FLAWED SOCIAL EXPERIMENTS

David Greenberg
University of Maryland, Baltimore County

Burt S. Barnow
George Washington University

May 2013

Author Information

David Greenberg (corresponding author)

PhD Massachusetts Institute of Technology

Professor of Economics Emeritus

University of Maryland, Baltimore County

Email: dhgreenb@umbc.edu

Burt Barnow

PhD University of Wisconsin-Madison

Amsterdam Professor of Public Service and of Economics

Trachtenberg School of Public Policy and Public Administration

George Washington University

Email: barnow@gwu.edu

FLAWED SOCIAL EXPERIMENTS

May 2013

Abstract

This paper describes10, somewhat overlapping, types of flaws that have occurred in social experiments. Each flaw is illustrated with examples from previous experiments. Some of these problems result in minor hurdles, while others cause experiments to fail—that is, the experiment is unable to provide a valid test of the hypothesis of interest. An accompanying summary table lists the flaws, indicates the circumstances under which they occur, their potential seriousness, and approaches for minimizing them. The most important of the flaws are response bias resulting from attrition; a failure to adequately implement the treatment as designed; and too small a sample to detect impacts. The third of these flaws can result from insufficient marketing, too small an initial target group, disinterest on the part of the target group in participating (if the treatment is voluntary), or attrition. To a considerable degree the flaws discussed in this article can be minimized. For instance, implementation failures and too small a sample can usually be avoided with sufficient effort and planning and response bias can often be mitigated—for example, through increased follow-up efforts in conducting surveys.

Key Words: Social experiments, experimental flaws, response bias**,** implementation, cross-over

INTRODUCTION

It is widely agreed among social scientists that social experiments, where units are randomly assigned to treatment and control status, provide the best opportunity for learning about the effectiveness of a social intervention. For this reason, as noted by [citation deleted for blind review], the number of social experiments has grown almost exponentially since the 1960s. However, although social experiments are rightly viewed as the "gold standard" in evaluation, almost all of them confront problems of some sort in their implementation or operation. Some of these problems result in minor hurdles, while others cause experiments to fail—that is, the experiment is unable to provide a valid test of the hypothesis of interest.[1] There is obviously a continuum between a minor flaw and a fatal problem. In this paper, we examine serious experimental flaws, but not all of these necessarily result in complete failure. As will be seen, in some circumstances it was possible for analysts to overcome the flaw, at least in part; and even in the case of a fatal flaw, an experiment may sometimes still provide useful information. In all the situations we describe, however, the flaw causes findings from comparisons between treatment and control groups to be subject to considerable uncertainty or unable to provide the information desired, and users of the experimental findings must exercise great caution.

In the remainder of this paper we consider 10, somewhat overlapping, types of flaws that have occurred in social experiments, some more frequently than others, and some more serious than others. We illustrate each with examples from previous experiments. We limit these illustrations to what we define as *social experiments*, drawing lessons from areas such as health, education, employment and job training, welfare, and housing. Areas excluded include experiments dealing with medicines and medical issues, utility pricing, marketing, and consumer behavior. Social experiments, which deal with social policies, are a sub-category of what Harrison and List (2004) have termed "framed field experiments," experiments that take place in the natural environment of the subjects of the experiment and in which the subjects know they are participating in an experiment and typically consent to participate. They contrast this with laboratory experiments, which take place in a controlled environment, and "natural field

[1] We do not view a "failed experiment" as one in which an experiment finds no impact for the treatment.

experiments," which take place "in the environment where the subjects are naturally undertaking certain tasks and where the subjects *do not know* that they are participants in an experiment" (List, 2011, p. 6, italics in the original)

1. Few Show Up.

This problem, which is a major concern, occurs when some potential participants in the program being tested are available for random assignment, but there are many less than planned for in the research design, resulting in an insufficient sample for hypothesis testing.[2] This can happen when participating in the treatment being tested is voluntary, but is unlikely when participation is mandatory. It may be due to insufficient outreach and marketing, because the target group of those who qualify to participate in the experiment is too small, or because few persons in the target population think it is potentially beneficial for them to participate. It is obviously important to determine the actual source of the problem. A lack of outreach and marketing can be overcome, but too small a target group or disinterest probably cannot be. Too small a target group generally implies too little research and planning prior to undertaking the experiment, but, in the case of disinterest, the experiment has provided important information.

