Establishing the experimenting society: The historical origin of social experimentation according to the randomized controlled design

TRUDY DEHUE
University of Groningen

This article traces the historical origin of social experimentation. It highlights the central role of psychology in establishing the randomized controlled design and its quasi-experimental derivatives. The author investigates the differences in the 19th- and 20th-century meaning of the expression "social experiment." She rejects the image of neutrality of social experimentation, arguing that its 20th-century advocates promoted specific representations of cognitive competence and moral trustworthiness. More specifically, she demonstrates that the randomized controlled experiment and its quasi-experimental derivatives epitomize the values of efficiency and impersonality characteristic of the liberal variation of the 20th-century welfare state.

In 1969, experimental psychologist Donald T. Campbell published an article that would become a classic in the social sciences. The article bears the style of a political appeal. "The United States and other modern nations," Campbell proclaimed, "should be ready for an experimental approach to social reform, an approach in which we try out new programs designed to cure specific social problems" (Campbell, 1969, p. 409; for discussions of the article as a classic article, see Rossi & Freeman, 1985; Bulmer, 1986; Shadish, Cook, & Leviton, 1991; Brewer & Cook, 1996; Dunn, 1998).

In the history of natural science and psychology, the word *experimental* has been used loosely for various procedures (Kuhn, 1962; Hacking, 1983; Galison, 1997; Danziger, 1990; Winston, 1990; Winston & Blais, 1996). In Campbell's vocabulary, however, it had a distinct meaning: "True" experimentation implies that particular groups of people are subjected to a treatment and are compared with an untreated control group. Most importantly, it also means that the comparability of the groups is guaranteed because they are composed on the basis of chance.

AMERICAN JOURNAL OF PSYCHOLOGY
Summer 2001, Vol. 114, No. 2, pp. 283–302

Whenever possible, the demands of valid policy evaluation should prevail over humanitarian or practical objections to random assignment of people to treatments or to sustained monitoring during an experiment. Only when the true experiment is absolutely unfeasible for practical or moral reasons, another procedure may be followed. Then researchers must revert to a second-best quasi-experimental design. Campbell meticulously discussed several such substitute procedures with accompanying alternative statistical techniques and the diverse "threats to validity" to which they are susceptible.

"Reforms as Experiments" was not the first publication by Campbell arguing that "the logic of the laboratory" should be extended into society. With statistician Julian C. Stanley he published a long chapter called "Experimental and Quasi-Experimental Designs for Research on Teaching" (Campbell & Stanley, 1963). In 1966 this chapter was reprinted as a separate book under the briefer but much more inclusive title *Experimental and Quasi-Experimental Designs for Research* (Campbell & Stanley, 1966). Eight years later, the 100,000th copy of this book was sold. In 1990, 3,168 worldwide citations were counted to both the chapter and the book (Campbell, 1992), and in 1999 the book—albeit facetiously—was still called the "historical standard" and the "Old Testament" of social research (Swanborn, 1999, p. 9). In a series of articles Campbell argued also that policy making always *is* social experimentation and pictured the social scientist as "methodological servant of the experimenting society" (Campbell, 1973; for a bibliography until 1988, see Overman, 1988, pp. 525–537). Publications by his many pupils and associates further propagated the idea of experimental evaluation of regulatory attempts. Already during his life several books were brought out in his honor. After his death in 1996 many obituaries and another honorary volume (Dunn, 1998) celebrated his grand contributions to social science.

In the United States as well as other Western countries, the randomized controlled experiment became the "true experiment" in social experimentation, and numerous policies have been tested according to this scheme. Educational attempts were evaluated ranging from instructional movies for schoolchildren to campaigns for safe sex and job training for welfare recipients, and social actions were tried out varying from welfare allowances, electricity pricing schemes, felon rehabilitation projects, and punishments for spouse-beaters. Tremendous sums of money were involved in such experiments, and countless people were monitored.

The present article traces the historical origins of social experimentation and challenges some of its central assumptions. I do not intend to cast doubt on the integrity of the social scientists involved, nor do I

dispute their methodological competence. However, I reject the widespread belief that social experimentation is just a matter of applying a transcendental "logic of science" by neutral scientific specialists. My historical analysis demonstrates that the contemporary definition of a true experiment as a randomized controlled experiment did not originate in a natural science lab but in the closely intertwined realms of social administration and social research. More specifically, I argue that this influential representation of scientific experimentation epitomizes the values of efficiency and impersonality that are central in the liberal variation of the 20th-century welfare state. Furthermore, I challenge the neutral image of social experimentation by describing it as a special instance of the so-called tools-to-theories heuristic.

Social experimentation as a tools-to-theories heuristic

Historian of statistics and psychology Gerd Gigerenzer used the phrase "tools-to-theories heuristic" in pointing to the remarkable phenomenon that sometimes the tools for investigating people are transformed into models representing them (Gigerenzer & Murray, 1987; Gigerenzer, 1991, 1992). Stated differently, Gigerenzer demonstrated that sometimes supposedly neutral research tools become depictions of reality. His main example is that of statistical inference as a model of human cognition. By the 1960s, inference testing had become such an obvious instrument in psychology that it was elevated to a general model of human thinking. Psychologists began to consider all human cognition as the intuitive application of the rules of hypothesis testing.

Gigerenzer also noticed that this theory of the statistical mind is not a purely descriptive one but is inherently prescriptive as well. Cognitive psychologists presume that people intuitively try to follow the rules of inferential statistics but also demonstrated that people hardly ever do it properly. Thus the human mind is described as a spontaneous but failing statistician.

