The Principles of Experimental Design and Their Application in Sociology

Michelle Jackson[1,2] and D.R. Cox[2]

[1]Institute for Research in the Social Sciences, Stanford University, Stanford, California 94305; email: mvjsoc@stanford.edu

[2]Nuffield College, University of Oxford, Oxford OX1 1NF, United Kingdom; email: david.cox@nuffield.ox.ac.uk

Annu. Rev. Sociol. 2013. 39:27–49

First published online as a Review in Advance on April 5, 2013

The *Annual Review of Sociology* is online at http://soc.annualreviews.org

This article's doi: 10.1146/annurev-soc-071811-145443

Keywords

principles of design, laboratory experiments, field experiments, survey experiments, visibility of experimental design, history of experiments in the social sciences

Abstract

In light of an increasing interest in experimental work, we provide a review of some of the general issues involved in the design of experiments and illustrate their relevance to sociology and to other areas of social science of interest to sociologists. We provide both an introduction to the principles of experimental design and examples of influential applications of design for different types of social science research. Our aim is twofold: to provide a foundation in the principles of design that may be useful to those planning experiments and to provide a critical overview of the range of applications of experimental design across the social sciences.

INTRODUCTION

The social sciences are largely observational, characterized by the application of nonexperimental methods such as social surveys, interviewing, or direct observation. But many social science questions may also be addressed using experimental methods, and indeed, some may be best approached this way. Particularly appropriate for experimental investigation are questions that aim to establish causal relationships between two or more variables.

Concepts needed for the design of successful experiments have been discussed for many years, mainly in the context of the natural sciences, initially in physics, and later especially in such applied subjects as agriculture, industrial technology, and clinical trials. Over the past 80 years, the discussion of experimental design in contexts where there is appreciable haphazard (that is, irregular and nonpredictable) variation has become a major chapter in statistical theory and application. In this article, we review the relevance of such ideas for the social sciences taken in a broad sense. We draw our illustrations largely from sociology and from sociologically relevant work in experimental economics and political science. Our focus is therefore on those social science disciplines that have been primarily observational since their inception; on the whole, we do not discuss fields that have always been dominated by experimental research, such as certain subdisciplines of psychology that might be classified under the auspices of social science (subdisciplines that include studies of personality, cognition, learning, and particularly experimental psychology).

The organization of the review is as follows: We begin by discussing the early use of experiments in the social sciences before presenting an empirical analysis showing a recent increase in their use. Next, we summarize the main statistical ideas involved in the design of experiments, ideas that are an important foundation for the types of application that we discuss in the remainder of the article (laboratory, field, and survey). We conclude with some general comments. Our aim is twofold: to provide a review of the principles of design that may be useful to those planning experiments and to provide an overview of the range of applications of experimental design in sociology and other social sciences. We review neither the many and diverse findings of social science experiments nor the huge swaths of social science literature that stem from experimental work. Rather, our focus is on the design of experiments as typically understood in the statistical literature, and we review such design in the context of applications within social science.

An initial difficulty in a broad-ranging discussion is that terminology varies between specific application fields as well as between these and the general statistical literature, with disagreement even on the question of what constitutes an experiment. We deal here with investigations in which the effects of a number of alternative conditions or treatments are to be compared. Broadly, the investigation is an experiment if the investigator controls the allocation of treatments to the individuals in the study and the other main features of the work, whereas it is observational if, in particular, the allocation of treatments has already been determined by some process outside the investigator's control and detailed knowledge. The allocation of treatments to individuals is commonly labeled manipulation in the social science context.

SOME HISTORY OF EXPERIMENTAL DESIGN IN THE SOCIAL SCIENCES

Although most social science disciplines developed during a period in which experimental design was well established in the statistical and scientific literature, such designs have a checkered history in the social sciences. We briefly review this history and highlight some of the design aspects of studies that are now celebrated as classics in the field.

As mentioned above, some social sciences have always had essentially experimental components. In social psychology, for example,

experimental designs have been employed throughout the history of the discipline.[1] Indeed, the first empirical paper published within the field of social psychology included an experimental study of the effect of competition on individual task performance. In an article in the *American Journal of Psychology*, Triplett (1898) reported the results from a laboratory experiment designed to shed light on the empirical finding that racing cyclists produced faster times in races that were paced or in competition with other riders than in unpaced races, one example of a more general phenomenon later labeled "social facilitation" (Allport 1920, Aiello & Douthitt 2001). In Triplett's experiment, two groups of 20 children were asked to complete a simple physical task—winding fishing line on a reel—as quickly as possible, either alone or in the presence of competition. The experimental design is therefore relatively simple, with manipulation of a single treatment (competition, absent or present) and focus on a single outcome variable (speed of task completion). Each child completed the task six times, with a different ordering of the trials (alone or in competition) for the two groups of children. The article does not describe the process through which individuals were allocated to groups; the systematic use of randomization had not been suggested until much later. After a comparison of task completion times between the two conditions, Triplett (1898, p. 533) concludes, "From the. . .laboratory races we infer that the bodily presence of another contestant participating simultaneously in the race serves to liberate latent energy not ordinarily available."

Although the beginning of any discipline is difficult to define with precision, Strube (2005, p. 281) writes that "[t]he publication of Triplett's experiment in 1898 was a watershed event in the history of social psychology. It marked the formal beginning of the field as an empirical enterprise and launched one of its most enduring and far-reaching research lines" (see Allport 1954 and Brehm et al. 1999 for similar plaudits). Experimental designs have subsequently dominated the field of social psychology, with other notable studies including Asch's (1951, 1956) tests of conformity, Milgram's (1974) tests of obedience to authority, Bandura et al.'s (1961) investigation into aggression as social learning, and Zimbardo's field experiment testing the effect of externally imposed social roles on individual behavior, better known as the Stanford Prison experiment (Haney et al. 1973; see also **http://www.prisonexp.org**).

By contrast, other areas of social science have been relatively slow to embrace experimental design as a central research method. In the first issue of *American Sociological Review*, Chapin (1936, pp. 7–8) found it necessary, in his presidential address, "to resolve a confusion in thought that is present in many discussions of the procedures of measurement and of experiment. The opinion that experiment cannot be used in social research is an erroneous judgment." The prevailing view criticized by Chapin that the social sciences are fundamentally nonexperimental appears to have also been present in political science, and the president of the American Political Science Association claimed that "[w]e are limited by the impossibility of experiment. Politics is an observational, not an experimental science" (Lowell 1910, p. 7, quoted in Druckman et al. 2006, p. 627). And Alfred Marshall (1890), in *Principles of Economics*, wrote that "[t]he natural sciences and especially the physical group of them have this great advantage as a discipline over all studies of man's action, that in them the investigator is called on for exact conclusions which can be verified by subsequent observation or experiment" (ch. 4, §5).

Despite such epistemological statements about the essential nature of the social sciences, even the very early histories of economics, political science, sociology, and other social

[1]At its inception, social psychology was an interdisciplinary enterprise. During the twentieth century, the subject of social psychology came to be studied in both psychology and sociology departments, with differences in focus related to the umbrella discipline. In this review we focus on social psychology as practiced within sociology.

science disciplines are in fact not lacking for examples of highly influential experimental studies. Perhaps the most well-known early sociological experiments were carried out at Western Electric's Hawthorne Works in the 1920s and 1930s and reported in *Management and the Worker* (Roethlisberger & Dickson 1939; see also Mayo 1933). The aim of the experiments was to evaluate the effect on workers' productivity of a range of stimuli, including the level of light in the factory, the length and timing of breaks, the provision of food, and other environmental factors. The illumination experiments became the most notorious of the studies as a result of a curious finding, later labeled the "Hawthorne effect."

The procedure in the Hawthorne illumination studies was as follows: Over a period of almost three years, the productivity of workers in three departments of the factory was measured through a simple count of the number of units (parts inspected, telephone relays assembled, wire coils wound) produced by each worker on each day. In the first stage of the experiment, all three groups of workers were subject to the experimental conditions, whereas in later stages a small group of workers was assigned to work in a special room, where more extreme manipulations could occur. Manipulation of lighting conditions allowed the researchers to test the hypothesis that better lighting conditions would lead to increased productivity. Dickson & Roethlisberger (1966, p. 20) later summarized the results of the experiments:

> [I]t was difficult to conclude just what the effect of illumination on productivity was. Some of the results of these experiments seemed to say that it was positive, some that it was zero, and some that it was screwy. For example, sometimes an improvement in illumination was accompanied by an improvement in output and then again, sometimes a positive increase in one did not produce an appreciable increase in the other. And then—this is the screwy part—quite often when the illumination intensity was decreased, output did not go down but remained the same and in some cases increased.

The finding that productivity increased even when the level of illumination was extremely poor (indeed, at times the illumination was so low as to be labeled moonlight) led the researchers to suspect that another factor was at work. From further experiments and observations, they argued that a psychological mechanism was operating to increase productivity: The workers were apparently responding positively to the increased attention to their productivity by researchers, and not to the changes in illumination. Note that Dickson & Roethlisberger's summary of the findings of the Hawthorne studies is rather more cautious than some of the later summaries and interpretations published by other social scientists [see Jones (1992, p. 453, footnote 2) for a discussion of this point].

