. 329. From about 1903 until about 1913 there was a strong shift in German and American psychology away from Wundt toward introspection. Leading dissidents like Oswald Külpe at Würzburg and Titchener maintained that collecting subjective reports was an important purpose of experimentation. Their point of view found many followers. One of them was Frank Angell, who was Titchener's close friend and his predecessor at the psychological laboratory at Cornell. See Boring, *History of Experimental Psychology* (cit. n. 12), pp. 410–420.

[29] Coover and Angell again included a pretest in their design for both groups. After some years the objection was raised that although this enhanced the "internal validity" of the experiment, it diminished the "external validity." With the so-called pretest-posttest control-group design, it remains unclear whether training that has a positive effect will also yield this effect in real life, when it is not preceded by a pretest. This point was raised in 1949 by R. L. Solomon, who dealt with it by implementing a four-group design. It was also discussed in 1957 by Campbell, who added that Fisher did not prescribe pretests either. See R. L. Solomon, "An Extension of the Control Group Design," *Psychol. Bull.,* 1949, *46:*137–149; and Donald T. Campbell, "Factors Relevant to the Validity of Experiments in Social Settings," *ibid.,* 1957, *54:*297–312, on p. 303.

[30] Another important change (which I will not discuss here) is that groups of subjects had now appeared on the scene. On this transition see Kurt Danziger, "Statistical Method and the Historical Development of Research Practice in American Psychology," in *The Probabilistic Revolution,* Vol. 2: *Ideas in the Sciences,* ed. Lorenz Krüger, Gerd Gigerenzer, and Mary S. Morgan (Cambridge, Mass.: MIT Press, 1987), pp. 34–49; and Danziger, *Constructing the Subject* (cit. n. 7), pp. 113–117.

This content downloaded from 129.125.148.19 on Mon, 10 Dec 2018 14:54:55 UTC
All use subject to https://about.jstor.org/terms

In describing these experiments, Coover repeated his reasons for employing control groups three more times. This repeated clarification and the use of quotation marks with the word *control* indicate that in 1907 the idea of a control group was virtually unknown in psychology.[31] Indeed, Coover and Angell's article offers a correction to the standard historiography of the control group. Ever since the historical reconstructions presented by R. L. Solomon and Edwin G. Boring, a 1908 British example of educational group comparison has been cited as the earliest use of control groups in psychology and a 1926 example from American psychophysics as the first application of the word *control* in relation to psychological group comparison. In fact, Coover and Angell, in their largely neglected article, used the term *control reagents* for their untreated controls nineteen years before Solomon's and Boring's example.[32] Coover is an intriguing figure in the history of science—and not just for his work in psychical research. (See Figure 3.)

SAVING MONEY: RANDOM GROUPING AT SCHOOL

In 1912, at the age of forty, Coover earned his Ph.D., again with a thesis entitled "Formal Discipline from the Standpoint of Experimental Psychology." In the same year, Stanford University finally made up its mind to accept a large sum offered decades before by the brother of its founder under the regretted condition that the money would be spent on psychical research. Angell, commissioned to find a suitable psychical researcher, drew up the profile of a typical *psychophysical* experimenter, and Coover, of course, fit the de-

[31] Coover and Angell, "General Practice Effect of Special Exercise" (cit. n. 28), pp. 331–332, 335. For early nonhuman controls see William G. Cochrane, "Early Development of Techniques in Comparative Experimentation," in *On the History of Statistics and Probability,* ed. D. B. Owen (New York: Dekker, 1976), pp. 1–26. Francis Galton's research proposal for testing the effects of praying, quoted in note 3, above, includes one of the exceptional early examples of human controls. Abraham M. Lilienfeld discussed some nineteenth-century cases of human controls in medicine; however, these cases mostly involved differently treated rather than untreated groups: Abraham M. Lilienfeld, "Ceteris Paribus: The Evolution of the Clinical Trial," *Bulletin of the History of Medicine,* 1982, *56:*1–18. Another example from medical research is described in William Coleman, "Experimental Physiology and Statistical Inference: The Therapeutic Trial in Nineteenth-Century Germany," in *Probabilistic Revolution,* Vol. 2, ed. Krüger *et al.,* pp. 201–229.