There is a striking similarity between the conception of the statistical mind and that of the experimenting society. In the first case an established research tool inspires a representation of human cognition, and in the latter the tool provides an image of social administration. Moreover, just like statistical inference as an exemplar of reasoning, scientific experimentation as a model of administration serves both descriptive and prescriptive purposes.

The idea of experimental government even is prescriptive in a double sense. Whereas reasoning along the lines of probabilistic statistics is considered an expression of cognitive capacities, it is crucial to the notion of social experimentation that adhering to the rules of science is also a token of trustworthiness. Campbell contrasted "experimental

administrators" to "trapped administrators." The first pave the way for randomized controlled trials (for instance by "allocating scarce resources by lottery" and "staged innovation"), whereas the second "have so committed themselves in advance to the efficacy of the reform that they cannot afford honest evaluation" (Campbell, 1969, p. 428). According to this view, adhering to the rules of science attests not only to intelligence, as in Gigerenzer's example, but also to morality.

The present history demonstrates that once the 20th-century liberal democrat values of efficiency and impersonality were embodied in the randomized controlled design, the design was promoted as the apogee of both rationality and reliability. For the sake of comparison, it begins in the 19th century.

Reading the thoughts of God

To be sure, 20th-century social scientists were not the first ones to merge the language of science and society. I. Bernard Cohen (1995) analyzed the natural science terminology in the writings of American society's founding fathers. Furthermore, Robert Brown (1997) discussed the use of the expression "social experimentation" in the 19th-century literature on science and society. Auguste Comte wrote on "social experiments" in relation to natural disturbances of social order such as avalanches or floods, and John Stuart Mill and George Cornewall Lewis also applied the expression to government actions.

However, Comte, Mill, and Lewis explicitly denied the possibility of deliberate experimentation with human beings for the sake of scientific research. Lewis expressed this view in the title of his book *Treatise on the Methods of Observation and Reasoning in Politics* (1852/1974). The volume itself amply explains why experimentation was not included. Scientific experimentation, Lewis argued, is "the physical mastery or manipulation of the object observed." This he considered inapplicable to humans because it would mean "destroying his life, or wounding his sensibility, or at least subjecting him to annoyance and restraint" (Lewis, 1852/1974, pp. 158–159, 161).

How to explain the difference? How to account for such prudence, extended even to the fear of "annoying" people, whereas in the next century group manipulation became an exemplar of sound thinking and trustworthiness? In his meticulous study of the 19th-century views, Brown (1997) considered it an incoherence that on one hand social actions were called social experiments whereas on the other hand scientific social experiments were deemed impossible or unwarrantable. However, I argue that 19th-century thinking on social experimentation seems confused only when judged from the standard of later beliefs on social relations and knowledge interests. Only by recognizing the sub-

stantial differences in the 19th- and 20th-century meaning of the expression "social experiment" can one gain understanding of both the earlier and the later beliefs.

Comte, Mill, and Lewis published at times when the strong tripartite relationship of politics, social amelioration, and social research characteristic of 20th-century welfare states was not yet established. In 19th-century Western societies, centralized government was much more limited. Direct social action was a responsibility primarily of local charity. Even John Stuart Mill, who impugned the power of aristocracy and devoted much of his writings to the subject of government responsibility, repeatedly stressed that authoritative interference should be limited to a very small range of human conduct (Kahan, 1992; Kurer, 1991).

Thus social science in the 19th century was not the sociotechnical affair it became in many 20th-century societies but predominantly a search for understanding of given social patterns. Generally speaking, 19th-century "social inquirers" (Haskell, 1977, p. 24) focused more on the conditions of societal coherence than those of change. Historians of statistics have described the gentlemen-statisticians, enlightened amateurs with a historical or political background who collected numbers in an attempt to read the thoughts of God. The 19th-century statistical techniques used for gathering knowledge to inform legislators (Kelman, 1987) were attuned mostly to the search for ideal types that were supposed to represent the given design of people and society. Quetelet's introduction of the average and normal distribution, in particular, aimed to give the true measure of things. It was used to examine population characteristics ranging from a male population's ideal chest girth to a community's fixed suicide or crime rate (Porter, 1986, 1994; Hacking, 1990; Desrosières, 1998).[1]

The task of social scientists and statisticians was to gain knowledge about society for the sake of prudent government rather than straightforward interference. To further elucidate the difference, social inquirers mainly looked at human communities in the detached but also respectful way entomologists looked at nests of ants. Thus we find in Lewis's (1852/1974) *Treatise on the Methods of Observation and Reasoning in Politics* considerate contemplations such as "whenever there is intelligence there is sensibility; and whenever there is sensibility, experiment as such, mere philosophical manipulation for the sake of truth, is inapplicable," followed shortly by cool reflections on famine and commercial crisis that have "an elective affinity with the rotten parts of the social fabric and try the strength of laws and institutions" (Lewis, 1852/1974, pp. 161–162, 172).

Like entomologists, these social scientists were fascinated by the natural social order, which they considered as delicate as important. This

explains why the 19th-century social experiment was merely a metaphor, borrowed from natural science, for describing events disturbing social order. The metaphor indicated that something can be learned about normal social life if events disrupting it are carefully observed. However, such disruptive social experiments were not something researchers should conduct: The expression "social experiment" did not yet refer to special methodological rules.

Changing definitions of government and experimentation

A different meaning of "social experimentation" rose together with changing views on government. The extreme poverty and general misery among the working people in turn-of-the-century America and in European industrialized countries gave rise to large-scale movements to extend centralized government. The rights of laborers were increasingly protected by minimum wage bills, child labor bills, and unemployment insurance. Social actions were introduced, from slum clearance projects to eugenic marriage laws. Increasingly and at a variety of government levels, administrative agencies were installed to initiate and regulate social changes.