An interesting example of insufficient outreach occurred in implementing Britain's Employment Retention and Advancement (ERA) demonstration. One of the three target groups of the intervention, single parents who were working part time and receiving Working Tax Credits,[3] could receive financial incentive bonuses under ERA by working full-time (at least 30 hours a week), as well as have access to caseworker services. During the initial recruitment effort, these mothers were not told about the incentive payments, clearly an important selling point of ERA, because of concern over their disappointment if they were randomly assigned to the control group and thus would be ineligible for the payments. Partially as a result, very few working single mothers volunteered to be randomly assigned. As a consequence, the

[2] This problem can also occur if the experimental design calls for too few subjects. The designers may have failed to conduct a power analysis to be sure that the experiment has enough observations to produce statistically significant results, or the designers may have made wrong assumptions in their power analysis such as the standard deviation of the outcome variable.

[3] The Working Tax Credit Program is similar to the Earned Income Tax Credit in the U.S.

policy of not mentioning the incentive payments was reversed, and considerable additional effort was put into recruiting these persons. Although this recruitment effort was in large part successful, the size of the ultimate research sample was still well under what was planned (Walker, Hoggart, and Hamilton, 2006). This was probably due in part because many female family head with children who are already working part-time are resistant to full-time work given their childcare responsibilities.

Although the ERA demonstration recovered from its initial recruitment problems, other social experiments have not. For example, the Madison and Racine Quality Employment Experiment, which was targeted at women in the WIN (Work Incentive) program (the forerunner of the today's welfare-to-work programs), was ultimately aborted because of a combination of the small size of its registrant pool, which meant that the potential sample that could be recruited was inadequate, and its slowness in getting the program it was testing underway (Leiman, 1982). In the Illinois Career Advancement Project, a substantial number of individuals were randomly assigned, but fewer than nine percent of the experimental group actually participated in the treatment, partially because little was done to encourage participation beyond a letter informing those in the group that they were eligible for financial assistance for education programs. Although some analysis was completed, the program was terminated one year early because of inadequate participation (citation deleted for blind review).

2. Failure to Properly Randomize.

By its very essence, social experimentation depends on the characteristics of the treatment group and the control group being similar, differing only as a result of the treatment being tested or by chance. One reason this may not occur is because of improper randomization. This most often happens when those administering the treatment or those who are supposed to be randomly assigned have some control over the randomization process, rather than complete control residing in persons with no interest in who is assigned to the treatment or control group. Even seemingly foolproof methods, such as assigning every other person who walks through the door or every person whose social security number ends in an odd

number to the treatment group, can be manipulated.[4] However, improper random assignment may also occur through inadvertent administrative errors. When those administering the treatment do have some control over random assignment, it is obviously important for those evaluating an experiment to interview these persons to determine if the assignment was not entirely random, although this can also sometimes be detected by comparisons of the observed characteristics of the treatment group with those of the control group at the time of random assignment.

The New Orleans Homeless Substance Abusers Project provides an interesting example of staff subversion of the random assignment process. Only those substance abusers considered sufficiently motivated were placed on the selection list; those who did not appear sufficiently motivated were assigned to the control group. As a result, under one-third of those entering the two treatment groups were actually randomized. Consequently, the analysis was conducted using non-experimental selection bias correction techniques. Surprisingly, these corrections actually increased the estimated impact of the treatment (Devine, Brody, and Wright, 1997).

A failure to properly randomize is not always purposeful. In one of the two sites of the United Kingdom's Supportive Caseloading experiment, a large number of unemployment benefit claimants who were ineligible for the treatment were assigned, apparently inadvertently, to the treatment group but not to the control group (Birtwhistle, Barnes, and Looby, 1994). The Harbinger Mental Health Project provides a somewhat less grievous example of failure to randomize: the 100 members of the research sample who approached the hospital for treatment were randomly assigned, but the 21 members of the sample who were long-term residents of the hospital were assigned to the treatment and control groups on a nonrandom basis. Both sets of individuals were included in the impact analysis (citation deleted for blind review).