The Hawthorne studies showed that interpersonal processes in the workplace could affect employee productivity. But the methodological contributions of the research were perhaps even more important, contributions that came about primarily because the initial experimental designs were defective and lacked sufficient control and because observer effects that should have been avoided through careful design were instead entangled with the effects of interest. The authors were commendably open about the challenges posed both by the environment and by experimental findings that were unexpected and unpredicted. It is not often appreciated that the Hawthorne studies encompassed several quite distinct designs, with substantial changes to design and procedure occurring as a result of unexpected findings from early experiments when the researchers discovered that a number of uncontrolled variables could have influenced the results (Roethlisberger & Dickson 1939, p. 19). One consequence of these changes to the experimental design was that there was insufficient independent replication across the experiments, and in retrospect a different design would have been a better starting point (a factorial design would have been most appropriate). But the recognition that the researcher may serve as a confounding factor when

experiments are conducted on human subjects was profoundly influential (Levitt & List 2011; see also Adair et al. 1989 for a meta-analysis of studies testing for the presence of a Hawthorne effect in educational research). Indeed, the studies were so influential that the findings have been subjected to scrutiny since the 1930s, and reanalyses of the Hawthorne data are still being published (e.g., Franke & Kaul 1978, Gillespie 1991, Jones 1992, Levitt & List 2011). Many of these reanalyses find that the original claims for the existence of a Hawthorne effect are not supported or are only weakly supported by the experimental data. Of course, the methods of statistical analysis available to those undertaking reanalyses are better than those available to the Hawthorne researchers.

The Hawthorne studies thus have a special place in the history of experimentation in the social sciences, and no other sociological experiment has achieved the same level of renown. We do not consider additional historical examples of experimental social science studies here, but for discussion of the history of experiments in sociology, see Oakley (1998); for a history of experiments in political science, see Green & Gerber (2003), and in economics, Roth (1993, 1995).

THE INCREASED USE OF EXPERIMENTAL DESIGN IN THE SOCIAL SCIENCES

Use of experiments in the social sciences has increased appreciably in recent years. Social science disciplines that have traditionally eschewed experimental designs have witnessed an increase in experimental research, and universities and research institutions have made infrastructural investments to support researchers collecting data using experiments.

One way to gauge the growth in experimental designs is to examine the publication frequency of research employing them. As an example, we measured the visibility of experimental design within economics, political science, and sociology through a simple content analysis of articles published in the top-ranked journals for each discipline between 1990 and 2010. We calculated the proportion of research articles that report results from experimental designs in the top-ranked journals for economics (*American Economic Review*, *Econometrica*), political science (*American Journal of Political Science*, *American Political Science Review*), and sociology (*American Journal of Sociology*, *American Sociological Review*). The top-ranked journals are defined as those with the largest number of total citations from 2008–2009, as reported in Thomson Reuters's Web of Knowledge Journal Citation Report®. We include only full research articles; research notes, replies, corrections, and similar pieces are excluded, leaving a total sample of 5,874 articles. An article was coded as employing experimental design if empirical results were reported that originated from the manipulation of human subjects. Articles that described a formal model of behavior with a hypothetical experimental test were not coded as experimental.

Figure 1 shows a three-year moving average of the proportion of all articles published in each of the top-ranked journals that employed experimental design (original journal by year coding is available from the authors on request). The figure demonstrates a clear growth over time in the number of experimental research articles. Experimental designs were present in top-ranked journals at the beginning of the period, but they represented only about 3% of the total number of research articles in the six journals. A rapid growth in the visibility of experimental designs occurred in the early 2000s, and by 2010 almost 8.5% of the total number of research articles in the top-ranked journals were experimental.

There are clear disciplinary differences. The growth of experimental designs was most pronounced in economics, and particularly in the *American Economic Review*, where by the end of the period 18% of research articles were experimental. Political science also witnessed a growth, albeit tempered in comparison with

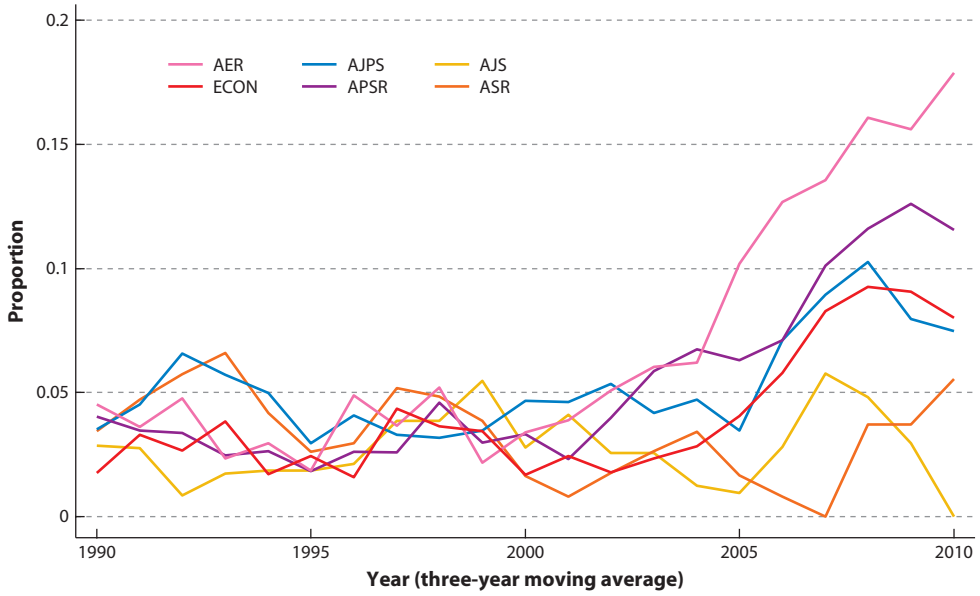

Figure 1

The proportion of research articles employing experimental design in top-ranked journals in economics, political science, and sociology (three-year moving average). Abbreviations: AER, *American Economic Review*; ECON, *Econometrica*; AJPS, *American Journal of Political Science*; APSR, *American Political Science Review*, AJS, *American Journal of Sociology*; ASR, *American Sociological Review*.

economics.[2] Indeed, the notable exception to the general pattern is sociology, where the visibility of experimental design fluctuates in a rather trendless fashion over the period considered.

An interesting qualitative feature of the content analysis was the extent to which nonexperimental papers increasingly referenced the language of experimental design. In particular, the term "natural experiment" appears to have gained in popularity in describing studies that would have been described as standard observational studies in the past (and were therefore coded as nonexperimental in our content analysis). For example, the implementation of a new social policy might today be described as a natural experiment, with the researcher drawing inferences about causal processes when comparing the pre– and post–policy change circumstances. The term "natural experiment" probably derives from the guidelines of Bradford Hill (1965), who specified conditions in which observational studies in epidemiology might reasonably point toward a causal interpretation; note that they were guidelines, and not criteria. A natural experiment in this context meant a large nonplanned intervention whose impact was so large that it could be separated from possible confounding effects. An extreme example is the use in radiation epidemiology of data from the survivors of the Hiroshima bomb. Such a version of the term "natural experiment" is too strong for many of the social science contexts in which it is applied.

The content analysis thus demonstrates a substantial increase in the visibility of experimental designs in the social sciences over the past 20 years. Other sources further

[2]Druckman et al. (2006) carried out a content analysis of articles published in the *American Political Science Review* over a much longer period than that considered here (between 1906 and 2004) and found a sharp growth toward the end of the period in the number of published articles reporting the results of experiments.

demonstrate that studies employing experimental designs are more frequently cited than other research articles, which suggests that the increased visibility of experimental studies in the top-ranked journals will be more than matched by the impact of these studies within the social scientific literature (Druckman et al. 2006, p. 632).

PRINCIPLES OF EXPERIMENTAL DESIGN

Sociologists who are planning experiments have available to them all manner of textbooks and practical guides (e.g., Campbell & Stanley 1963, Brown & Melamed 1990), alongside existing experimental research that provides convenient templates for new studies. But when considering the details of the design of any particular experiment, it is helpful to return to the small set of basic principles that lie at the heart of strong experimental design. Detailed discussion of general principles of experimental design in the presence of substantial uncontrolled variation largely dates from Fisher (1926), who developed his ideas in a book (Fisher 1935) still in print. He in effect suggested the four main principles that are the touchstones of experimental design to this day.

First, it is important to eliminate possible systematic errors. This may be achieved by ensuring that major predictable sources of unwanted variability affect all treatments equally and, importantly, by random allocation of treatments to units, ensuring that each unit is equally likely to receive each treatment. In this context, randomization always means use of an impersonal device with known probabilistic properties. Randomization is in principle required at every stage of a study at which appreciable uncontrolled variability might enter and is often importantly combined with concealment to deal with various kinds of subjective preference, which in some kinds of studies may easily distort conclusions.