[32] Solomon, "Extension of the Control Group Design" (cit. n. 29); and Edwin G. Boring, "The Nature and History of Experimental Control," *Amer. J. Psychol.,* 1954, *67:*573–589. Whereas Coover and Angell's 1907 article was ignored, Thorndike and Woodworth's 1901 experiments were very well received. Kolesnik reported that by the 1930s thousands of American teachers were taught, by reference to "Thorndike and Woodworth (1901)," that the doctrine of formal discipline was a myth. In 1930 their article was still described as "a veritable bombshell into the educational camp": Kolesnik, *Mental Discipline in Modern Education* (cit. n. 23), pp. 7, 34. As a matter of fact, Woodworth seems to be the one who first denied Coover and Angell their rightful place in the historiography of the control group. Whereas Boring mostly borrowed his discussions of the control group in psychology from Solomon, Solomon appears to have copied his history from Woodworth. His reconstruction is just like that on the pages of Woodworth's *Experimental Psychology* to which his bibliography refers. Woodworth's history of the use of control groups began with the research he conducted with Thorndike in 1901 (this was repeated by Solomon, but Boring pointed out that Thorndike and Woodworth, in their thick report of countless experiments, spent only eight casual lines discussing a trial with a control group). Woodworth's next instances are the now-standard 1908 British and 1926 American cases. See Robert S. Woodworth, *Experimental Psychology* (New York: Holt, 1938), pp. 178–181. It is highly unlikely that at the time he was writing his overview he had simply forgotten Coover and Angell's dismissal of his own work. Elsewhere in *Experimental Psychology,* Woodworth overtly expressed his annoyance with Coover's dissertation on transfer experiments (p. 198), and a mimeographed version of his book was already used for teaching at Columbia in 1909. See A. S. Winston, "Robert Session Woodworth and the 'Columbia Bible': How the Psychological Experiment Was Redefined," *Amer. J. Psychol.,* 1990, *103:*391–401. Of course, both Woodworth's maneuverings and all these laborious searches for the "origins" of the control group (including my own) only underline the pivotal role ascribed to it from the 1930s to the present.

This content downloaded from 129.125.148.19 on Mon, 10 Dec 2018 14:54:55 UTC
All use subject to https://about.jstor.org/terms

Figure 3. *John Edgar Coover. The date and source are unknown, but a caption states that in his spare time Coover developed a new system of shorthand "specially for students to simplify notetaking." (Courtesy Department of Special Collections, University Archives, Stanford University Libraries, Stanford, California.)*

scription.[33] Coover's part in the history of randomization after his appointment has been described by Hacking: in his experiments on thought transference Coover meticulously randomized all kinds of orders.[34] (See Figure 4.)

[33] For entertaining information on this financial offer with strings attached see Hacking, "Telepathy" (cit. n. 4); and F. C. Dommeyer, "Psychical Research at Stanford University," *Journal of Parapsychology,* 1975, *39:*173–205. The events have been briefly described by Angell, who reported that he looked for an experimenter with "special extensive training in the psychology of motor automatisms and of subliminal impressions, in the ideational and affective processes underlying belief and conviction, in illusions of perception and the value of evidence": Frank Angell, "Introduction," in Coover, *Experiments in Psychical Research* (cit. n. 16), pp. v–xxiii, on pp. xx, xxi.

[34] There is an indication that Coover was aware of the imperative of random ordering before he went into psychical experimentation. The 1907 article coauthored with Angell expressly states that the order of the stimuli was not randomized but that therefore precautions were taken to trick the subjects into believing that the orders were determined by chance: Coover and Angell, "General Practice Effect of Special Exercise" (cit. n. 28), p. 331. As to Coover's further history: after his 1917 work on psychical research he virtually ceased publishing, and after Angell's retirement in 1922 he became a displaced person in a department led by Lewis Terman, an influential educational psychologist and friend of Thorndike and Woodworth. Until Coover's own retirement in 1937, he "remained an irritant" for Terman, who saw him as an "Angell holdover" whom he could not dismiss because of another very large sum donated to the department for psychical research. However, Quinn McNemar, a famous methodologist himself, remembered him as "owl-wise Coover," with whom one had to be patient in the face of his endless elaborations but who was willing to help with the most complicated logical problems. See Dommeyer, "Psychical Research at Stanford University"; Mauskopf and McVaugh, *Elusive Science* (cit. n.

This content downloaded from 129.125.148.19 on Mon, 10 Dec 2018 14:54:55 UTC
All use subject to https://about.jstor.org/terms