Simultaneously, many felt that too much protection would mean discharging people of their individual accountability. If government help was to be offered, it should be directed at individual behavioral change. Moreover, many regarded interference in the free market as a hardly admissible intervention into private affairs. If private money was to be spent on public actions, indisputable benefits would have to justify the costs. Particularly in contexts where such liberal views on the welfare state dominated, hard proof was demanded of the efficacy of government interventions.[2]

Most importantly, in the present context, the proof requested was of a special kind. Suspicion about administrative inefficiency expressed itself in distrust of administrative beliefs. Whereas 19th-century government elites could still appeal to personal grace and discretion, the authority of 20th-century bureaucratic officials depended largely on adherence to impersonal standards, procedures, facts, and figures. In other words, formal rather than practical, theoretical, or substantial rationality was postulated (Weber, 1978; see also Kalberg, 1980).

Administrators who had to justify their decisions in impersonal terms increasingly appealed for help from the social sciences. As Theodore Porter cogently argued, the transition to mechanical objectivity in large parts of 20th-century social science was imposed by the needs of administrative officials who are easily accused of arbitrariness. Social scientists rapidly adapted to the new demands and began to focus on knowledge that was instrumental rather than reflexive, standardized rather than

discretionary. In social science, too, free reasoning became increasingly associated with unconstrained judgment and unconstrained judgment with arbitrariness and whim (Porter, 1995).

A close alliance was established with statistics that also adapted to the new demands. Mathematically trained statistical technicians replaced the former genteel amateurs. In the new statistics, the mean outcome of repeated measurements no longer represented a given ideal type. Deviations from the mean were studied as an indication of real population differences rather than errors. The change of focus from what binds people to what separates them induced the development of the tools of random sampling and calculating confidence intervals, of correlation and regression. In sum, a new kind of probabilistic state reasoning was worked out for calculating and controlling chance rather than regarding it as mere destiny (Porter, 1986; Hacking, 1990; Desrosières, 1998).

Economists, psychologists, political scientists, and sociologists began working on a network of methodological rules for banning "subjectivity." From the beginning of the 20th century, the majority of them preferred the role of an administrative advisor providing sheer methods and facts to that of a contemplative intellectual contributing interpretations and views. In these years, the predominant self-image of social scientists became that of mere technical servants to social administration (Bannister, 1987; Haskell, 1977; Lyons 1969; Ross, 1991).

The introduction of the randomized controlled design in psychology

With the increasing fear of arbitrariness, higher demands were made on what might constitute a scientific experiment. As a consequence, many social scientists concluded that in their field scientific experimentation is not practicable. Arguing that for true experimentation an unfeasible degree of control would be needed, economists embarked on the further development of nonexperimental techniques and mathematical models (Morgan, 1990). Sociologists held comparable views. F. Stuart Chapin, an American engineer and sociologist, argued that the state is exempted from the ban on experimentation with human beings but that, nevertheless, truly scientific experiments are impossible: "Fundamental differences in race, government, political ideals, and standard of living constitute the uncontrolled conditions which invalidate conclusions that may be drawn from much social experimentation" (Chapin, 1917, p. 244).

The solution of deliberately composing experimental and control groups was introduced in settings that enabled stringent control. Unlike other human sciences, psychology already had a tradition in active experimental manipulation of human subjects. From about the 1870s,

psychophysical researchers had developed elaborate skills in subjecting volunteers to the strictest methodological regimes (Danziger, 1990; Coon, 1993; Benschop & Draaisma, 2000). During the combined administrative turn of government and social science, some of these psychologists were quick to draw from their psychophysical background in developing procedures geared to the new requirements. Elaborating on their established designs, they were the first ones to devise experiments comparing deliberately composed experimental and control groups (Dehue, 1997, 2000).

Schoolchildren were the typical subjects of research in the early human experimental versus control group experiments. Whereas Edwin Boring found no articles reporting human controls in the 1916 volume of the *American Journal of Psychology*, Kurt Danziger found 14–18% in the 1914–1916 volumes of the *Journal of Educational Psychology* (Boring, 1954, p. 587; Danziger, 1990, pp. 113–115). According to Boring (1954, p. 588), children (and rats) were the standard subjects of early controlled experiments because children, like rats, were "inexpensive and plentiful." However, one might add that education as a means of creating self-supporting individuals was a pivotal concern of liberalism. From 1910, a forceful movement started in American schools for efficiency and scientific engineering (Callahan, 1962; Brown, 1992). And, most importantly, the school population—again like rats—was easy to handle. It was feasible to assign subjects to experimental or control groups and make them adhere to research protocols. The children's compliance was enforced by the teachers and the teachers' acquiescence by the powerful school management.[3]

Experiments were conducted to establish the results of large versus small classes, fresh versus ventilated air, the teachers' ways of teaching, their sex, or whatever educational variable the school management could think of. In the 1920s, the definition of a valid experiment was further refined. Methods were sought for excluding the possibility that effects had to be ascribed to some other difference between the groups than the educational measure to be tested. At first it became customary to handle the problem by subjecting children to preliminary tests on suspected factors and forming groups with equal test results. However, this matching procedure collided with the guiding principles of the liberal welfare state, that is, with the values of economy and impersonality. Matching was time-consuming and expensive, and choosing the factors to be suspected still depended on personal imagination (the sex of pupils might make a difference, or their ethnic background, or any other factor). At this point, educational psychologists invoked the methodological use of chance. The idea rose to cancel out unwanted variation by composing the groups at random.