We require conclusions of reasonable precision: that is, conclusions that are based on sufficient data to stand out above random error. Precision is to be distinguished from accuracy, which also requires freedom from systematic error. The second principle is therefore to enhance precision by comparing like with like. That is, the experimental units are grouped into sets or blocks expected to respond similarly, arranged such that each treatment occurs the same number of times in each block. The broad idea is that control through comparison against some null situation is essential to establish that a certain intervention produces a particular effect. Further, if an intervention produces a particular type of effect, we need to compare the consequences of various changes to the intervention in order to probe the nature of that effect.

Thirdly, replication is typically important both to verify directly repeatability of conclusions and also to enhance precision and to allow empirical estimation of the reliability of the estimated treatment effects. Replication, provided it is genuinely independent, gives a guide to the reproducibility of conclusions and ensures that confidence limits and other aspects of statistical analysis deliver a reasonable assessment of the reliability of conclusions.

The fourth notion is that of a factorial experiment. In a factorial experiment, several questions are addressed within one experiment rather than by a one-at-a-time approach, and thus the factorial route may be much more efficient in use of resources. Each question is represented in the design as a factor to be varied, and every combination of factors is assigned to the experimental units. One example is two different kinds of change in policy, each implemented for either a short, medium, or long period and each implemented in either a weak or strong form. This yields $2 \times 3 \times 2$ possibilities, and a factorial experiment would involve all these, each typically implemented on a number of distinct experimental units (see Ledolter & Swersey 2007 on factorial designs). Factorial designs allow for the estimation of statistical interactions that may be crucial for interpretation. In an extension of the idea, some of the factors may refer to baseline properties of the study individuals: age, gender, and ethnicity, for

example. An absence of important interactions of treatment effects with these baseline properties may be an important source of support for generalizability. Factorial experiments are frequently used in the social sciences, particularly in the context of survey experiments.

The Design of Experiments

There are several stages in the design of an experiment. Keeping the research objectives in mind, the researcher begins by clearly outlining the alternative treatments under investigation, treatments that often consist of all or most combinations of a number of factors (that is, variables). The researcher chooses the experimental units (or subjects) and records appropriate baseline features of the units to allow for comparison of pre- and post-treatment conditions. Subsequently, the researcher allocates one treatment to each experimental unit. In a narrow sense, the procedure for assigning treatments to experimental units is in fact what constitutes the specific choice of experimental design. The outcome(s) of interest must be identified, and procedures for measuring the outcome(s) must also be determined. As the design is implemented in the experimental setting—laboratory, field, or survey—careful monitoring will ensure both that the design is accurately implemented and that it can be adjusted where necessary.

In planning the design, the researcher aims to keep four key requirements in mind. First, the design should be capable of delivering conclusions that are free of systematic error and more broadly of ambiguities of interpretation. The importance of this requirement was highlighted in the Hawthorne experiments, where the findings of the experiments could not be unambiguously interpreted with respect to the research questions of interest (an aspect that, somewhat ironically, secured the study's place as a sociological classic). Second, the statistical errors of estimation connected with the design should be reduced to a reasonable level and it should be possible to estimate the magnitude of such errors, for example by calculating confidence intervals for the size of any comparisons estimated. Ideally, adequate precision will be achieved as economically as is feasible, and the confidence intervals will not be too wide. Third, it should be relatively simple to implement the design reliably. Overly complex designs are to be avoided. And fourth, especially for investigations intended as a basis for practical application on a broad front, the conclusions should have a broad range of validity. In major investigations, especially those of a novel kind, small pilot investigations are often carried out to allow the design to be tested and refined before the final implementation and data collection. The extent to which the four requirements can be met is likely to differ depending on whether designs are applied in the laboratory, field, or survey context, as we discuss below. For example, it may be more difficult to satisfy the third requirement—that it should be relatively simple to implement the design reliably—in a field experiment than in a laboratory study.

Choice of Experimental Units and External Validity

An experimental unit is the physical entity that is subject to treatments and upon which outcomes will be recorded; it corresponds to the smallest subdivision of the system under investigation such that any two units might receive different treatments. Sometimes it may be desirable or even essential to use the same physical entity several times as distinct experimental units, commonly known as a within-subject design in the social sciences. Thus, in a laboratory experiment, an experimental unit will often be a subject-session combination, each subject taking part in several sessions. This arrangement may be efficient in the sense that treatment contrasts can be studied free of simple intersubject systematic differences, and it is particularly appealing if the number of subjects available is limited. The classic social psychology experiment designed by Triplett (1898) included a within-subject component, with each subject completing the same task six times under two different

treatments (competition, absent or present); an important feature of the analysis was a within-subject comparison of performance under the different treatments. A potential difficulty with within-subject designs is that the outcome observed on a particular unit may depend not only on the treatment experienced in that session but possibly on the treatments that the subject received in earlier sessions. Suitable spacing of sessions may alleviate the problem, and if the so-called carryover or residual effect is confined to one session, appropriate design and analysis will correct for that. The possibility of more complex dependencies on the previous experience of the subject would be a major concern. As we discuss below, a standard protocol for experimental economics laboratories is that deception is forbidden, a protocol that has arisen, in part, to ensure that the previous experience of an experimental subject does not contaminate subsequent studies.

The choice of experimental units is consequential for what social scientists refer to as external validity, known also as range of validity and, most recently, as transportability (Pearl & Bareinboim 2011). External validity concerns the extent to which conclusions of a study can be applied more generally, and it depends both on the generalizability of the sample of experimental units and on the generalizability of the context within which the experiment takes place.

For studies bearing directly on general social science theory or public policy, range of validity regarding experimental units potentially depends on three aspects. The first is the possibility of support from theory or external information, which provides the basis for a claim that the experimental findings are likely to be constant across individuals and contexts, and therefore generalizable. A well-constructed sociological theory will specify the scope conditions under which the theory will hold and to which the findings will be generalizable (on scope conditions see especially Walker & Cohen 1985, Cohen 1988). The second and most immediately important aspect in the design is the use of key characteristics of the study individuals with the hope of showing stability of the primary effects with respect to these characteristics (see, for example, Cox 1958, pp. 9–11). Thus, depending on context, verifying that the conclusions applied regardless of gender, age, and ethnicity would give some basis for extrapolation. The third is that the individuals studied are, if not a respectable sample in a formal sense, at least broadly representative of the target population. The emphasis on each of these three aspects differs between laboratory, field, and survey experiments, which we discuss in more detail below.

Range of validity regarding the experimental context depends only on whether the context of the experiment—e.g., laboratory, field, survey—provides an appropriate basis for generalizability. Highly artificial environments may be a threat to external validity even when the experimental units are perfectly representative of the population of interest.

Importantly, for the detailed investigation of a new phenomenon, external validity may be of relatively little concern. For example, to study in depth the spread of a new infection or the causes of social unrest, it would be appropriate to concentrate on areas where the infection or unrest was common regardless of general representativeness.

Choice of Treatments

In some ways, the most crucial aspect of experimental design is the specification of treatments, the comparison of which will yield interesting conclusions free of ambiguities of interpretation. The simplest treatment structure has just two treatments, often regarded as a control or null treatment and a positive treatment or intervention. It is important that both are clearly specified, and in such a way that any difference in outcome found has a clear underlying interpretation. It is possible to have a number of qualitatively different treatments with or without a control. As described above, a common approach is to use a factorial design in which treatments are identified by several clearly defined features, each of which could define a new treatment on its own. Typically,

the various treatments will be equally replicated, although exceptions may be made when some treatments are more expensive or time-consuming to apply, or when there is a control and several other treatments and the main emphasis is on comparing the other treatments individually with the control when extra replication of the control is desirable.

CONTEMPORARY EXPERIMENTAL DESIGN IN THE SOCIAL SCIENCES

In this section, we consider different classes of experiments commonly applied in the social sciences and discuss the design issues associated with each. The classes of experiments that we focus on are laboratory experiments, field experiments, and survey experiments. To some extent, this division of experimental designs into different classes is arbitrary, but as the categories are well understood by social scientists, this categorization is appropriate here. For each class of experiment, we provide examples of the type of work carried out using these designs in sociology and other social sciences, as well as an assessment of the strengths and weaknesses of the class of experiments in relation to the principles of design.

Laboratory Experiments

Laboratory experiments are generally understood to come closest to the types of experiment common in some parts of the natural sciences, in that they take place in a physical location determined by the researcher, and the researcher has a high degree of control over treatments and other experimental conditions. Laboratory experiments in sociology are typically described as empirically driven or theory driven, alternatively labeled descriptive or formal (e.g., Martin & Sell 1979, Thye 2007, Walker & Willer 2007). Empirically driven experiments aim to establish empirical regularities, so that general conclusions about real-world phenomena can be drawn. In contrast, theory-driven experiments aim to isolate causal processes and test theory. Where well-formulated social scientific theories exist, theory-driven experiments allow researchers to create the conditions under which causal relationships can be identified. The use of the word theory in these contexts is different from the usage in the natural sciences, however, where the word hypothesis would be used for a proposal suggested provisionally as a basis for interpretation.