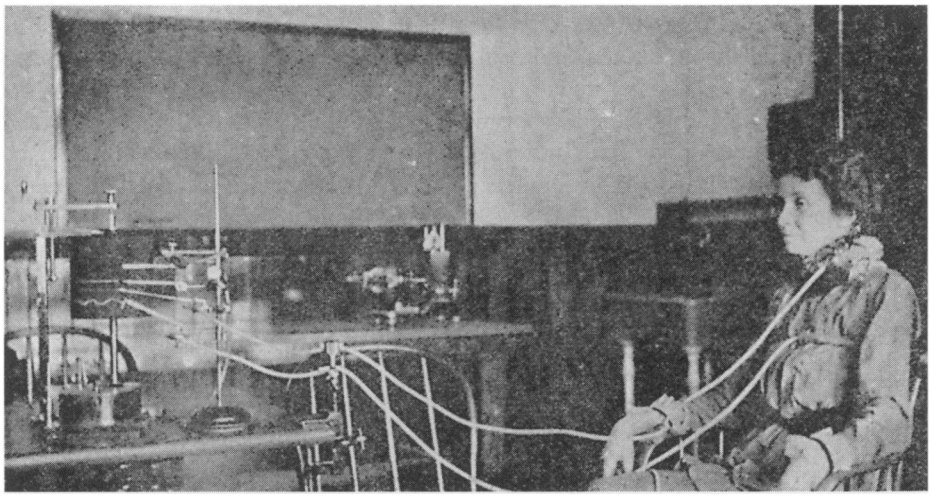

Figure 4. *Psychical research in Coover's laboratory. The apparatus is a kymograph, which graphically records responses. (From John Edgar Coover,* Experiments in Psychical Research at Leland Stanford Junior University *[1917; rpt., New York: Arno, 1975], facing page 508.)*

While Coover applied himself to psychical research, experimental educational research thrived. Educational psychologists, armed with contracts from educational administrators, left their university laboratories for experimentation in real schools. "Treatments" now included ways of teaching, educational measures like punishment and praise, and all kinds of teacher behaviors and classroom circumstances. Whereas Boring counted no control groups at all in reports of experiments published in the 1916 volume of the *American Journal of Psychology,* Kurt Danziger found that 14 percent of those described in the same year's *Journal of Educational Psychology* used control groups.[35] In the world outside the orderly laboratory, however, many more variables had to be controlled; moreover, a bigger and more varied audience had to be convinced of the merits of psychological research. Educational psychologists soon began to discuss the price they paid for comparing the results of different groups rather than comparing subjects' results before and after treatment: while they had abolished the possible unwanted effect of the pretest, they had also introduced possible unwanted differences between the groups. The experimenters faced the problem of ascertaining that distinct outcomes for the experimental and control groups were caused by the treatments and not by some other variables.[36]

13), pp. 253–254, 285; Minton, *Lewis M. Terman* (cit. n. 27), p. 135; and Quinn McNemar, "Autobiography," in *A History of Psychology in Autobiography,* Vol. 7, ed. Gardner Lindzey (San Francisco: Freeman, 1980), pp. 316–317.

[35] Danziger scrutinized the *Journal of Educational Psychology* by taking samples from the volumes of 1914–1916, 1924–1926, and 1934–1936; between 14 percent and 18 percent of the articles published in these periods used control groups. In the *Journal of Applied Psychology* volumes of 1917–1920 and 1934–1936, between 1 percent and 6 percent reported control groups. As to the postwar period, in 1951 Boring found that only 52 percent of articles in the *American Journal of Psychology* mentioned control groups. See Boring, "Nature and History of Experimental Control" (cit. n. 32), p. 587; and Danziger, *Constructing the Subject* (cit. n. 7), pp. 113–115.

[36] Coover again was among the first to raise the issue. Whereas his 1907 paper with Angell had shown no awareness of the problem of dissimilar groups, his 1912 Ph.D. dissertation expressed doubts on the topic: "Our trained reagents were more mature than the control reagents and were experienced in laboratory work." And in

This content downloaded from 129.125.148.19 on Mon, 10 Dec 2018 14:54:55 UTC
All use subject to https://about.jstor.org/terms

In educational research conducted in the 1920s, it became customary to handle this problem by subjecting children to preliminary tests on suspected factors and composing matched groups on the basis of the test results. A slightly easier way of matching was to experiment with existing school classes with the same mean scores and standard deviations on particular tests. When matching was not viable, the procedure of the so-called rotation experiment was sometimes followed: several groups were subjected to a full series of different treatments given in variable order. An early rotation experiment, undertaken by Thorndike and his Ph.D. student William Anderson McCall, investigated whether fresh or recirculated air would make children perform better on mental tests.[37] It seems to have been McCall who first suggested composing the groups at random.

As biographical and autobiographical notes report, McCall was raised in a very poor family that worked in the Kentucky and Tennessee coal mines. From early childhood, he had to contribute to the family income by laboring half of each year as a coal digger. In 1913, at the age of twenty-two, he borrowed some money and left for New York. "With the black dust of years of summer work in the coal mines imbedded in his hands," he arrived at Columbia University, where he gained his third bachelor's degree, as well as an M.A. and a Ph.D., from Teachers College within three years.[38]

At Columbia, with its instrumental approach to teaching, McCall was the right man at the right place. He acquired a job as an instructor and in 1923 published a methodological treatise entitled *How to Experiment in Education.* In the introduction, McCall estimated that increased efficiency of education could save a full year of teaching per person, and he worked out that psychological experimentation to determine the best teaching methods could save $134,680,000,000,000 over the next 100 generations of Americans, which would amount to "790 times the costs of the first World War" and "390 times the costs of all wars in recorded history." Then he advanced a way of making the work of his profession still more profitable to society. The book considers seven options for equalizing groups; the first is "groups equated by chance," which McCall discussed as "an economical substitute" for matching. Experimental psychology, he claimed, could tip the economic balance still more favorably if the expensive method of matching were exchanged for the cheaper procedure of randomization.[39]