[4] For a good discussion of properly randomly assigning individuals, see Orr, 1999. Bruhn and McKenzie (2009) found that simple pure randomization may perform poorer than other methods such as pair-wise matching and stratification when sample size is less than 300.

3. Control Cross-Over.

Sometimes called "control contamination," control cross-over occurs when some members of the control group receive the treatment being tested that they are supposed to be denied. This obviously diminishes the estimated impact of the treatment. However, unless the cross-over is rampant, it is unlikely to result in the complete failure of the experiment. Moreover, Orr (1999) has suggested a simple correction exists that can be used when the proportion of the control group that crossed over is known to the evaluators.

In the Alternative Schools Demonstration, for example, 13 percent of the controls in one site and 39 percent of the controls in another site attended the alternative high schools that only members of the treatment group of high-risk youths were supposed to attend; however, the evaluators corrected for the resulting cross-over in estimating the impacts of the alternative schools (Dynarski and Wood, 1997). In Bolivia's School Facility Improvements experiment, an experiment in which schools, rather than individuals were randomized, some control schools received funds to improve their physical facilities, while only treatment schools were supposed to receive these funds. The evaluation attempted to address this cross-over issue by using an approach developed by Manski (1990) that puts bounds on the impact estimates (Newman, Pradham, and Rawlings, 2002). However, the evaluation of Denmark's Job Training Demonstration apparently did not correct for cross-overs, although almost one-quarter of the control group received job training that they were supposed to be denied, (citation deleted for blind review).

4. Adverse Publicity Resulting in Canceling an Experiment.

While rare, this has happened. For example, the New Deal for Disabled Persons (NDDP) was a voluntary welfare-to-work program for incapacity (disability) claimants in the United Kingdom. Original plans called for NDDP to be evaluated with a random assignment experimental design at the time it was introduced nationally. Although the effectiveness of the program was unproven, a decision was made shortly before NDDP was introduced to drop the planned experimental evaluation. The experiment was terminated largely as a result of concern over the denial of services to a control group of disabled persons.

Although an evaluation was conducted, it was non-experimental (Orr, Bell, and Lam, 2007). Due to adverse publicity, the random assignment Matriculation Awards Demonstration in Israeli high schools, in which entire schools were randomly assigned and cash awards for reaching achievement goals were offered to students at the treatment schools, was suspended after one year of a planned three-year experiment (Angrist and Lavy, 2002).[5]

5. Failure to Implement the Treatment Properly.

This is a potentially serious, although not usually ruinous, problem that has occurred in a number of social experiments. In such instances, the experiment does not test what it was designed to test. Implementation (or process) analysis that involves observation of the program being tested and interviews with staff administering the treatment and members of treatment groups can be critical to detecting whether the treatment was implemented as planned.

The failure to implement the designed treatment is well illustrated by the Quantum Opportunity Program Pilot and the Quantum Opportunity Program Demonstration. These were sequentially run, multi-site experiments, which were intended to test the effects of comprehensive services for high school students with a high probability of dropping out. Neither of the experiments implemented the full complement of planned services, although the extent of the deviations from the planned treatment varied among the sites. Indeed, one of the five sites in the first experiment completely failed to implement the program and subsequently was dropped from the evaluation analysis. In the later experiment, no site implemented the education or the community service components of the tested program as prescribed (Maxfield, Castner, Maralani, and Vencill, 2003; [citation deleted for blind review]).