Laboratory experimentation has historically had a relatively small role in social science data collection, albeit an important one.[3] The social psychology field in sociology has a long-standing experimental tradition, with significant areas of research based around testing central theories, particularly expectation states theory and exchange theory (for summaries of the former theory, see Berger et al. 1977, Berger & Wagner 1993, Berger & Webster 2006; for summaries of the latter theory, see Cook et al. 1990, Molm & Cook 1995, Molm 2007). On the former, experimental tests of expectation states theory manipulate status characteristics within and between recognized social groups (e.g., men and women, young and old) in the laboratory to examine whether social differences form the basis of status distinctions that subsequently influence behavior (e.g., Ridgeway 1991, 2001; Ridgeway et al. 1998, 2009; Ridgeway & Erickson 2000; Ridgeway & Correll 2006; Berger 2007). On the latter, social psychologists working in the exchange theory tradition argue that human interactions must be understood in the context of exchange relationships, in which individuals offer material and nonmaterial goods in the expectation that they will receive similar goods in exchange (e.g., Blau 1955, 1964; Homans 1958; Cook 1987). Laboratory experiments are designed to manipulate both the social networks within which exchanges will be undertaken as well as the form of social exchange (Cook & Emerson 1978; Cook et al. 1983; Molm et al. 2000, 2001;

[3] We do not consider laboratory experimentation in political science here, but for an introduction, see McDermott 2007, Palfrey 2009, Morton & Williams 2010, Iyengar 2011.

Cook & Cooper 2003). A useful summary of the role of laboratory experimentation in testing expectation states theory and exchange theory can be found in Bonacich & Light (1978).

One type of laboratory experiment that has come to prominence in recent years, and with influence across the social sciences, is that carried out under the auspices of experimental economics. On the whole, experimental economists are engaged in explicitly theory-driven experiments that speak to the central tenets of economic theory. Levitt & List (2007, p. 153) write that "[t]he allure of the laboratory experimental method in economics is that, in principle, it provides *ceteris paribus* observations of individual economic agents, which are otherwise difficult to obtain." Special areas of interest in experimental economics include auction theory, choices made under risk and uncertainty, the limits to rationality, public goods, social preferences, trust, and asset markets. See Friedman & Sunder (1994, appendix 1) for readings in experimental economics, split by subject area, and Kagel & Roth (1995) for reviews of experimental findings, again by subject area. Levitt & List (2007) also give a thoughtful summary of economics experiments on social preferences.

For economics, a discipline with some strongly embedded theory and assumptions about human behavior, the theory-driven experiment has much appeal, as the degree of control afforded by the laboratory allows for situations that mimic those specified by economic theory, ensuring internal validity. Many economics experiments (or games) are of relatively simple design. For example, in an ultimatum game, there are two paired participants. One participant (the proposer) is provided with a sum of money and asked to propose a division of the total sum between herself and the other participant (the responder). The responder is asked whether he is prepared to accept the proposed division; if he accepts, the sum is divided as proposed, but if he rejects the proposal, neither player receives any money (Eckel 2007, p. 507). An important feature of this game, and indeed of many other games carried out under the auspices of experimental economics, is that the study is not, in fact, experimental in the sense used in the present paper because the comparative element that is crucial to the concept of an experiment is often missing. Rather, investigations of this type are careful observational studies of the behavior of subjects in a tightly defined context mimicking aspects of the real economic world. In a corresponding experimental setting, the context would be specified by a limited number of component elements capable of separate variation and the emphasis would be on estimating the impact and relative importance of these components, with the aim of understanding how the context generates its effect. Use of the same subject for two or more different settings would often be precluded by the possibility of complex dependencies of later responses on the earlier setting, although such carryover effects may sometimes be reduced by deceiving the subjects as to the objective of the investigation. However, such deceptions are usually forbidden by the protocols of the experimental economics laboratory.

The internal validity afforded by laboratory experiments is seen to be the greatest strength of the method, enabling precise and clear tests of social scientific theories. But a standard objection to the laboratory experiment, familiar to anyone who has ever opened a social science methods textbook, is that external validity is lacking because the environment in the laboratory is artificial, and experimental results are therefore not generalizable. As external validity is so contentious a topic within discussions of laboratory experiments, we here discuss the issues at stake in some detail. The external validity issues are in fact seen to be rather different for empirically driven and theory-driven experiments (Thye 2007).

In an empirically driven experiment, which aims to establish empirical regularities, generalizability is seen to be enhanced through the application of statistical principles: In an experiment conducted on a representative sample of individuals, we would observe differences between control and treatment groups and ask

whether these differences were likely to have been observed by chance (Walker & Willer 2007). Thus, results from empirically driven experiments are purportedly generalizable in just the same way as nonexperimental results (Thye 2007). But there are two potential threats to external validity and generalizability that we see as important for empirically driven laboratory experiments. First, if empirical regularities are to be established and immediately generalized to the population, clearly the sample of experimental units within the experiment should be broadly representative of that population to which the researcher wishes to generalize (Martin & Sell 1979). The second threat to external validity is one of context and comes from the artificiality of the laboratory environment. If the aim is to generalize from a laboratory experiment to claim an empirical regularity, there must be some confidence that what happens in the environment of the laboratory is a reasonable representation of what happens in the real world. There is evidence that some real-world contexts can be reasonably approximated in a laboratory setting (e.g., Correll et al. 2007, p. 1333).

In theory-driven experiments, the artificial environment of the laboratory is embraced in order to fully isolate causal processes of interest and thus to test elements of scientific theories. The aim of such experiments is not to generalize to a wider population, but rather to hone and test a scientific theory that, if validated, could later be applied to a real-world context to explain human behavior (Webster & Kervin 1971, Zelditch 2007). As Martin & Sell summarize,

> Experimental tests are artificial so their results cannot be directly generalized to natural phenomena as they operate in the real world. Instead, the results are pertinent only to the law-like statements that the experiment was designed to test. . . . The relation of the results to the real world is indirect in that the results of an experimental test allow only for the assessment of the support or nonsupport for the law-like statements. If the law-like statements are supported, then the statements may be used to explain and predict phenomena of the real world. . . . In other words, the specific experimental results are not generalized to the real world; however, when empirically supported, the law-like statements are. (Martin & Sell 1979, p. 587; see Webster & Kervin 1971, p. 269, for a similar argument)

From this perspective, the two threats to external validity described in relation to empirically driven laboratory experiments—the representativeness of the sample of individuals and the artificiality of the laboratory environment—are thus not applicable to theory-driven experiments.

The literature on laboratory experiments suggests that there is a broad level of agreement that the external validity concerns differ between empirically driven and theory-driven experiments. But it is not at all clear that, first, theory-driven experiments are immune from the requirements of external validity by virtue of their special role in testing theory or, second, that all social scientists who conduct theory-driven experiments explicitly limit themselves to making nongeneralizable, theoretically based claims. The first of these criticisms centers on the principle that a theory-driven laboratory experiment may indeed be able to demonstrate that x causes y but that it demonstrates this causal relationship only under condition l, the laboratory condition, and for sample s. To claim that a theory-driven experiment can be generalized through the law-like statements supported by the results is simply to claim that condition l and sample s are not relevant to the applicability of the law-like statements (rather than to establish this in reference to external theory and data). In the natural sciences, such assumptions may be somewhat less controversial, but in the social sciences we are bound to ask whether, say, the experimental context of the economics laboratory can simply be ignored because the results will speak to theory, particularly when the step from theory to the application of theory in explaining empirical phenomena is sometimes rather small. Only when we can find support for the assumptions that l and

s do not influence our findings is external validity assured (e.g., Zelditch 2007). Our second criticism accepts, for the sake of argument, that theory-driven experiments are less susceptible to concerns about external validity, but it questions whether theory-driven experimenters are always clear that the empirical results are not generalizable in any simple way. In practice, the distinction between empirically driven and theory-driven experiments is often blurred, and the research results of experiments that are described as theory-driven are generalized to a wider population without a full discussion of the role of experimental units and context.

A final topic worthy of discussion is how far the practical arrangements of laboratories help or hinder external validity. In many institutions, the laboratory condition and perhaps even the experimental subjects will be determined by someone other than the researcher. Laboratory protocols are particularly stringent in experimental economics, where to a large degree there is consensus on how laboratories should be set up and how experiments should be conducted. One rule present in almost all economics laboratories states that participants must not be deceived, even though deception has long been a rather standard feature of laboratory experiments in other social science disciplines (and particularly in social psychology). Friedman & Sunder (1994) argue that "experimental economists require complete credibility because salience and dominance are lost if subjects doubt the announced relation between actions and rewards, or if subjects hedge against possible tricks" (p. 17; see also Eckel 2007). More practically, many laboratories outlaw deception in order that a finite subject pool is not contaminated for future research by deception in previous experiments (e.g., Holt 2007). But the strategies used to avoid deception in the laboratory may be so artificial, or so different from strategies that would be employed were deception to be permitted, that the experimental results are no longer adequate to speak to the validity of a theory. As Cook & Yamagishi (2008) state, a consequence of the nondeception protocol is that some nonrational elements of human behavior are "ruled out by design" (p. 216; see also Sell 2008). Other laboratory (and journal) protocols put in place to protect subject pools and uphold standards (including an understanding that participants should always be paid in hard cash, as cash rewards are seen to most closely mimic the conditions of a market) also have the potential to alter the context of the experiment in theoretically relevant ways (e.g., Guala 2005, ch. 11). To some extent, the laboratory protocols within experimental economics should be of concern only to economists, but in institutions where resources are pooled, common protocols may be imposed for disciplines across the social sciences, resulting in harmful consequences for sociological experimental design.