McCall did not take the task of randomization lightly. For example, he rejected the procedure of writing numbers on pieces of paper because papers with larger numbers contain more ink than those with smaller ones and are therefore likely to sink further to

the final section of the thesis he suggested that in fact "the control reagents should equal the trained reagents in number, in initial efficiency, and in the facility in introspection, or their results may not be comparable with those of the trained reagents." See Coover, *Formal Discipline* (cit. n. 26), pp. 83, 223–224.

[37] This experiment, conducted in 1916, is described in William Anderson McCall, *How to Experiment in Education* (1923; rpt., New York: Macmillan, 1926), p. 194.

[38] McCall published a moving thirty-four-page autobiography that he wrote as a young child, with a biographical preface and afterword by Helen Duder: William Anderson McCall, *I Thunk Me a Thaut* (New York: Teachers College Press, 1975), quotation from p. 37. Duder explains that McCall gained his first bachelor's degree in 1911 at Cumberland College in Williamsburg, Kentucky. His father allowed him to attend the college because it offered a one-term telegraphy course, but the college president arranged for him to stay for four years. Next, the president helped him gain a position at Lincoln Memorial University at Cumberland Gap to teach two terms of psychology. Here he earned his second bachelor's degree. Biographical information on McCall is also available in the Milbank Memorial Library, Special Collections, Teachers College, Box 44, Finding Aid "Will McCall" and Finding Aid "Teachers College News," Columbia University, New York. Some biographical facts and McCall's publications up to 1975 are listed in Clare D. Kinsman, ed., *Contemporary Authors: A Bio-bibliographical Guide to Current Authors and Their Works* (Detroit: Gale, 1975), p. 426.

[39] McCall, *How to Experiment in Education* (cit. n. 37), pp. 41–42.

This content downloaded from 129.125.148.19 on Mon, 10 Dec 2018 14:54:55 UTC
All use subject to https://about.jstor.org/terms

the bottom of a container. He nevertheless maintained that "any device which will make the selection truly random is satisfactory."[40]

EXPERIMENTAL PSYCHOLOGY AND R. A. FISHER

Experimental psychology has long staked its claim to the status of a science by emphasizing its methodological inheritance from the natural sciences. But random group experimentation, nowhere more the apogee of methodological rigor than in psychology,[41] appears to be firmly rooted in the discipline's own professional and social history. Moreover, rather than the instantaneous discovery of an ingenious individual, the random group design seems to be an unlooked-for outcome of a long historical process.

There are some indications that Fisher, as he developed his methods and techniques for agricultural research, was acquainted with the ways of experimental psychology. The first substantial chapter in his 1935 book *The Design of Experiments* is entitled "The Principles of Experimentation Illustrated by a Psycho-Physical Experiment." The "psycho-physical" experiment of the title is the now-famous "tea experiment," which Fisher had worked out after a colleague at the agricultural research station declined a cup of tea because Fisher had put the milk in the cup before pouring the tea. She maintained that a cup with the wrong milk-tea order did not taste as good as a cup with the right tea-milk order.[42] To test this claim of delicate sensory discrimination, Fisher explained, a series of trials is necessary, and it must be randomly established before each trial whether the milk or the tea should be poured in first.

Fisher's little joke in the title of this chapter points to his awareness of the characteristics of psychophysical experimentation. As Hacking discovered, Fisher was also quite knowledgeable about the methodology of psychical research. There is no evidence, however, that Fisher derived his random group design directly from psychology. Although he advances reasons for randomization that were also presented by the psychologists I have discussed (including McCall's argument that it would save money), he derived the randomization requirement primarily from his new techniques of statistical data analysis. In both his discussion of the tea experiment, where random ordering was at issue, and in his next chapter on a comparative study of cross-fertilized and self-fertilized plants, where