Implementation problems plagued the Targeted Negative Income Tax demonstration, which was run for public assistance recipients in seven sites in Germany from 1999 to 2002. For example, in most,

[5] Angrist and Lavy (2002) do not indicate the reason why there was negative publicity, only that the program was presented to reporters as an attempt to increase scores on a test that is a pre-requisite for university admission, and "this led to extensive and mostly critical media coverage," resulting in the suspension of the program (p. 11, f.n. 8).

but not all sites, those eligible for the tested program, which was quite complex, were initially informed about the program by letter, with no further attempt at follow-up. Because of the implementation problems, no conclusions about impacts were possible in six of the seven sites.[6]

Sometimes the implemented treatment (in contrast to the planned treatment) does not differ sufficiently from the treatment provided to controls to result in a useful test. For instance, one goal of the San Diego Homeless Research demonstration was to compare traditional case management with comprehensive case management, which was supposed to have smaller caseloads and provide additional services. In practice, the differences between the two types of case management were minimal. Perhaps as a result, statistically significant differences in outcomes did not result (citation deleted for blind review). Something similar occurred in Britain's Intensive Gateway Trailblazers demonstration, which targeted young adults who had been unemployed for at least six months and who were receiving benefit payments. The mandatory tested program was supposed to require individuals assigned to the treatment group to participate in a course and to receive more intensive training and counseling than controls. In practice, the services actually received by the treatment and control groups were similar (Davies and Irving, 2000).

6. Failure to Adequately Communicate the Treatment.

If members of a treatment group are to respond to a treatment, they presumably must understand what the treatment is. In a sense, inadequate communication of the treatment to the treatment group is a type of implementation failure. In fact, as mentioned above, this was one of the problems with the German Targeted Negative Income Tax demonstration. However, it is not always evident that a lack of understanding of the treatment threatens the validity of an experiment. This is especially true of a demonstration program in which the same lack of understanding would exist were the tested program actually adopted.

[6] Information based on correspondence and discussions in 2003 and 2013with Alexander Spermann, one of the key persons who conducted the Targeted Negative Income Tax demonstrations.

A good example is a recent random assignment demonstration program run in a district in India. In this experiment, families were to bring their grain to local millers who fortified the resulting flour with iron at no additional cost to the families. This was intended to offset iron deficiency anemia that causes low productivity and health problems in much of the developing world. Although the millers did this in the early days of the experiments, they soon stopped, perhaps, in part, because of a misunderstanding on their part. As a result, while the anemia rate fell in the first part of the experiment, there was no difference in the rate by the end of the experiment (Banerjee, Duflo, and Glennerster, 2011).

In the Seattle-Denver Income Maintenance Experiment, a U.S. test of a negative income tax program, a survey was administered to determine the treatment group's understanding of the rather complex program being tested. The results indicated no more than a "moderate" understanding (SRI International, 1983, p. 34). This probably was not due to implementation failures because, in contrast to the German Targeted Negative Income Tax demonstration, a rather intense effort was made to educate program participants. For example, upon enrollment and a year after enrollment, participants were visited by a trained counselor who spent more than an hour describing the treatment and who provided tables that participants could use to determine their payments under the negative income tax. In addition, help in answering questions was also available at field offices throughout the experiment. Interestingly, accuracy on the survey depended on the experiences of the respondents—for example, accuracy tended to be greater among persons who had become unemployed, an event that affected their payments under the program. This result suggested to the evaluators that "people will find out about the effects of different behaviors when those behaviors or activities become relevant" (SRI International, 1983, p. 35). Moreover, inclusion of comprehension scores constructed from the survey responses in a regression model of labor supply (the key behavior being tested by the experiment) found no relation between the variable and labor supply (SRI International, 1983, p. 35).

The Primary Prevention Initiative, which was a test of school attendance, physical exam, and immunization requirements for the children of AFDC recipients, provides another illustration of a failure to understand the treatment. A telephone survey of over 200 members of the treatment group indicated

that over 80 percent of them could not identify either of the mandatory requirements of the program (Wilson, Stoker, and McGrath, 1999, Table 1). Knowledge among those who had been sanctioned through a grant reduction for failing to meet the requirements of the tested program was greater but only slightly so (Wilson, Stoker, and McGrath, 1999, Table 2). Not surprisingly, the impact of the program on school attendance and immunization was negligible. Thus, the PPI, as implemented, did not test the impact of the intervention on a knowledgeable population. If the state believes that the welfare population could be educated about the PPI rules, then the efficacy of the strategy cannot be ascertained from the experiment undertaken.