Field Experiments

The distinction between laboratory and field experiments is seemingly well understood by social scientists, who if asked to describe the difference between the two would perhaps state something in line with the following: Field experiments are distinguished from laboratory experiments by virtue of taking place in a real-world context. But it is, in fact, extremely difficult to pin down the distinction between laboratory and field experiments. What is a real-world context? Are the Hawthorne experiments examples of field experiments, because they were undertaken in a real-world factory? Or does the fact that a group of workers was separated from other factory workers to be observed in a special room mean that the studies should be characterized as laboratory experiments?

One important difference between field and laboratory experiments lies in the degree of control that the investigator holds. Humphreys & Weinstein (2009, p. 369) argue that in a field experiment, "Researchers typically maintain control over the assignment to treatment while forgoing a measure of control over the treatment itself. Features such as the characteristics of subjects, the information available to them, and the precise manner and context in which the treatment is applied are more likely to

take on values given by 'nature' rather than being set at the discretion of the investigator." Although many researchers would provide similar definitions (e.g., de Rooij et al. 2009), there is by no means complete agreement on the classification of experiments as laboratory or field. For example, Harrison & List (2004) provide a taxonomy of experiments with four categories, and with six classification rules to determine into which category the experiment falls (pp. 1012–14). Here, however, we settle on Humphrey & Weinstein's (2009) definition of a field experiment, as this succinctly captures the distinction that we wish to highlight. Thus, the real-world quality that social scientists intuitively recognize as the distinguishing feature of field experiments stems from the researcher's lack of control over the experiment once the treatment has been implemented within a naturalistic setting. In this light, the Hawthorne experiments are best described as field experiments, albeit ones with a tighter degree of control than might be possible in other studies.

Concerns about the external validity of laboratory experiments have always acted to increase the appeal of field experiments to sociologists (for a discussion of influential field experiments outside of sociology, see de Rooij et al. 2009, Humphreys & Weinstein 2009, Green & John 2010). Field experiments are understood to have the capacity to test causal questions in a setting that is largely generalizable to human behavior as it occurs in the wild, so to speak. One area in which field experiments have been particularly favored is the study of discrimination, a phenomenon that is typically construed as wrong, unfair, and in many cases illegal. Given the deviant nature of discriminatory practices, almost all alternative observational research methodologies can provide only poor (and possibly misleading) evidence of the extent and nature of discrimination.

A standard design for an experimental study of discrimination is the correspondence test, and we here describe this design as it would be applied to study race discrimination in recruitment. The researcher takes the role of a potential employee, making applications for a number of jobs. Typically, two or more letters of application are sent to a single employer. The letters are substantively the same, varying only in respect to the treatment condition; thus, all characteristics of an individual that might be deemed relevant to job performance (e.g., qualifications or experience) and all characteristics of the firm are controlled, allowing the researcher to evaluate the effect of the race of the applicant alone (a matched-pairs design). In studies of race discrimination, the treatment is most commonly the name of the applicant, because names are taken to be a strong signal of race; in Bertrand & Mullainathan (2004), for example, half of the applications were attributed to Emily Walsh or Greg Baker and half to Lakisha Washington or Jamal Jones. Interest is in how far employers differ in their behavior toward the applicants: If identically qualified applicants receive systematically different responses from the employer, the difference can be attributed to the treatment, and race discrimination is established. Virtually all studies employing a correspondence test design find evidence of discrimination (e.g., Daniel 1968; McIntosh & Smith 1974; Newman 1978, 1980; Commission for Racial Equality 1980, 1996; Firth 1981; Howitt & Owusu-Bempah 1990; Esmail & Everington 1993; Noon 1993; Hoque & Noon 1999; Bertrand & Mullainathan 2004; see also Pager & Shepherd 2008). The correspondence test design has been adapted to study other forms of discrimination; for example, Correll et al. (2007) demonstrate that employers discriminate against mothers, whereas Jackson (2009) uses a factorial design to show that characteristics related to class origin are used as a basis for discriminatory judgments by employers.

Audit studies have employed similar experimental designs to test whether employers discriminate when faced with job candidates in person rather than written applications; in these studies, equally matched actors of different racial backgrounds apply for the same job, and if invited, subsequently attend job interviews. Once again, results from these experiments provide robust evidence of racial

discrimination by employers (McIntosh & Smith 1974, Cross et al. 1990, Turner et al. 1991, Commission for Racial Equality 1996, Riach & Rich 2002). In an influential sociological audit study, Pager (2003) extends the standard audit study design to assess the extent of discrimination related both to race and to possession of a criminal record. She finds strong evidence of employer discrimination against African Americans and against those with a criminal record (Pager 2003, p. 958).

The studies of discrimination throw into sharp relief some of the advantages of experimental design as applied to social science questions. Most obviously, discrimination studies directly test the level of discrimination faced by potential employees by using a simple experimental design with easily interpretable results. Furthermore, the experiments have a strong claim to external validity because they so closely mimic the situation that would obtain in the real world, and thus the conclusions can be readily applied to a wider societal context. But discrimination studies also demonstrate some of the weaknesses of field experiments. A lack of control over treatment is a potential problem, particularly for audit studies, where actors play the role of job applicants and where the experimental results will therefore be influenced by the quality and consistency of the actor's performance. By their nature, audit studies are not double-blind, as the actor must know to which condition they have been assigned in order to play the role, and his or her behavior might be influenced by this knowledge (e.g., Pager 2003, pp. 969–70; 2007). Furthermore, Heckman & Seligman (1993) argue that while, for example, two actors are matched on characteristics related to productivity in order to isolate the effect of race on the chances of a successful job application, productivity-relevant characteristics of the actors that are unobserved by the researcher but observed by employers mean that this test of discrimination is not a pure one (see also Heckman et al. 1998, Neumark 2011). A final concern voiced in relation to discrimination studies is not a design issue but an ethical one: Is it morally acceptable to deceive the participants in the experiment, who are indeed unaware that they are participants at all? Although detailed discussion of this point may be best left to those with special expertise in ethical issues, it bears noting that field experiments relying on a degree of deception for their success face practical hurdles put in place by universities, professional bodies, and presses.

A form of field experiment that we do not consider in detail here is the trial, or policy intervention study, also labeled the social experiment (e.g., Greenberg & Shroder 2004). Trials are to all intents and purposes field experiments, but they tend to be conducted on a larger scale than most social science field experiments, and they aim to evaluate the effects of particular policy interventions on social outcomes. These experiments are particularly important for disciplines in which there is significant interest in altering human behavior and/or social processes, such as criminology and educational research. For a discussion of design issues in the context of social experiments, see Bloom (2005), as well as Michalopoulos (2005), Riecken & Boruch (1978), Cook & Shadish (1994), and for summaries of specific social experiments, see Greenberg & Shroder (2004). For a thoughtful discussion of the importance of experimentation within criminological research, see Sherman (2007, 2009); examples of experimental criminology are described in Barnes et al. (2010) and Farrington & Welsh (2005). On experiments in educational research, the influential work of Mosteller and colleagues on the Tennessee class size project demonstrates the utility of experimental designs in assessing and guiding educational policy (Mosteller 1995, Mosteller et al. 1996, Cook 2002).

Survey Experiments

The final class of experiments that we consider is the survey experiment, also known as the population-based survey experiment. A basic definition comes from Mutz's (2011, p. 2) recent book on survey experiments: "[A]

population-based survey experiment is an experiment that is administered to a representative population sample. . . . [It] uses survey sampling methods to produce a collection of experimental subjects that is representative of the target population of interest for a particular theory." Just as with field experiments, in a survey experiment the investigator has full control over the experimental treatments and over assignment of treatments to experimental units, but little control over the treatments or subjects once the experiment is under way. Survey experimentation has been made more feasible for social scientists with a significant improvement in the infrastructure available to conduct such experiments; computer-aided survey techniques allow for greater flexibility of design than in the past, and organizations that support data gathering through survey experiments are growing in prominence (for example, Knowledge Networks and TESS: Time-Sharing Experiments for the Social Sciences).[4]

In a survey experiment, the researcher embeds the experimental design within a sample survey, often alongside standard survey instruments (such as those measuring demographic information or voting intentions). Treatments are assigned randomly to individuals participating in the survey. One common design, known as the factorial survey, employs vignettes to measure attitudes toward subjects or groups (for more on the use of vignettes in survey experiments, see Alves & Rossi 1978, Rossi 1979, Rossi & Anderson 1982, Rossi & Nock 1982). Vignettes are short, hypothetical descriptions of individuals or situations, and in a factorial survey, the researcher designs a number of vignettes in order to manipulate respondents' attitudes toward an "object of judgment" (Mutz 2011, p. 54). For example, Jasso & Opp (1997) used a factorial survey to measure norms around political protest in the city of Leipzig. They identified several potential features that could lead individuals to feel duty-bound to undertake a political protest, including the level of discontent with the system, the types of protest actions available, and the level of personal risk expected by participating in political protest. They then designed vignettes that varied these features: 504 different vignettes were constructed to represent all combinations of the desired treatments, of which 36 vignettes were excluded for reasons of plausibility. Each respondent was asked to evaluate 10 of the vignettes chosen at random, and these evaluations comprise the outcome variable of the analysis. Interest is therefore in assessing the effect of the treatments, both alone and in combination with other treatments, on evaluative judgments (or attitudes).