[40] *Ibid.* I do not know whether McCall was aware of randomization of order in psychophysics or psychical research. In his introduction of random grouping, he referred to a third kind of randomization (not yet discussed in this essay): randomly drawing a representative sample. "Just as representativeness can be secured by the method of chance, . . . so equivalence can be secured by chance": *ibid.,* p. 41. Once again the role of psychology in the development of contemporary standard research techniques seems to have been more significant than is generally acknowledged. McCall had discussed this third kind of randomization in an earlier book on psychological test construction, where he explained that composing a psychological aptitude test involves drawing random samples of the ability to be tested: William Anderson McCall, *How to Measure in Education* (New York: Macmillan, 1922), pp. 201–202. Historical studies of random sampling in scientific research report only two Scandinavian "pre-McCall" examples of random sampling; they do not mention any work in psychology. See F. F. Stephan, "History of the Uses of Modern Sampling Procedures," *J. Amer. Statist. Assoc.,* 1948, *43:*12–39, on p. 21; W. C. Kruskal and Frederick Mosteller, "Representative Sampling, IV: The History of the Concept," *Statistical Review,* 1980, *48:*169–195, on p. 179 n 6; and Alain Desrosières, "The Part in Relation to the Whole: How to Generalize? The Prehistory of Representative Sampling," in *The Social Survey in Historical Perspective, 1880–1940,* ed. Martin Bulmer, Kevin Bales, and Kathryn Kish Sklar (Cambridge: Cambridge Univ. Press, 1991), pp. 217–245.

[41] See Andrew S. Winston and Daniel J. Blais, "What Counts as an Experiment? A Trans-disciplinary Analysis of Textbooks, 1930–1970," *Amer. J. Psychol.,* 1996, *109:*599–616.

[42] Fisher did not mention that the example was derived from his own life. This has been revealed by his daughter and biographer: see Joan Fisher Box, "R. A. Fisher and the Design of Experiments, 1922–1926," *American Statistician,* 1980, *34:*1–7.

This content downloaded from 129.125.148.19 on Mon, 10 Dec 2018 14:54:55 UTC
All use subject to https://about.jstor.org/terms

random grouping was involved, Fisher strongly emphasized that randomization is part of his techniques of statistical significance testing.[43]

Whereas the impact of psychology on Fisher may be uncertain, an arrow may certainly be drawn in the opposite direction. As Alexander Lovie has argued, by the 1930s the psychologists' experimental designs had become so complex that the resulting data were hard to manage and interpret. Fisher's solutions reached them via several routes.[44] Some psychologists began to translate them into the language of their own field, which led to some painstaking prose: "To understand better the meaning of the experimental group in terms of blocks and plots, let us examine the two possible arrangements: (1) where a single individual or a group is a block, and (2) where a single individual or group is a plot." Soon a range of handbooks on statistics, many of them written by educational psychologists from the Stanford-Columbia network, passed on Fisher's instructions to wider circles. As one of these psychologists would observe some years later, Fisher's ways quickly "became epidemic just as though psychologists had never previously planned an experiment."[45] Although McCall was no stranger to the psychology community, the introduction of the experimental design with randomly composed experimental and control groups has been attributed to Fisher ever since.[46]

[43] This is not to suggest that Fisher was the stereotypical aloof mathematician. His main sources of inspiration were eugenics and biometrics. On the merging of Fisher's eugenical and agricultural interests see MacKenzie, *Statistics in Britain* (cit. n. 2). On his familiarity with psychophysical experimentation see Hacking, "Telepathy" (cit. n. 4), pp. 449–450. Stigler drew a straight line between Fechner's 1860 *Elemente der Psychophysik* and Fisher's 1935 *Design of Experiments,* saying that Fechner's book on experimental design is "the most comprehensive treatment of that topic" before the one by Fisher: Stigler, *History of Statistics* (cit. n. 2), p. 244.

[44] Alexander D. Lovie, "The Analysis of Variance in Experimental Psychology: 1934–1945," *British Journal of Mathematical and Statistical Psychology,* 1979, *32:*151–178, on p. 152. For instance, in the 1930s Fisher was twice visiting professor at Iowa State College, where the statistician George Snedecor was employed. Snedecor first applied Fisherian methods in an article on the causes of class differences in mathematical achievement, and somewhat later he published a book on Fisher's statistical methods that psychologists found easier to study than the original work. In 1940 Snedecor's Iowa colleague, the educational psychologist Everett Lindquist, followed with the book *Statistical Analysis in Educational Research,* which became a much-cited source for educational experimenters. See George W. Snedecor, *Statistical Methods* (Ames, Iowa: Collegiate, 1937); Everett F. Lindquist, *Statistical Analysis in Educational Research* (Boston: Houghton Mifflin, 1940); Anthony J. Rucci and Ryan D. Tweney, "Analysis of Variance and the 'Second Discipline' of Scientific Psychology: A Historical Account," *Psychol. Bull.,* 1980, *87:*166–184; and Lovie, "Analysis of Variance".