In 1923, William A. McCall, an educational psychologist at Columbia University, published a manual titled *How to Experiment in Education* (McCall, 1923). The book flamboyantly exemplifies the dictates of efficiency and impersonality in psychology. In his introduction, McCall estimated that increased efficiency of education could save a full year of teaching per person and calculated that psychological experimentation would save "$134,680,000,000,000 for the next 100 generations of Americans." Then he advanced a way of making his profession still more cost-effective, which was to equate the groups by chance (McCall, 1923, pp. 41–42).

This was not the first time in history that randomization was used for experimental purposes. From the 1870s, psychophysical experimenters had randomized orders for thwarting the expectations of experimental subjects on the stimuli to come (Dehue, 1997). And from the early 1900s British economists and political scientists discussed the feasibility of drawing representative population samples (Stephan, 1948; Desrosières, 1998; see also note 1). McCall himself had already used random sampling to pick representative items to be included in a psychological test (McCall, 1922). Yet with his proposal of randomly composing experimental and control groups, randomization was methodologically deployed in a new and revolutionary way. Random allocation to groups epitomized the combined pursuit of efficiency and impersonality: It was an economical way to cancel out the individuality of both experimental researchers and their subjects.

Psychology joins the social sciences

Meanwhile, social scientists successfully extended their claims of expertise. Enduring problems of poverty, labor rebellion, drug abuse, illiteracy, and crime were tackled by administrators in cooperation with scientific experts. In America, a large endowment from the Rockefeller family further stimulated the involvement of social scientists with such issues. In the early 1920s, Beardsly Ruml, a psychology graduate from Chicago University, was appointed director of the Laura Spelman Rockefeller Memorial. Ruml was a fervent proponent of science-based administration. In a memorandum to the Rockefellers he urged that their money should not be spent on humanitarian organizations but on the development of "knowledge which in the hands of competent technicians may be expected in time to result in substantial social control" (quoted by Samelson, 1985, p. 39).

An alliance was established between the Memorial and the Social Science Research Council (SSRC), newly founded to enhance standardized social science by Chicago political scientist Charles E. Merriam. Psychology, as a means of social change through behavioral interven-

tions, was proclaimed the basic discipline of all social science. The authoritative Merriam in particular had high expectations for psychology. He "bombarded" his colleagues with texts on psychological methods and introduced psychologists into social science circuits (Ross, 1991, p. 455; see also Lyons, 1969; Samelson, 1985; Platt, 1996).

Psychologists became involved with administrative research projects outside their traditional domain of mental testing and education. One of them was Louis L. Thurstone, who had developed scales for measuring "attitudes" on particular races and nationalities, for instance, or the seriousness of crimes. Impressed by Thurstone's methodological capabilities, Merriam arranged a full professorship for him in the Chicago Social Science Research Building. In his new position as a social scientist, Thurstone supplemented his work on attitude scales with that on educational means of attitude change and evaluated the results. For instance, he handed out free tickets to schoolchildren for films on subjects such as bootlegging or war and told the children to validate their tickets by writing their names on them. This allowed him to compare the attitudes of children who had seen the movies with those of control groups who had not (Thurstone, 1952; Danziger, 1997).

Not all social scientists accepted the new experimental strategy. SSRC member F. Stuart Chapin, for one, kept to his former view that watertight social experiments cannot be done. After he learned of randomized controlled experiments, he rejected them, arguing that experimental allocation to social measures collides with the humanitarian mores of reform. Chapin made a career of developing procedures for comparing natural groups (Chapin, 1938, 1947, 1949–1950).

But Merriam had no such qualms. Whereas Thurstone experimented with small groups of children, Merriam initiated the first megaproject comparing adult experimental and control groups (Merriam & Gosnell, 1924; Gosnell, 1927). In the mid-1920s, his pupil Harold F. Gosnell investigated the effect of attempts to enhance the use of voting rights. Six thousand Chicago citizens were selected from districts with Blacks, Poles, Irish, Swedes, Germans, Russian Jews, Czechs, Italians, and native-born Whites from a "Gold Coast," a "depreciated residential area," and a "good residential neighborhood" (Gosnell, 1927, p. 14). In each district, random experimental and control groups were composed. The experimental groups received a written call to vote and information on the voting procedures in their own languages. With the help of city pollbooks that listed the names and addresses of those who voted, the experimenters concluded that the information had made a difference.

The launch of the paradigm of experimental and quasi-experimental designs

By the end of the 1920s, American social scientists became associated with the highest political levels and acquired social authority them-

selves. The equation of rationality with efficiency and reliability with impersonality enabled them to market their assistance as a testimony of capability and trustworthiness. Whereas originally formal rationality shaped the style of social research, now adherence to social science methodology became the ultimate proof of true impartiality.

A committee installed by President Hoover and chaired by Merriam published a two-volume report, *Recent Social Trends in the United States* (Merriam, 1933). The report vigorously propagated the view that all aspects of life should be organized according to the standards of science. In his concluding chapter, Merriam rhetorically asked, "What is to be the attitude of the great democracy toward the expert, the technician, the scientist in determining the course and speed of the ship of state?" and prophetically answered, "Recent social trends in America and elsewhere drive us inexorably toward a still closer relationship between education, science and government" (Merriam, 1933, pp. 1500–1501).