Implementation failure is not limited to complex treatments. Moreover, it is sometimes unavoidable. For example, the Arkansas Welfare Waiver demonstration tested the effects of subjecting AFDC recipients to a family cap in which their benefits would no longer increase with the birth of an additional child. The experiment was limited to 10 small rural counties, while all AFDC recipients in the remainder of the state were simply made subject to the family cap. At least partially due to this and the fact that the family cap was widely publicized, many controls in the 10 experimental sites believed they were subject to the cap, although they were not (Turturro, Benda, and Turney, 1997).

7. Inadequate Sample Size.

Although they usually receive less publicity than large experiments, many social experiments rely on small research samples. For example, drawing on a data base containing information on 143 social experiments that were completed between 1962 and 1996, Greenberg, Shroder, and Onstott (1999)[7] found that 18.6 percent had samples of fewer than 200,[8] 16.4 percent had samples ranging from 200 to 499; and 15.0 percent had samples ranging from 500 to 999. Because the impacts of the treatments tested in social experiments are often modest in magnitude, small experiments, especially those with samples of only a

[7] The authors believed that their database included most of the social experiments completed between 1962 and 1996.

[8] As used by Greenberg, Shroder, and Onstott (1999), "sample" includes both the treatment and control groups and usually refers to the number of individuals or households randomly assigned.

few hundred, tend to be underpowered. Consequently, even if the treatments produce true impacts, the estimated impacts are likely to be statistically insignificant.[9] For example, evaluators of the Tulsa Individual Development Account demonstration conducted a 10-year follow-up survey in which they located 855 of the households that were originally randomly assigned (Grinstein-Weiss, et al., 2012). Almost all the impacts estimated with the resulting data were statistically insignificant. However, this can be mitigated to some extent by using a regression framework to account for baseline covariates correlated with the outcome of interest. For example, in their analysis of a recently fielded experiment with 106 persons in the treatment group and 130 individuals in the control group, Cook, O'Brien, Braga, and Ludwig (2012) note that in the absence of accounting for covariates, their minimum detectable effect size would be .37 standard deviations; but if the covariates can account for one-third of the variance in the outcome, this would fall to .30 standard deviations.

There are several reasons why sample sizes are small in some social experiments, and, as a result, detecting the impacts of the tested treatment may be unsuccessful. First, as previously discussed, when participation in an experiment is voluntary, relatively few persons may volunteer. Second, as discussed further below, some of the original sample may be lost when follow-up data are collected.[10] Third, and perhaps most importantly, sample size may be constrained because of budgetary considerations. The cost of both administering treatments and collecting follow-up data often increase as sample size increases. To minimize these costs, experiments with very large samples—14.3 percent of the 143 experiments mentioned in the previous paragraph had samples of over 10,000—typically rely on existing automated administrative databases and often test modest incremental changes in existing programs.

If social experiments were solely intended to provide information on program impacts, it would not be obvious why most small, underpowered experiments have been undertaken. After all, power tests can be, and often are, conducted before experiments are initiated. However, there may sometimes be

[9] This is not surprising because the ability of the data to detect even a moderate true impact is very weak at even a relatively low level of statistical significance. For example, at an alpha of 0.100 and a 5 percentage point impact at a control mean of .5, the power is about .4.

[10] This does not affect the figures mentioned in the previous paragraph, because they pertain to the number of individuals initially randomly assigned.

political reasons for undertaking experiments with small samples—for example, delaying a decision on a policy change or responding to pressure to give the impression of scientifically testing a change that has already been decided upon. Or perhaps those designing some small experiments are overly optimistic about the size of the effect the treatment will produce.

There are, in fact, a few small experiments that have resulted in large, statistically significant impacts and, as a result, have been influential. The large, statistically significant findings could have resulted because the treatments produced larger impacts than most social programs that have been tested experimentally or, possibly, because the findings were spurious. One example is several experiments that relied on small samples to test job clubs. These experiments, which involved supervised job-search activities in a group, resulted in large impacts and wide-spread adoption of job clubs. Because the various replications of the first job club experiment consistently produced positive and statistically significant impact estimates, there was reason to have some confidence in the findings.[11] Another example is the evaluation of the Perry Preschool Program, which was based on a program group of 58 children and a control group of 65 children (Cook and Wong, 2007); however, as a consequence of its large, favorable, and often statistically significant impacts, this demonstration has been influential in promoting early-intervention programs targeted at young children. Although the small sample is troubling,[12] its exceptionally large estimated impacts were of sufficient size to often be statistically significant.