[45] Brent Baxter, "Problems in the Planning of Psychological Experiments," *Amer. J. Psychol.,* 1941, *54:*270–280, on p. 273; and Quinn McNemar, "At Random: Sense and Nonsense," *Amer. Psychol.,* 1960, *15:*295–300, on p. 297. These handbooks include J. P. Guilford, *Fundamental Statistics in Psychology and Education* (New York: McGraw Hill, 1942); Helen Mary Walker, *Elementary Statistical Methods* (New York: Holt, 1943); Truman Lee Kelley, *Fundamentals of Statistics* (Cambridge, Mass.: Harvard Univ. Press, 1947); Allen Louis Edwards, *Statistical Analysis for Students in Psychology and Education* (New York: Rinehart, 1946); Henry Edward Garrett, *Statistics in Psychology and Education* (New York: Longmans, 1947); and Quinn McNemar, *Psychological Statistics* (New York: Wiley, 1949).

[46] Nevertheless, I traced McCall because he was honored in at least one text on educational experimentation for introducing the random group design before Fisher did: Donald T. Campbell and Julian C. Stanley, "Experimental and Quasi-experimental Designs for Research on Teaching," in *Handbook of Research on Teaching,* ed. N. L. Gage (Chicago: Rand McNally, 1963), pp. 171–247; rpt. as Campbell and Stanley, *Experimental and Quasi-experimental Designs for Research* (cit. n. 1). As to his further history: in 1927, McCall had become a professor of measurement and statistics. In the same year he caught tuberculosis, which took years to cure and confined him to a wheelchair for the rest of his life. After his return to work, he did not join forces with the other statisticians who passed on Fisher's statistical techniques to psychologists. He devoted most of his time to educational test construction and to cooperation with universities in China. In 1939 he published a revised version of *How to Measure in Education* under the title *Measurement* (New York: Macmillan). In his spare time, he fought for a world government in which all nations had two votes, "one proportional to population and another proportioned to gross national product or degree of education of the citizenry": Special Collections, Teachers College, Columbia University, Box 44, letter by McCall, 17 Oct. 1970.

This content downloaded from 129.125.148.19 on Mon, 10 Dec 2018 14:54:55 UTC
All use subject to https://about.jstor.org/terms

EPILOGUE: THE MARCH OF THE UNCONTROLLED VARIABLES

In addition to the gradual emergence of the random group design, this essay demonstrates that its development was connected to the instrumental turn of experimental psychology. It originated during the transition of academic psychophysical experimentation into research serving administrative decision making in education. Thus the story of the development of the present-day ideal design is another significant example supporting a central thesis in contemporary historiography of statistics and methodology: that the increase of rules and numbers in the sciences reflects the increase of rules and numbers in bureaucratic management. Theodore Porter, in particular, has recognized that the widespread reliance on overt procedures and quantitative data in the sciences has been strongly encouraged by the needs of twentieth-century democratic societies. The bureaucratic organizations that arise in tandem with these societies, Porter argues, are run by expanding groups of policymakers and administrators. Unlike the authoritative elites of the old days, these officials cannot count on public trust in their personal wisdom or refer to their personal intuition in making decisions. In order to avoid charges of arbitrariness and to gain support for their policies, they have to offer assurance of the impartiality of their views. Thus they have increasingly come to appeal to the sciences, most notably the social sciences, for an impersonal basis of authority. In their turn, these sciences have developed intricate systems of standard procedures for administrative knowledge production.[47]

The random group design is clearly a result of these processes. The story of its origins is an account of methods, initially developed for isolated laboratory research, that were transported into the realm of bureaucratic management in education. There they proliferated and were minutely codified in order to ensure as much algorithmic rationality as possible.[48] I will conclude with a rough sketch of developments from the 1930s to the 1970s. On the one hand, this condensed history will demonstrate that the discipline of psychology was a vital factor in the design's dissemination in other social sciences and other areas of human management and control.[49] On the other hand, it will show that, far from settling methodological issues, the establishment of the ideal of random group design gave impetus to substantial extension of the set of codified routines in the social sciences.

In the 1930s, American social science was booming. More than ever, social amelioration

[47] Theodore M. Porter, *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life* (Princeton, N.J.: Princeton Univ. Press, 1995). See also Ian Hacking, *The Taming of Chance* (Cambridge: Cambridge Univ. Press, 1990).

[48] It should be noted, though, that in a sense American psychology was directed by technosocial ideals from the very early stage of psychophysical laboratory research; see Deborah J. Coon, "Standardizing the Subject: Experimental Psychologists, Introspection, and the Quest for a Technoscientific Ideal," *Technology and Culture,* 1993, *34:*757–783.