The possibility of multifactorial designs, in which multiple treatments are identified and every combination of treatments is tested, is a significant advantage of survey experiments. The mode of administration of the survey experiment allows for much larger sample sizes than would be feasible in a laboratory or perhaps even a field experiment, and these larger sample sizes provide the statistical power to detect even modest effects. In addition, each respondent to the survey will evaluate a set of vignettes, the size of the set being determined by the researcher, allowing for tests of within- and between-subject differences in responses. In statistical analysis of the treatments, both main effects and interaction effects can be examined, which may uncover significant but previously unexpected results. Thus, survey experiments fulfill many of the criteria identified earlier as features of strong experimental design.

But perhaps the most attractive feature of survey experiments for social scientists is that, like field experiments, they are seen to

[4]Many present-day survey experiments are conducted online, although not all online (or Web-based) experiments are survey experiments. Increasingly, researchers are implementing designs online that have been previously implemented in the laboratory, with no attempt to reach a representative sample of respondents as in a standard survey experiment. Online experiments of this type are likely to be subject to design considerations that differ somewhat from experiments implemented in other contexts, not least because of technological constraints (see, e.g., Kohavi et al. 2009). It is not yet clear whether online experiments will earn a place alongside laboratory, field, and survey experiments as a standard context for sociological data collection.

benefit from a high degree of external validity because the experiments are carried out on a representative sample of people outside of a laboratory. This high degree of external validity is coupled with a similarly high degree of internal validity; the researcher maintains control over the experimental conditions in much the same way as in the laboratory. Consequently, the results of the survey experiment are seen to be both easily generalizable and appropriate for testing causal hypotheses.

Although statements asserting the internal and external validity of survey experiments would be seen as largely uncontroversial among social scientists, the validity issues raised by survey experiments are in fact more complicated than one would at first suppose. As discussed, external validity is influenced both by experimental units and by context, and although a representative sample speaks to the generalizability of findings from the chosen experimental units, the artificiality of the survey environment does raise questions about generalizability of context: The survey experiment is not conducted in the real world, as with field experiments, but in an environment at least partially constructed by the researcher. Some research has indicated that the survey environment might in fact be more problematically artificial than that of the laboratory: For example, Collett & Childs (2011, p. 520) show that subjects respond differently to laboratory and vignette treatments that are designed to be equivalent, with stronger (and theoretically expected) affective responses observed in the laboratory, where the subjects experience the treatments in a more emotionally intense way. Experiments conducted in the laboratory might also encourage increased engagement and involvement on the part of subjects. Therefore, although surveys are administered in what might appear to be more naturalistic settings than the laboratory, behavior may still be influenced by the mode of administration of the survey, a lack of care and attention to the questions, the necessary artificiality of the questions, or social desirability effects, as well as by other potential problems associated with standard survey designs (e.g., Finch 1987, Eifler 2010). For a detailed discussion of external validity in survey experiments, see Mutz (2011, ch. 8).

Social science studies employing survey experiment designs include Sniderman et al. (1991), Williams et al. (2008), and Hopkins & King (2010). For impressive reviews of the design of survey experiments, particularly in relation to the study of political and social attitudes, see Sniderman & Grob (1996) and Sniderman (2011), and for an up-to-date summary of the use of survey experiments in the social sciences, see Mutz (2011).

SOME GENERAL CONSIDERATIONS

Despite the primacy of observational methods within the social sciences, and sociology in particular, experimental methods have had substantial and growing influence. In this review, we have provided examples of the types of social science questions amenable to experimental investigation and the design tools available to address those questions. But before embarking on a headlong rush to experimentation, the cautious sociologist may wish to consider not just the issues of experimental design reviewed here, but also the role of experiments in the production of scientific knowledge. The experimental method is extremely powerful, particularly in largely eliminating the contaminating role of unobserved confounders and thereby making causal explanation more secure, but a good part of that power derives from the integration of experimental methods with established scientific practices. Three (related) features of scientific practices should be highlighted: the role of theory, the importance of replication, and the role of disciplinary boundaries.

Rigorous research questions and/or hypotheses linked to theory are an important foundation of scientific discovery, and where the aim is to draw theoretically relevant and meaningful conclusions from empirical research, the theory and method must be integrated. For example, experiments may be used to test hypotheses derived from

well-established theories or to establish baseline empirical regularities that will potentially form the basis of new theories. One consequence of the integration of theory and method is that results from different studies are commensurable, and thus cumulative knowledge is made possible. In the social sciences, perhaps with the exception of economics and social psychology, there are few commonly accepted theories from which hypotheses can be derived and tested. Instead, there is a proliferation of theories, and it is not uncommon to see experimental tests of theories that have been established only a few pages before the experimental results are reported. In part this may be attributed to the perceived higher status attached to theory development in contrast to the testing of established theories. But the result is a lack of coherence, an incomprehensible pattern of small points of light in a dark sky.

The career incentives to invest in the replication of existing experimental results are also low, yet independent replication of experimental findings is absolutely essential for scientific progress. Fisher identified independent replication as one of the four central principles of experimental design, and Campbell & Stanley (1963, p. 3) similarly emphasize the importance of replication in their classic volume on experimentation in the social sciences:

> [W]e must increase our time perspective, and recognize that continuous, multiple experimentation is more typical of science than once-and-for-all definitive experiments. The experiments we do today, if successful, will need replication and cross-validation at other times under other conditions before they can become an established part of science, before they can be theoretically interpreted with confidence.

Independent replication does not simply entail exact repetition of the same design. For example, a factorial experiment may be set up partly to confirm previous conclusions but also to examine new features. It is largely through independent replication that the experimental method can be allowed to claim a privileged role in the testing of scientific theory in comparison with observational methods, although without some change in the rewards that replication confers, any simple injunction in support of replication will have little effect. But without independent replication, we are unable to judge the stability of the experimental conclusions or the extent to which those conclusions can be extended outside the scope conditions of the initial experiment. From the perspective of a cumulative approach to scientific knowledge, a single experiment is not likely to offer conclusive evidence in support of a general theory, and it is important that findings from both identical and distinct (but related) studies reinforce the same finding. The value of experimental research is to be found in the collectivity.

Disciplinary boundaries within the social sciences may to some extent limit the current usefulness of experimental methods. These boundaries discourage the establishment of general theories of human behavior by embedding individual theories within discipline-specific debates and terminology, consequently increasing the cost of independent replication by those outside the discipline. Disciplinary boundaries also limit the extent to which results of experiments are shared across specialized fields, so that a test of, say, cognitive dissonance reduction theory in economics may take place without reference to the extensive body of social psychological work testing the theory. Finally, even within social science disciplines, methodological and theoretical overspecialization works to limit the scientific usefulness of experiments. Social scientists are apt to describe themselves and others as *x* methodologists or *y* theorists, sometimes with an implication of all-inclusive explanation, but it is dangerous indeed if a discipline places the power to test theories in the hands of experimentalists, and the power to describe patterns of empirical regularities in the hands of observationalists.

The emphasis throughout this discussion has been on the design of individual experimental studies leading to secure conclusions as

far as possible free of worrying caveats. Nevertheless, progress typically depends on the synthesis of conclusions from various sources. Cochran (1965) remarked that he had earlier asked R.A. Fisher how he thought the conclusions from observational studies could be made more secure. The characteristically gnomic reply: "Make your theories elaborate." Cochran reported that Fisher meant that the evidence used should be as wide-ranging in nature as feasible. This must surely apply to both observationally and experimentally based evidence and strongly encourages mixtures of approaches.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We thank Karen Cook, Shelley Correll, David Grusky, Oscar Grusky, Colin Mills, Cecilia Ridgeway, and the anonymous reviewers for their very helpful comments on earlier drafts of this review.