8. Sample Attrition.

Sample attrition occurs when some of those who are randomly assigned are unavailable when follow-up data are collected. This is most likely to occur when the follow-up data are obtained through surveys, rather than through administrative records. To take a rather extreme, but important, example, in the Food Stamp Employment and Training Program demonstration, 12-month survey data were

[11] See [citation deleted for blind review} for a summary.

[12] For example, the statistically significant 18 percentage point increase in high school graduation or high school equivalency, which was found for the Perry program, occurred because just seven more members of the treatment sample graduated from high school or received high school equivalency than members of the control sample.

successfully collected for only 50 percent of the research sample of 13,086. This was attributable to difficulties in obtaining addresses from local Food Stamp offices, high mobility among the sample population, and the large number of homeless persons who were randomly assigned (Puma, Burstein, Merrell, and Silverstein, 1990).

Administrative data are not completely immune to sample attrition. For example, many social experiments involving the welfare population have relied on the records of state welfare agencies and earnings data reported by employers to state agencies administering the unemployment insurance system. The unemployment insurance records do not include individuals who move out of the state in which the experiment was conducted and individuals who are self-employed. Moreover, attrition of this sort can systematically differ between the treatment and control group—for example, some persons in the treatment group may respond to the program being tested by moving out-of-state in order to obtain employment.

A rather rare, but instructive, instance of probable non-comparability of treatment and control groups in an experiment that relied on administrative data occurred in the Wisconsin's Self-Sufficiency First/Pay for Performance Program (SSF/PFP). In this test of a mandatory welfare-to-work program, AFDC applicants who were assigned to the treatment group were required to participate in both the SSF and PFP components of the tested program, while persons who were already in the AFDC system at the beginning of the experiment were required to participate in only the PFP component. A data system that was being developed at the time of random assignment allowed the staff administering the AFDC (now TANF) program to exempt some AFDC applicants who were assigned to the treatment group from SSF. Unfortunately, those who were exempted disappeared from the data available for analysis. Thus, the evaluators had confidence that the non-comparability problem was minimal for active AFDC recipients, but not for AFDC applicants. As they explain,[13]

[13] Cancian, Kaplan, and Rothe (2000). The evaluators were hired after the experiment was completed. The firm originally employed to evaluate the experiment was terminated.

> …we are concerned that while initial assignment was random, the ultimate placement of [AFDC applicants] into the control or experimental groups may have been nonrandom. In particular, there is some evidence that cases with the greatest barriers to employment may have been exempted from SSF/PFP and deleted from the data set, whereas similar cases were not exempted from the control group and were retained in the data set. Thus, differences in outcomes may be attributed to: (1) the impact of the SSF/PFP programs, *or* [italics in original] (2) differences in the characteristics of individuals assigned to the groups.

The experimental findings for the active AFDC cases appear in the text of the evaluation report, while those for AFDC applicants are discussed in an appendix that also prominently includes the warning quoted above.

However, the sort of attrition that occurred in Wisconsin's Self-Sufficiency First/Pay for Performance Program (SSF/PFP) is somewhat unusual. Attrition is usually smaller and less serious with administrative data than with survey data.

Two problems result from sample attrition. First, as mentioned above, the sample size is reduced, sometimes greatly reducing the power of the data to detect impacts resulting from the treatment. For example, the Project Hope demonstration began with a small sample of 140. In an attempted follow-up survey of 116 of these individuals, only 24 were ultimately interviewed, partially as a result of 55 telephone numbers having been disconnected (Office of Community Services, 1992). Similarly, in the Partnership for Hope demonstration, only 58 persons of the 109 individuals randomly assigned returned a questionnaire that was mailed out at the end of the experimental treatment (Office of Community Services, 1994).