Indeed, in the 1930s administrative social science was booming. More than any president before him, Franklin D. Roosevelt attracted techno-scientific support. He surrounded himself with social scientists as his personal advisors, of whom Merriam even became his "uncle Charley" (Ross, 1991, p. 455). By the end of the decade some 8,000 social scientists were working for the American federal government, and when the United States became involved in World War II, another 8,000 were added. The army became an important field for the application and further development of social scientific research methods (Lyons, 1969, p. 83).

Samuel Stouffer, a former student of Merriam and Thurstone, was appointed head of the army's Morale Division, charged with investigating the soldiers' attitudes toward the war. At first the division applied itself to only measuring these attitudes, but it soon was charged with improving morale and evaluating the results of these educational attempts. The film series *Why We Fight* was made, and the Morale Division's experimental section, staffed mainly by psychologists, designed experiments for assessing the impact of the movies (Hovland, Lumsdaine, & Sheffield, 1949).

In the meantime, American social researchers had adopted the technique of significance testing as introduced by British biometrician Ronald A. Fisher. As Fisher had repeatedly stressed, random allocation to groups was a condition of the valid application of his technique (Fisher, 1935). In this way, a forceful extra argument was added for random allocation. However, during the experiments, the soldiers became suspicious when some were summoned to view a movie whereas others were not. Random assignment appeared to elicit a new source of bias. Although upon pretesting the units turned out to differ in many consequential respects, it was decided that existing army units should consti-

tute the experimental versus control groups. And that was only one of many methodological trespasses. After the war, four volumes were published of *Studies in Social Psychology in World War II,* of which the third book dealt with the activities of the experimental section (Hovland et al., 1949). The latter book mostly communicated the view that much more methodological expertise is needed before solid statements can be made.

Nevertheless, after the war the members of the army experimental section acquired good positions (Buck, 1985; Herman, 1995). Stouffer became head of the new Laboratory for Social Relations at Harvard University. In the *Studies in Social Psychology* and in several articles, he drew up the methodological balances of his wartime experiences. Inferior methodological designs, he argued, are like poor defense lines: "There is all too often a wide-open gate through which uncontrolled variables can march" (Stouffer, 1950, p. 357). Stouffer's strategy was to improve the defense rather than quit the fight. To him the difficulties of social experimentation confirmed the need for enhanced methodological expertise. He carefully scrutinized the kinds of bias caused by various deviations from the randomized controlled design.

In the 1950s, new methodological treatises acknowledged that it is not always possible to assign people at will to experimental conditions. Series of alternative procedures were analyzed carefully. To continue Stouffer's metaphor, these methodological texts read as catalogs of an ever more sophisticated armory against a never completely arrestable invasion. Among the first methodologists to write in this vein was young psychologist Donald T. Campbell, who also had started his career as an army attitude and propaganda researcher (Campbell, 1981). In 1957, Campbell established his reputation as a keen methodologist with an astute article on a variety of second-best designs and accompanying factors that may invalidate the results (Campbell, 1957). With this work he began his career as a champion of social experimentation according to the methodology of experimental and quasi-experimental designs described earlier in this article.

The expansion of program evaluation

In the 1960s, President Johnson initiated his War on Poverty and the nationwide Planning, Programming, Budgeting System for finding the most effective and least costly alternatives in achieving social progress. At (again) a much larger scale than ever before, social scientists were employed as technical experts (Williams, 1971). Gigantic experiments with welfare allowances were initiated to find out whether financial aid reduced people's motivation to work. Some 1,500 poor families were recruited for the notorious New Jersey Negative Income Tax Experi-

ment. This sample was randomly divided into a control group that did not receive an allowance and experimental groups that differed as to the combination of the amount received and the income level at which help was stopped. All groups were regularly monitored throughout the 4-year term of the experiment. Many comparable experiments followed. The total costs of 10 such experiments amounted to $1.1 billion, of which $450 million was spent on research and administration. The largest project was the Seattle–Denver Income Maintenance Experiment, in which nearly 5,000 families were involved (more figures are available in Bulmer, 1986; Haveman, 1987).

Although in a sense people on welfare are as dependent on authorities as hospitalized patients, schoolchildren, or soldiers, keeping experimental welfare recipients under control was not feasible. The researchers had to deal with sources of bias such as selective agreement to participate, dropping out of annoyed families (in the control group particularly), incorrect reports of family income, and subjects moving to other states. After the experiments, special conferences were organized and books were published to scrutinize all aspects of the experimental design, the implementation of the experiments, and the reported results (Kershaw & Fair, 1976; Watts & Rees, 1976; Rossi & Lyall, 1976).

More than a few social science methodologists concluded that experimentation was not the best track to follow in the field of program evaluation (as it was now called). In taking issue with the experimental paradigm, renowned psychologist and psychometrician Lee Cronbach went as far as to argue that "we should blur lines that separate 'values' from 'facts,' 'humanities' from 'sciences,' and 'quantitative' from 'qualitative,' or 'applied' from 'basic' research" and to advance "architecture, music, and philosophy" as the exemplary disciplines to social science (Cronbach, 1987, p. 421).

Albeit with increasing caution, Campbell kept on defending the idea of an experimental society (for a thorough analysis of Campbell's thinking through the years, see Shadish et al., 1991).

The handbook *Quasi-Experimentation* by Cook and Campbell (1979) increased the number of "threats to validity" from 12 to 33, which motivated many social scientists to keep as closely as possible to the true design. In the 1970s and 1980s, the Ford Foundation supported randomized controlled experiments with 65,000 welfare recipients in 20 American states (Gueron & Pauly, 1991). To many, such experiments still are "the Rolls Royces or Cadillacs of evaluative research design both because of their superior inferential power and because of their glamour" (Bulmer, 1986, p. 169; see also Boruch, 1997; Orr, 1999). Indeed, to the present day randomized controlled experiments are widely conducted,

including projects such as the one currently running in the Netherlands in which heroin addicts receive their drug for free and are compared with a control group getting only methadone (Dehue, in preparation).