LITERATURE CITED

Adair JG, Sharpe D, Huynh C-L. 1989. Hawthorne control procedures in educational experiments: a reconsideration of their use and effectiveness. *Rev. Educ. Res.* 59:215–28

Aiello JR, Douthitt EA. 2001. Social facilitation from Triplett to electronic performance monitoring. *Group Dyn.* 5:163–80

Allport FH. 1920. The influence of the group upon association and thought. *J. Exp. Psychol.* 3:159–82

Allport FH. 1954. The historical background of modern social psychology. In *The Handbook of Social Psychology*, ed. G Lindzey, pp. 3–56. Reading, MA: Addison-Wesley

Alves WM, Rossi PH. 1978. Who should get what? Fairness judgments in the distribution of earnings. *Am. J. Sociol.* 84:541–64

Asch SE. 1951. Effects of group pressure upon the modification and distortion of judgment. In *Groups, Leadership and Men*, ed. H Guetzkow, pp. 177–90. Pittsburgh, PA: Carnegie Press

Asch SE. 1956. Studies of independence and conformity: a minority of one against a unanimous majority. *Psychol. Monogr.* 70:416

Bandura A, Ross D, Ross SA. 1961. Transmission of aggression through imitation of aggressive models. *J. Abnorm. Soc. Psychol.* 63:575–82

Barnes GC, Ahlman L, Gill C, Sherman LW, Kurtz E, Malvestuto R. 2010. Low-intensity community supervision for low-risk offenders: a randomized, controlled trial. *J. Exp. Criminol.* 6:159–89

Berger J. 2007. The standardized experimental situation in expectation states research: Notes on history, uses, and special features. See Webster & Sell 2007, pp. 353–78

Berger J, Fiçek MH, Norman RZ, Zelditch M. 1977. *Status Characteristics and Social Interaction*. New York: Elsevier

Berger J, Wagner D. 1993. Status characteristics theory: the growth of a program. In *Theoretical Research Programs: Studies in the Growth of Theory*, ed. J Berger, M Zelditch, pp. 23–63. Stanford, CA: Stanford Univ. Press

Berger J, Webster M. 2006. Expectations, status, and behaviour. In *Contemporary Social Psychological Theories*, ed. PJ Burke, pp. 268–300. Stanford, CA: Stanford Univ. Press

Bertrand M, Mullainathan S. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *Am. Econ. Rev.* 94:991–1013

Blau PM. 1955. *The Dynamics of Bureaucracy*. Chicago: Univ. Chicago Press

Blau PM. 1964. *Exchange and Power in Social Life*. New York: Wiley

Bloom HS, ed. 2005. *Learning More from Social Experiments. Evolving Analytic Approaches.* New York: Russell Sage Found.

Bonacich P, Light J. 1978. Laboratory experimentation in sociology. *Annu. Rev. Sociol.* 4:145–70

Bradford Hill A. 1965. The environment and disease: association or causation? *Proc. R. Soc. Med.* 58:295–300

Brehm SS, Kassin SM, Fein S. 1999. *Social Psychology.* Boston, MA: Houghton Mifflin. 4th ed.

Brown SR, Melamed LE. 1990. *Experimental Design and Analysis* (*Quantitative Applications in the Social Sciences*). Newbury Park, CA: Sage

Campbell D, Stanley J. 1963. *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand-McNally

Chapin FS. 1936. Social theory and social action. *Am. Sociol. Rev.* 1:1–11

Cochran WG. 1965. The planning of observational studies of human populations. *J. R. Stat. Soc. A* 128:234–66

Cohen BP. 1988. *Developing Sociological Knowledge: Theory and Method*. Chicago: Nelson-Hall

Collett JL, Childs E. 2011. Minding the gap: meaning, affect, and the potential shortcomings of vignettes. *Soc. Sci. Res.* 40:513–22

Commission for Racial Equality. 1980. *Half a Chance? A Report on Job Discrimination Against Young Blacks in Nottingham*. Nottingham: CRE

Commission for Racial Equality. 1996. *We Regret to Inform You. . .* London: CRE

Cook KS. 1987. *Social Exchange Theory*. Beverly Hills, CA: Sage

Cook KS, Cooper RM. 2003. Experimental studies of cooperation, trust and social exchange. In *Trust, Reciprocity and Gains from Association: Interdisciplinary Lessons from Experimental Research*, ed. E Ostrom, J Walker, pp. 277–333. New York: Russell Sage Found.

Cook KS, Emerson RM. 1978. Power, equity and commitment in exchange networks. *Am. Sociol. Rev.* 43:721–39

Cook KS, Emerson RM, Gillmore MR, Yamagishi T. 1983. The distribution of power in exchange networks: theory and experimental results. *Am. J. Sociol.* 89:275–305

Cook KS, O'Brien J, Kollock P. 1990. Exchange theory: a blueprint for structure and process. In *Frontiers of Social Theory: The New Synthesis*, ed. G Ritzer, pp. 158–81. New York: Columbia Univ. Press

Cook KS, Yamagishi T. 2008. In defense of deception. *Soc. Psychol. Q.* 71:215–21

Cook TD. 2002. Randomized experiments in educational policy research: a critical examination of the reasons the educational evaluation community has offered for not doing them. *Educ. Eval. Policy Anal.* 24:175–99

Cook TD, Shadish WR. 1994. Social experiments: some developments over the past fifteen years. *Annu. Rev. Psychol.* 45:545–80

Correll SJ, Benard S, Paik I. 2007. Getting a job: Is there a motherhood penalty? *Am. J. Sociol.* 112:1297–38

Cox DR. 1958. *Planning of Experiments*. New York: Wiley

Cross H, Kenny G, Mell J, Zimmerman W. 1990. *Differential Treatment of Hispanic and Anglo Job Seekers: Hiring Practices in Two Cities*. Washington, DC: Urban Inst. Press

Daniel WW. 1968. *Racial Discrimination in England. Based on the PEP Report*. Harmondsworth: Penguin

de Rooij EA, Green DP, Gerber AS. 2009. Field experiments on political behavior and collective action. *Annu. Rev. Polit. Sci.* 12:389–95

Dickson WJ, Roethlisberger FJ. 1966. *Counselling in an Organization: A Sequel to the Hawthorne Researches*. Boston, MA: Harv. Bus. Sch. Press

Druckman JN, Green DP, Kuklinski JH, Lupia A. 2006. The growth and development of experimental research in political science. *Am. Polit. Sci. Rev.* 100:627–35

Eckel C. 2007. Economic games for social scientists. See Webster & Sell 2007, pp. 497–515

Eifler S. 2010. Validity of a factorial survey approach to the analysis of criminal behavior. *Methodology* 6:139–46

Esmail A, Everington S. 1993. Racial discrimination against doctors from ethnic minorities. *BMJ* 306:691–92

Farrington D, Welsh B. 2005. Randomized experiments in criminology: What have we learned in the last two decades? *J. Exp. Criminol.* 1:9–28

Finch J. 1987. The vignette technique in survey research. *Sociology* 21:105–14

Firth M. 1981. Racial discrimination in the British labour market. *Ind. Labor Relat. Rev.* 34:265–72

Fisher RA. 1926. The arrangement of field experiments. *J. Minist. Agric. G. B.* 33:503–13

Fisher RA. 1935. *Design of Experiments*. Edinburgh: Oliver & Boyd

Franke RH, Kaul JD. 1978. The Hawthorne experiments: first statistical interpretation. *Am. Sociol. Rev.* 43:623–43
Friedman D, Sunder S. 1994. *Experimental Methods: A Primer for Economists*. Cambridge, UK: Cambridge Univ. Press
Gillespie R. 1991. *Manufacturing Knowledge: A History of the Hawthorne Experiments*. New York: Cambridge Univ. Press
Green DP, Gerber AS. 2003. The underprovision of experiments in political science. *Ann. Am. Acad. Polit. Soc. Sci.* 589:94–112
Green DP, John P. 2010. Field experiments in comparative politics and policy. *Ann. Am. Acad. Polit. Soc. Sci.* 628:6–10
Greenberg DH, Shroder M. 2004. *The Digest of Social Experiments*. Washington, DC: Urban Inst. Press
Guala F. 2005. *The Methodology of Experimental Economics*. Cambridge, UK: Cambridge Univ. Press
Haney C, Banks C, Zimbardo P. 1973. Interpersonal dynamics in a simulated prison. *Int. J. Criminol. Penol.* 1:69–97
Harrison GW, List JA. 2004. Field experiments. *J. Econ. Lit.* 42:1009–55
Heckman J, Ichimura H, Smith J, Todd P. 1998. Characterizing selection bias using experimental data. *Econometrica* 6:1017–99
Heckman J, Seligman P. 1993. The Urban Institute Audit Studies: their methods and findings. In *Clear and Convincing Evidence: Measurement of Discrimination in America*, ed. M Fix, RJ Struyk, pp. 187–258. Washington, DC: Urban Inst. Press
Holt CA. 2007. *Markets, Games, and Strategic Behavior*. Boston, MA: Pearson
Homans GC. 1958. Social behavior as exchange. *Am. J. Sociol.* 63:597–606
Hopkins D, King G. 2010. Improving anchoring vignettes: designing surveys to correct interpersonal incomparability. *Public Opin. Q.* 74:201–22
Hoque K, Noon M. 1999. Racial discrimination in speculative applications: new optimism six years on? *Hum. Resour. Manag. J.* 9:71–82
Howitt D, Owusu-Bempah J. 1990. The pragmatics of institutional racism: beyond words. *Hum. Relat.* 9:885–99
Humphreys M, Weinstein JM. 2009. Field experiments and the political economy of development. *Annu. Rev. Polit. Sci.* 12:367–78
Iyengar S. 2011. Laboratory experiments in political science. In *Cambridge Handbook of Experimental Political Science*, ed. JN Druckman, DP Green, JH Kuklinski, A Lupia, pp. 73–88. New York: Cambridge Univ. Press
Jackson M. 2009. Disadvantaged through discrimination. the role of employers in social stratification. *Br. J. Sociol.* 60:669–92
Jasso G, Opp K-D. 1997. Probing the character of norms: a factorial survey analysis of the norms of political action. *Am. Sociol. Rev.* 62:947–64
Jones SRG. 1992. Was there a Hawthorne effect? *Am. J. Sociol.* 98:451–68
Kagel JK, Roth AE, eds. 1995. *Handbook of Experimental Economics*. Princeton, NJ: Princeton Univ. Press
Kohavi R, Longbotham R, Sommerfield D, Henne RM. 2009. Controlled experiments on the web: survey and practical guide. *Data Min. Knowl. Discov.* 18:140–81
Ledolter J, Swersey AJ. 2007. *Testing 1–2–3*. Stanford, CA: Stanford Univ. Press
Levitt SD, List JA. 2007. What do laboratory experiments measuring social preferences reveal about the real world? *J. Econ. Perspect.* 21:153–74
Levitt SD, List JA. 2011. Was there really a Hawthorne effect at the Hawthorne plant? An analysis of the original illumination experiments. *Am. Econ. J.* 3:224–38
Lowell AL. 1910. The physiology of politics. *Am. Polit. Sci. Rev.* 4:1–15
Marshall A. 1890. *Principles of Economics: An Introductory Volume*. London: Macmillan
Martin MW, Sell J. 1979. The role of the experiment in the social sciences. *Sociol. Q.* 20:581–90
Mayo E. 1933. *The Human Problems of an Industrial Civilization*. New York: Macmillan
McDermott R. 2007. Experimental political science. See Webster & Sell 2007, pp. 483–96
McIntosh N, Smith DJ. 1974. *The Extent of Racial Discrimination*. PEP, Broadsh. No. 547. London: PEP