[49] Of course, there is much more to say about the dissemination of the random group design in other disciplines. For a somewhat different account of its route from psychology to other social sciences see Robert Brown, "The Delayed Birth of Social Experiments," *History of the Human Sciences,* 1997, *10:*1–23, on pp. 17–19. In a history of its introduction in American medicine, Harry Marks has shown that randomization was applied briefly after World War II. As Marks argues, here the design was chiefly seen as a means of disciplining clinicians in their decisions about which patients should be given and which denied a particular treatment: Harry M. Marks, "Notes from the Underground: The Social Organization of Therapeutic Work," in *Grand Rounds: One Hundred Years of Internal Medicine,* ed. Russell C. Maulitz and Diana E. Long (Philadelphia: Univ. Pennsylvania Press, 1988), pp. 297–336. See also Lilienfeld, "Ceteris Paribus" (cit. n. 31); Winston and Blais, "What Counts as an Experiment?" (cit. n. 41); and Trudy Dehue, "De beleidsmaker: Maatschappelijke experimenten en het dilemma van de verzorgingsstaat [The policy-maker: Social experiments and the dilemma of the welfare state]," in *Met zachte hand: Opkomst en verbreiding van het psychologisch perspectief* [*With gentle force: The emergence and establishment of the psychological perspective*], ed. Jeroen Jansz and Peter van Drunen (Utrecht: Lemma, 1996), pp. 241–259.

This content downloaded from 129.125.148.19 on Mon, 10 Dec 2018 14:54:55 UTC
All use subject to https://about.jstor.org/terms

was seen as a matter of centralized government and scientific control. Economists, sociologists, and political scientists such as William Ogburn, Charles Merriam, and Samuel Stouffer were engaged as advisors on Roosevelt's governmental initiatives. These highly influential men introduced psychological methods into the realm of social policy. For instance, Charles Merriam ("Uncle Charley" to the president) brought psychologists to his meetings and conferences and "bombarded" his colleagues in the political science department of the University of Chicago with texts on psychological methods.[50] Through such connections, psychologists became involved with administrative research projects outside their traditional province of education. Psychologists like Louis Thurstone and Paul Lazarsfeld were employed in social science departments, where they devised methods not only for measuring people's attitudes and opinions but also for assessing the success of radio talks and movies in influencing their feelings and beliefs on subjects like immigration, voting, bootlegging, and war.[51] The investigative routines of educational psychology seemed applicable in these contexts, and here too random group design was seen as the paragon procedure. However, this ideal soon proved to be more of a guiding than a practicable principle. Its unfeasibility became particularly clear during World War II in the research branch of the army's Morale Division.

At first the research branch, led by Samuel Stouffer, specialized in gathering information about soldiers' opinions and attitudes; somewhat later it was also charged with improving morale and with evaluating the results of such educational attempts. A film series, *Why We Fight,* was devised, and an experimental section staffed mostly by psychologists was set up for the objective assessment of the films' effects. It was agreed that the movies should be evaluated by comparing randomly composed groups of soldiers that had and had not seen them. However, a problem arose because the soldiers became suspicious when some of them were summoned to view a movie and others were not. Random assignment appeared to cause an unexpected new kind of experimental bias. Although pretesting showed that existing army units differed in many consequential respects, these units nonetheless had to constitute the experimental and the control groups. And that was only one of many methodological trespasses. After the war, researchers published four volumes entitled *Studies in Social Psychology in World War II.* The third, on the activities of the experimental section, was hardly a self-assured presentation of collected knowledge. Rather, the volume offered an overview of dilemmas researchers face when experimental research moves out of the laboratory into real life.[52]

Nevertheless, the experimenters insisted that amateur assessments of the effectiveness

[50] Dorothy Ross, *The Origins of American Social Science* (Cambridge: Cambridge Univ. Press, 1991), p. 455. See also Gene Martin Lyons, *The Uneasy Partnership: Social Science and the Federal Government in the Twentieth Century* (New York: Russell Sage, 1969); and Jennifer Platt, *A History of Sociological Research Methods in America, 1920–1960* (Cambridge: Cambridge Univ. Press, 1996).

[51] See Louis Thurstone, "Autobiography," in *A History of Psychology in Autobiography,* Vol. 4, ed. Edwin G. Boring (Englewood Cliffs, N.J.: Prentice-Hall, 1952), pp. 295–321; and Paul F. Lazarsfeld, "An Episode in the History of Social Research: A Memoir," in *The Intellectual Migration: Europe and America, 1930–1960,* ed. Donald Fleming and Bernard Bailyn (Cambridge, Mass.: Harvard Univ. Press, 1969), pp. 270–337. For more early examples of psychological field experiments in areas other than education see Gardner Murphy, Lois Barclay Murphy, and Theodore M. Newcomb, *Experimental Social Psychology: An Interpretation of Research upon the Socialization of the Individual* (New York/London: Harper, 1937).

[52] Social Sciences Research Council, *Studies in Social Psychology in World War II,* Vol. 3: *Experiments on Mass Communication,* by Carl Hovland, Arthur A. Lumsdaine, and Fred D. Sheffield (Princeton, N.J.: Princeton Univ. Press, 1949). For a history of social science research in World War II see Peter Buck, "Adjusting to Military Life: The Social Sciences Go to War, 1941–1950," in *Military Enterprise and Technological Change: Perspectives on the American Experience,* ed. Merritt Roe Smith (Cambridge, Mass.: MIT Press, 1985), pp. 203–253.