Conclusion: Making policy while testing policy

The present historical reconstruction demonstrates that the most prominent paradigm of experimentation in social science and psychology was not derived from a transcendental logic of science, nor does it stem from any research lab. Whereas originally the word *experiment* was introduced into social thought as a natural science metaphor, its central meaning in social science exemplifies the aspiration of ruling by technique rather than tradition, of replacing the individuality of both the governors and the governed by impersonality.

Moreover, proponents of the paradigm of experimental and quasi-experimental research marketed adherence to its procedures as a proof of overall good thinking and moral creditability. They appealed to the conscience of policy makers and tried to make them comply with the rules of experimentation. In that sense, one might say that the randomized controlled experiment was not just a tool-to-theory but also a tool-to-practice heuristic. Whereas cognitive psychologists in Gigerenzer's example did not try to change people's strategies for drawing conclusions, adherents of social experimentation actively prodded society to become experimental. In the words of Campbell associates Riecken and Boruch (1974), their arguments were "frankly aimed at influencing the thinking of policy developers and decision makers in Federal, State, and local governments."

The tools of scientific policy evaluation are not neutral but recreate society in their image. In 1974, the American General Accounting Office established a special department for professional evaluation research, which was even charged with evaluating evaluations (Dekker & Leeuw, 1989). Governments in other countries followed shortly, as did the boards of other institutions such as hospitals, universities, and businesses. Moreover, small and large companies specializing in social program evaluation mushroomed in the private sector.

And it is not just administrators who are expected to adhere to the mores of research. This holds even more so for the people voluntarily or involuntarily participating in the experiments. Experimental subjects watch prescribed television programs, undergo particular educational training, are put on diets, live in experimental houses, and take prescribed heroin or methadone. For measuring the effects of such treatments, the subjects' freedom is limited to keep them from "contaminating" influences, and they are observed or interrogated in standardized ways. If subjects can give their opinions freely, they are recorded according to prestructured schemes.

Most importantly, in the design of every experiment countless nonmechanical decisions must be made. In the experiment with free heroin, currently running in the Netherlands, definitions are needed for diagnosing the condition and the behavior of the subjects. Have they improved significantly when their bodily weight is back to normal? If so, how normal is normal enough? Can it justly be claimed that distribution of free heroin diminishes social malfunctioning if addicts no longer mug fellow citizens? And is a change from mugging with violence to nonviolent shoplifting a slight improvement? Experiments, even those with the true design, can never be neutral, if only because they demand classifications based on conventions and agreements.

The benefits of mechanical objectivity should not be underestimated. Protocols offer protection against individual whim and unequal treatment. The obligation to ground decisions on facts and procedures has often served as an antidote to wishful thinking and grinding injustice. Few people would like to be placed back in the 19th century, when capriciousness and exploitation reigned. However, questioning current solutions does not entail a plea to return to former abuses. As history has taught, complete exclusion of personal discretion and sheer reliance on rules may cause more damage than changes for the good. The words *impersonal* and *formal* retain negative connotations, indicating that injustice may be done when complexity and variety are neglected. Life becomes gray when standardization reigns.

It is beyond the scope of this article to embark upon the complicated discussions about when precisely trust should be put on persons rather than procedures. However, it is within its scope to counter untenable claims of neutrality and thus help recreate space for theory, reflection, creativity, and speculation and briefly for intellectual social research. Without denying the many benefits of mechanical objectivity, relating it to the ideals of sociotechnical administration reminds one of its limited significance as a general definition of rationality. In both science and society, objectivity also can be a matter of expertise based on individual reflection and mutual consultation; persons may be more important than procedures.

Notes

I thank Yvette Bartholomée, Anne Beaulieu, Harro Maas, Ted Porter, Ad Prins, and Nico Randeraad for their helpful comments on a previous version. Correspondence about this article should be addressed to Trudy Dehue, Heymans Institute for Psychological Research, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands (e-mail g.c.g.dehue@ppsw.rug.nl). Received for publication May 15, 2000; revision received July 8, 2000.

1. In a very illuminating chapter, Desrosières (1998) discussed the question of how 19th-century (and early 20th-century) social scientists could attach value to data acquired from nonrepresentative samples. Desrosières arrived at conclusions similar to the ones presented in this article. He argued that it is wrong to stigmatize the early methods as biased. Surveys for which subjects were selected on the basis of familiarity rather than randomness were conducted for goals other than those of 20th-century research. The investigators' aim was to describe situations, such as the unrest or poverty in workers' communities. As Desrosières expresses it, "The intent was not yet to take measurements in order to prepare for the measures to be taken, as would be the case when the welfare state developed. It was to assemble elements capable of grounding the characters in the story to be told or organized, through, in particular, typological works: Classification, which created collective actors, was one of the products of these surveys" (Desrosières, 1998, p. 213).

2. Esping-Andersen (1990) contrasted three welfare state regimes. The liberal variation encourages market efficiency, emphasizes work ethic norms, and offers only minimum benefits. The United States, Canada, and Australia offer the archetypes of this model. In the other variations, the corporatist welfare state (Germany, Italy) and the social democratic welfare states (Scandinavian countries), granting social rights is a less contested issue.