Michalopoulos C. 2005. Precedents and prospects for randomized experiments. In *Learning More from Social Experiments. Evolving Analytic Approaches*, ed. HS Bloom, pp. 1–36. New York: Russell Sage Found.

Milgram S. 1974. *Obedience to Authority; An Experimental View*. New York: Harper & Row

Molm LD. 2007. Experiments on exchange relations and exchange networks in sociology. See Webster & Sell 2007, pp. 379–406

Molm LD, Cook KS. 1995. Social exchange theory. In *Sociological Perspectives on Social Psychology*, ed. KS Cook, G Fine, J House, pp. 209–35. Needham, MA: Allyn & Bacon

Molm LD, Peterson G, Takahashi N. 2001. The value of exchange. *Soc. Forces* 80:159–84

Molm LD, Takahashi N, Peterson G. 2000. Risk and trust in social exchange: an experimental test of a classical proposition. *Am. J. Sociol.* 105:1396–427

Morton RB, Williams KC. 2010. *Experimental Political Science and the Study of Causality: From Nature to the Lab*. New York: Cambridge Univ. Press

Mosteller F. 1995. The Tennessee study of class size in the early school grades. *Crit. Issues Child. Youths* 5:113–127.

Mosteller F, Light RJ, Sachs JA. 1996. Sustained inquiry in education: lessons from skill grouping and class size. *Harv. Educ. Rev.* 66:797–842

Mutz DC. 2011. *Population-Based Survey Experiments*. Princeton, NJ: Princeton Univ. Press

Neumark D. 2011. *Detecting discrimination in audit and correspondence studies.* NBER Work. Pap. 16448, Natl. Bur. Econ. Res., Cambridge, MA. **http://www.nber.org/papers/w16448**

Newman J. 1978. Discrimination in recruitment: an empirical analysis. *Ind. Labor Relat. Rev.* 32:15–23

Newman J. 1980. Reply. *Ind. Labor Relat. Rev.* 33:547–50

Noon M. 1993. Racial discrimination in speculative application: evidence from the UK's top 100 firms. *Hum. Resour. Manag. J.* 3:35–47

Oakley A. 1998. Experimentation and social interventions: a forgotten but important history. *BMJ* 317:1239–42

Pager D. 2003. The mark of a criminal record. *Am. J. Sociol.* 108:937–75

Pager D. 2007. The use of field experiments for studies of employment discrimination: contributions, critiques, and directions for the future. *Ann. Am. Acad. Polit. Soc. Sci.* 609:104–33

Pager D, Shepherd H. 2008. The sociology of discrimination: racial discrimination in employment, housing, credit, and consumer markets. *Annu. Rev. Sociol.* 34:181–209

Palfrey TR. 2009. Laboratory experiments in political economy. *Annu. Rev. Polit. Sci.* 12:379–88

Pearl J, Bareinboim E. 2011. External validity and transportability: a formal approach. *58th World Stat. Congr., Int. Stat. Inst., Aug., Dublin*. The Hague, Neth.: Int. Stat. Inst.

Riach P, Rich J. 2002. Field experiments of discrimination in the market place. *Econ. J.* 112:480–518

Ridgeway CL. 1991. The social construction of status value: gender and other nominal characteristics. *Soc. Forces* 70:367–86

Ridgeway CL. 2001. Gender, status and leadership. *J. Soc. Issues* 57:637–55

Ridgeway CL, Backor K, Li YE, Tinkler JE, Erickson KG. 2009. How easily does a social difference become a status distinction? Gender matters. *Am. Sociol. Rev.* 74:44–62

Ridgeway CL, Boyle EH, Kuipers K, Robinson D. 1998. How do status beliefs develop? The role of resources and interaction. *Am. Sociol. Rev.* 63:331–50

Ridgeway CL, Correll SJ. 2006. Consensus and the creation of status beliefs. *Soc. Forces* 85:431–54

Ridgeway CL, Erickson KG. 2000. Creating and spreading status beliefs. *Am. J. Sociol.* 106:579–615

Riecken HW, Boruch RF. 1978. Social experiments. *Annu. Rev. Sociol.* 4:511–32

Roethlisberger FJ, Dickson WJ. 1939. *Management and the Worker*. Cambridge, MA: Harvard Univ. Press

Rossi PH. 1979. Vignette analysis: uncovering the normative structure of complex judgments. In *Qualitative and Quantitative Social Research: Papers in Honor of Paul F. Lazarsfeld*, ed. RK Merton, JS Coleman, PH Rossi, pp. 176–86. New York: Free Press

Rossi PH, Anderson AB. 1982. The factorial survey approach: an introduction. See Rossi & Nock 1982, pp. 15–67

Rossi PH, Nock SL, eds. 1982. *Measuring Social Judgments: The Factorial Survey Approach*. Beverly Hills, CA: Sage

Roth AE. 1993. On the early history of experimental economics. *J. Hist. Econ. Thought* 15:184–209
Roth AE. 1995. Introduction to experimental economics. See Kagel & Roth 1995, pp. 3–109
Sell J. 2008. Introduction to deception debate. *Soc. Psychol. Q.* 71:213–14
Sherman LW. 2007. The power few: experimental criminology and the reduction of harm. *J. Exp. Criminol.* 3:299–21
Sherman LW. 2009. Evidence and liberty: the promise of experimental criminology. *Criminol. Crim. Justice* 9:5–28
Sniderman PM. 2011. The logic and design of the survey experiment: an autobiography of a methodological innovation. In *Cambridge Handbook of Experimental Political Science*, ed. JN Druckman, DP Green, JH Kuklinski, A Lupia, pp. 102–14. New York: Cambridge Univ. Press
Sniderman PM, Grob DB. 1996. Innovations in experimental design in attitude surveys. *Annu. Rev. Sociol.* 22:377–99
Sniderman PM, Piazza T, Tetlock PE, Kendrick A. 1991. The new racism. *Am. J. Polit. Sci.* 35:423–47
Strube MT. 2005. What did Triplett really find? A contemporary analysis of the first experiment in social psychology. *Am. J. Psychol.* 118:271–86
Thye SR. 2007. Logical and philosophical foundations of experimental research in the social sciences. See Webster & Sell 2007, pp. 57–86
Triplett N. 1898. The dynamogenic factors in pacemaking and competition. *Am. J. Psychol.* 9:507–33
Turner M, Fix M, Struyk R. 1991. *Opportunities Denied, Opportunities Diminished: Racial Discrimination in Hiring*. Washington, DC: Urban Inst. Press
Walker HA, Cohen BP. 1985. Scope statements: imperatives for evaluating theory. *Am. Sociol. Rev.* 50:288–301
Walker HA, Willer D. 2007. Experiments and the science of sociology. See Webster & Sell 2007, pp. 25–56
Webster JR, Kervin JB. 1971. Artificiality in experimental sociology. *Can. Rev. Sociol.* 8:263–72
Webster JR, Sell J, eds. 2007. *Laboratory Experiments in the Social Sciences.* Amsterdam: Academic/Elsevier
Williams MTE, Turkheimer E, Magee E, Guterbock T. 2008. The effects of race and racial priming on self-report of contamination anxiety. *Personal. Individ. Differ.* 44:746–57
Zelditch M. 2007. Laboratory experiments in sociology. See Webster & Sell 2007, pp. 517–31

Annual Review
of Sociology
Volume 39, 2013

Contents

Urban and Rural Community Sociology

Indexes

Errata

An online log of corrections to *Annual Review of Sociology* articles may be found at http://soc.annualreviews.org/errata.shtml