This content downloaded from 129.125.148.19 on Mon, 10 Dec 2018 14:54:55 UTC
All use subject to https://about.jstor.org/terms

of social policies are always worse than professionally devised second-best procedures. As Stouffer put it, in the absence of professional experimental design "there is all too often a wide-open gate through which uncontrolled variables can march."[53] In the 1950s, methodological treatises began to appear that still advanced the random group design as the ideal experiment but simultaneously recognized that generally people cannot be assigned to experimental conditions as haphazardly as plants can be randomly assigned to plots. As a remedy they presented long series of alternative designs. To continue Stouffer's metaphor, these methodological texts catalogued an ever more sophisticated and proliferating armory against a continuing invasion of belligerent variables.

Among the first methodologists to write in this vein was the young psychologist Donald T. Campbell, who had started his career as an army attitude and propaganda researcher. In 1957 Campbell established his reputation as a keen methodologist with an astute article on a variety of experimental designs and accompanying factors that may bias the results. With the statistical assistance of Julian C. Stanley, he subsequently published an essay entitled "Experimental and Quasi-experimental Designs for Research on Teaching" that introduced the term *quasi experiments* for designs that do not meet the requirement of random assignment. In 1966 the essay was reprinted as a separate book under the briefer and more embracing title *Experimental and Quasi-experimental Designs for Research.*[54] This book, now widely known as "Campbell and Stanley," established Campbell's status as a routinely consulted methodologist of the social sciences. Other publications by Campbell, on ranges of quasi-experimental designs, were to follow, some under flamboyant titles like "The Social Scientist as Methodological Servant of the Experimenting Society" and "Reforms as Experiments" that by themselves testify to the close relation between elaborate methodology and social administration.[55]

[53] Samuel A. Stouffer, "Some Observations on Study Design," *American Journal of Sociology,* 1949–1950, *55:*335–361, on p. 357.

[54] Donald T. Campbell, "Factors Relevant to the Validity of Experiments in Social Settings," *Psychol. Bull.,* 1957, *54:*297–312; Campbell and Stanley, "Experimental and Quasi-experimental Designs for Research on Teaching" (cit. n. 46); and Campbell and Stanley, *Experimental and Quasi-experimental Designs for Research* (cit. n. 1). See also Campbell, "Comment: Another Perspective on a Scholarly Career," in *Scientific Inquiry and the Social Sciences: A Volume in Honor of D. T. Campbell,* ed. Marilynn B. Brewer and Barry E. Collins (San Francisco: Jossey-Bass, 1981), pp. 454–486; his army work is discussed on pp. 460–461.

[55] Donald T. Campbell, "The Social Scientist as Methodological Servant of the Experimenting Society," *Policy Studies Journal,* 1973, *2:*72–75; and Campbell, "Reforms as Experiments," *Amer. J. Psychol.,* 1969, *24:*409–429. The most renowned title is Thomas D. Cook and Campbell, *Quasi-experimentation: Design and Analysis Issues for Field Settings* (Boston: Houghton Mifflin, 1979). In the 1973 issue of the journal *Evaluation* Campbell was interviewed as "one of the foremost expositors of experimental design" and "the author the readers have most frequently requested as a subject for an interview": Susan Salasin, "Experimentation Revisited: A Conversation with D. T. Campbell," *Evaluation,* 1973, *1:*7–13. Campbell's wide reputation is also acknowledged in, e.g., James A. Caproso and Leslie L. Roos, eds., *Quasi-experimental Approaches: Testing Theory and Evaluating Policy* (Evanston, Ill.: Northwestern Univ. Press, 1973); Henry W. Riecken and Robert F. Boruch, eds., *Social Experimentation: A Method for Planning and Evaluating Social Intervention* (New York: Academic, 1974); Stuart S. Nagel, ed., *Policy Studies and the Social Sciences* (New Brunswick, N.J.: Transaction, 1979); Brewer and Collins, eds., *Scientific Inquiry and the Social Sciences;* Lee J. Cronbach, *Designing Evaluations of Educational and Social Programs* (San Francisco: Jossey-Bass, 1983); William M. K. Trochim, ed., *Advances in Quasi-experimental Design and Analysis* (San Francisco: Jossey-Bass, 1986); William R. Shadish and Charles S. Reichardt, "Overview," in *Evaluation Studies Review Annual,* Vol. 12 (London: Sage, 1987), pp. 13–30; and Martin Bulmer, ed., *Social Science and Social Policy* (London: Allen & Unwin, 1986), pp. 155–179.

This content downloaded from 129.125.148.19 on Mon, 10 Dec 2018 14:54:55 UTC
All use subject to https://about.jstor.org/terms