3. The typical subjects of early medical comparative experiments with humans also lived in restricted conditions. In his book on the origin of the randomized controlled design in American medicine, Harry Marks (1997) reported on early experimental and control groups of patients, soldiers, or prisoners. In medicine, however, the design did not begin to find acceptance until the 1950s, and only after the authority of statisticians was called in. The call for comparative experimentation, which came also primarily from government institutions, met with much resistance. Practicing physicians appealed to their time-honored clinical discretion and held that assigning diseased people to control groups or trying out treatments on healthy people collides with medical ethics (Marks, 1997). Of course, doctors had much more social power than the schoolchildren and the humble, mostly female teachers who were enrolled for educational control group experiments in the early 20th century.

References

Bannister, R. C. (1987). *Sociology and scientism: The American quest for objectivity, 1880–1940.* Chapel Hill: University of North Carolina Press.

Benschop, R. J., & Draaisma, D. (2000). In pursuit of precision. The calibration of minds and machines in late 19th-century psychology. *Annals of Science, 57,* 1–25.

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

Boruch, R. (1997). *Randomized experiments for planning and evaluation.* London: Sage.

Brewer, M. B., & Cook, T. D. (1996). Donald T. Campbell (1916–1996). *American Psychologist, 52,* 267–268.

Brown, J. (1992). *The definition of a profession: The authority of metaphor in the history of psychology.* Princeton, NJ: Princeton University Press.

Brown, R. (1997). Artificial experiments on society: Comte, G. C. Lewis and Mill. *Journal of Historical Sociology, 10,* 74–97.

Buck, P. (1985). Adjusting to military life: The social sciences go to war, 1941–1950. In M. R. Smith (Ed.), *Military enterprise and technological change: Perspectives on the American experience* (pp. 203–253). Cambridge, MA: MIT Press.

Bulmer, M. (1986). Evaluation research and social experimentation. In M. Bulmer, K. G. Banting, S. S. Blume, M. Carley, & C. H. Weiss (Eds.), *Social science and social policy* (pp. 155–179). London: Allen & Unwin.

Callahan, R. E. (1962). *Education and the cult of efficiency.* Chicago: University of Chicago Press.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54,* 297–312.

Campbell, D. T. (1969). Reforms as experiments. *American Psychologist, 24,* 409–429.

Campbell, D. T. (1973). The social scientist as methodological servant of the experimenting society. *Policy Studies Journal, 2,* 72–75.

Campbell, D. T. (1981). Comment: Another perspective on a scholarly career. In M. B. Brewer & B. E. Collins (Eds.), *Scientific inquiry and the social sciences: A volume in honor of D. T. Campbell* (pp. 454–486). San Francisco: Jossey-Bass.

Campbell, D. T. (1992). *Stanley's contributions to measurement and experimental design.* Unpublished manuscript.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching.* Chicago: Rand McNally.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

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

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

Chapin, F. S. (1947). *Experimental designs in social research.* New York: Harper.

Chapin, F. S. (1949–1950). Experimental designs in social research. *American Journal of Sociology, 55,* 401–403.

Cohen, I. B. (1995). *Science and the founding fathers: Science in the political thought of Thomas Jefferson, Benjamin Franklin, John Adams, and James Madison.* New York: Norton.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation. Design and analysis issues for field settings.* Chicago: Rand McNally.

Coon, D. J. (1993). Standardizing the subject: Experimental psychologists, introspection, and the quest for a technoscientific ideal. *Technology and Culture, 34,* 757–783.

Cronbach, L. J. (1987). Social inquiry by and for earthlings. In W. R. Shadish & C. S. Reichardt (Eds.), *Evaluation studies* (Vol. 12, pp. 40–425). London: Sage.

Danziger, K. (1990). *Constructing the subject.* New York: Cambridge University Press.

Danziger, K. (1997). *Naming the mind. How psychology found its language.* London: Sage.

Dehue, T. (1997). Deception, efficiency, and random groups. *Isis, 88,* 653–673.

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

Dehue, T. (in preparation). *A Dutch Treat. Social experimentation and the case of heroin maintenance in the Netherlands.* Groningen, The Netherlands: University of Groningen.

Dekker, P. J., & Leeuw, F. L. (1989). *Beleids- en programma-evaluaties* [Policy and program evaluations]. Leiden, The Netherlands: DSWO Press.

Desrosières, A. (1998). *The politics of large numbers. A history of statistical reasoning.* Cambridge, MA: Harvard University Press.

Dunn, W. N. (1998). Campbell's experimenting society: Prospect and retrospect. In W. N. Dunn (Ed.), *The experimenting society: Essays in honor of Donald T. Campbell* (pp. 20–21). New Brunswick, NJ: Transaction Publishers.

Esping-Andersen, G. (1990). *The three worlds of welfare capitalism.* Cambridge, UK: Polity Press.

Fisher, R. A. (1935). *The design of experiments.* Edinburgh: Oliver & Boyd.

Galison, P. L. (1997). *Image and logic: A material culture of microphysics.* Chicago: University of Chicago Press.

Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review, 98,* 254–267.

Gigerenzer, G. (1992). A discovery in cognitive psychology: New tools inspire new theories. *Science in Context, 5,* 329–350.

Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics.* London: Erlbaum.

Gosnell, H. F. (1927). *Getting out the vote: An experiment in the stimulation of voting.* Chicago: University of Chicago Press.

Gueron, J., & Pauly, E. (1991). *From welfare to work.* New York: Russell Sage Foundation.

Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of natural science.* New York: Cambridge University Press.

Hacking, I. (1990). *The taming of chance.* New York: Cambridge University Press.

Haskell, T. (1977). *The emergence of professional social science: The American Social Science Association and the 19th-century crisis of authority.* Urbana: University of Illinois Press.

Haveman, R. H. (1987). Social experimentation and "social experimentation." In W. R. Shadish & C. S. Reichardt (Eds.), *Evaluation studies* (Vol. 12, pp. 608–627). London: Sage.

Herman, E. (1995). *The romance of American psychology: Political culture in the age of experts.* Berkeley: University of California Press.

Hovland, C. I., Lumsdaine, A. A., & Sheffield, F. E. (1949). *Experiments on mass communication (Studies in social psychology in World War II, 3).* Princeton, NJ: Princeton University Press.

Kahan, A. S. (1992). *Aristocratic liberalism.* New York: Oxford University Press.

Kalberg, S. (1980). Max Weber's types of rationality: Cornerstones for the analysis of rationalization processes in history. *American Journal of Sociology, 85,* 1145–1179.

Kelman, S. (1987). The political foundations of American statistical policy. In W. Alonso & P. Starr (Eds.), *The politics of numbers* (pp. 275–303). New York: Russell Sage.

Kershaw, D., & Fair, J. (1976). *The New Jersey income-maintenance experiment* (Vol. 1). New York: Academic Press.

Kuhn, T. S. (1962). *The structure of scientific revolutions.* Chicago: University of Chicago Press.

Kurer, O. (1991). *John Stuart Mill. The politics of progress.* New York: Garland.

Lewis, C. G. (1974) *A treatise on the methods of observation and reasoning in politics.* New York: Arno. (Original work published 1852)

Lyons, G. M. (1969). *The uneasy partnership: Social science and the federal government in the 20th century.* New York: Russell Sage Foundation.

Marks, H. M. (1997). *The progress of experiment. Science and therapeutic reform in the United States, 1900–1990.* New York: Cambridge University Press.

McCall, W. A. (1922). *How to measure in education.* New York: Macmillan.

McCall, W. A. (1923). *How to experiment in education.* New York: Macmillan.

Merriam, C. E., & Gosnell, H. F. (1924). *Non-voting: Causes and methods of control.* Chicago: University of Chicago Press.

Merriam, C. E. (1933). Government and society. In President's Research Committee on Social Trends, *Recent social trends in the United States* (Vol. 1, pp. 1489–1543). New York: McGraw-Hill.

Morgan, M. (1990). *The history of econometric ideas.* Cambridge, UK: Cambridge University Press.

Orr, L. L. (1999). *Social experiments. Evaluating public programs with experimental methods.* London: Sage.

Overman, E. S. (Ed.). (1988). *Methodology and epistemology: Selected papers of D. T. Campbell.* Chicago: Chicago University Press.

Platt, J. (1996). *A history of sociological research methods in America, 1920–1960.* Cambridge, UK: Cambridge University Press.

Porter, T. M. (1986). *The rise of statistical thinking, 1820–1900.* Princeton, NJ: Princeton University Press.

Porter, T. M. (1994). From Quetelet to Maxwell: Social statistics and the origins of statistical physics. In I. B. Cohen (Ed.), *The natural sciences and the social sciences* (pp. 345–363). Dordrecht, The Netherlands: Kluwer.

Porter, T. M. (1995). *Trust in numbers: The pursuit of objectivity in science and public life.* Princeton, NJ: Princeton University Press.

Riecken, H. R., & Boruch, R. F. (Eds.). (1974). *Social experimentation: A method for planning and evaluating social intervention.* New York: Academic Press.

Ross, D. (1991). *The origins of American social science.* New York: Cambridge University Press.

Rossi, P. H., & Freeman, H. E. (1985). *Evaluation: A systematic approach.* London: Sage.

Rossi, P. H., & Lyall, K. C. (1976). *Reforming public welfare: A critique of the negative income tax experiment.* New York: Russell Sage Foundation.

Samelson, F. (1985). Organizing for the kingdom of behavior: Academic battles and organizational policies in the twenties. *Journal of the History of the Behavioral Sciences, 21,* 33–47.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). Donald T. Campbell: Methodologist of the experimenting society. In W. R. Shadish, T. D. Cook, & L. C. Leviton (Eds.), *Foundations of program evaluation* (pp. 73–119). London: Sage.

Stephan, F. S. (1948). History of the uses of modern sampling procedures. *Journal of the American Statistical Association, 43,* 12–39.

Stouffer, S. A. (1950). Some observations on study design. *American Journal of Sociology, 55,* 355–361.

Swanborn, P. G. (1999). *Evalueren* [Evaluating]. Amsterdam: Boom.

Thurstone, L. L. (1952). Autobiography. In E. G. Boring, H. S. Langfeld, H. Werner, & R. M. Yerkes (Eds.), *A history of psychology in autobiography* (Vol. 4, pp. 295–321). Englewood Cliffs, NJ: Prentice Hall.

Watts, H., & Rees, H. (Eds.) (1976). *The New Jersey income-maintenance experiment* (Vols. 1 and 2). New York: Academic Press.

Weber, M. (1978). *Economy and society* (Vols. 1 and 2). Berkeley: University of California Press.

Williams, W. (1971). *Social policy research and analysis: The experience in the federal agencies.* New York: Elsevier.

Winston, A. S. (1990). Robert Sessions Woodworth and the "Columbia Bible": How the psychological experiment was redefined. *American Journal of Psychology, 103,* 391–401.

Winston, A. S., & Blais, D. J. (1996). What counts as an experiment?: A transdisciplinary analysis of textbooks, 1930–1970. *American Journal of Psychology, 109*(4), 599–616.