The other problem with sample attrition, and often the more serious one, is that it is unlikely to be random. This can make the remaining sample unrepresentative of the group originally included in the research sample. Moreover, those who attrite from the treatment group may differ from those who attrite

from the control group, often in ways that are not observable but yet are associated with the outcomes of interest. This causes response bias in the impact estimates.[14]

One recent example of response bias is the United Kingdom's Employment Retention and Advancement demonstration, where both survey data and administrative data were collected five years after random assignment for two of the program's three target groups. The survey response rate was 62 percent for one of these groups and 69 percent for the other. The sample size was sufficiently large for both groups that lack of power to detect outcomes was not a serious problem even after attrition. However, the possibility of response bias was a concern. Fortunately, it was possible to examine the survey data for possible response bias by using the administrative data to compare the earnings impacts for those who responded to the survey with the earnings impacts for the full sample. It turned out that the earnings impacts were markedly larger for the survey respondents than for the full sample, strongly suggesting the presence of response bias (Hendra et al., 2011). Because some of the key outcomes, such as earnings, employment status, and some government benefit payments, were available from the administrative data, as well as from the survey data, the estimates of impacts from the former could be emphasized in reporting findings from the experiment. However, data on other outcomes, such as health status, wage rates, and hours, were only available from the survey data. Thus, it was not possible to determine whether estimates of impacts on these outcomes were subject to response bias or to provide alternative estimates of impacts based on these outcomes.

Even when sample attrition is relatively low, response bias can still result. For example, the evaluators of the Tulsa Individual Development Account (IDA) demonstration, which attempted to encourage households to accumulate assets by subsidizing home purchases, home repairs, post-secondary education, business investments, and retirement savings, collected survey data around 10 years after random assignment. Rather remarkably, 855 of the 1,103 individuals originally randomly assigned were

[14] There is another type of response bias that arises because respondents give the wrong answers either deliberately or because of recall problems or cognitive issues. In this paper, however, we use the term "response bias" to refer to biases caused because survey respondents in treatment groups differ from respondents in control groups in ways that cannot be adjusted through statistical means..

located and included in the 10-year survey. To determine whether there are observable differences between respondents in the 10-year treatment and control groups that are not attributable to the experimental treatment, the evaluators compared their characteristics at the time of random assignment. When this was done, some small observed differences in the characteristics of treatment and control group 10-year respondents became apparent. For example, home ownership was one key outcome. As it turned out, respondents in the control group were more likely to own homes at the time of random assignment than members of the treatment group, although this difference is not statistically significant at conventional levels. Unfortunately, unlike the U.K. Employment Retention and Advancement demonstration, administrative data were not available to substitute for the survey data. Thus, the evaluators used a variety of methods to attempt to control for response biases including regression adjustments, propensity score matching, and difference-in-difference analysis (see Grinstein-Weiss et al. 2012 for details). Consequently, the analysis was necessarily non-experimental.

9. Changes in the Environment.

Unlike laboratory experiments, the environment surrounding social experiments cannot be controlled while the experiment is undertaken. Instead, the environment within the institution in which an experiment is taking place (e.g., a school, welfare office, police station, or hospital) or the external environment (e.g., the state of the economy or programs affecting the target population) may change, and this may affect findings from the experiment. Environmental changes differ from the flaws listed above because they do not result from the implementation or design of an experiment. Nonetheless, they are included in the discussion because they sometimes represent an important challenge to those conducting experiments.

As an example of a change in the environment internal to experiments, consider Boruch, Merlino, and Porter's (2012) discussion of the substantial turnover among teachers that occurred in a number of experiments in school systems—for example, 42 percent of the teachers left their jobs within about a year after a randomized trial in which they were involved began. They assert that such "churning" weakens

the ability of experiments to detect impacts on student learning, both because students learn less under unstable conditions and because teachers in a new position are so busy acquiring job-specific human capital that they cannot focus on implementing the treatment being tested by the experiment. They further suggest that the ability of experiments to detect impacts in other sorts of institutions is similarly weakened by job churning. This may well be true. However, if turnover does not differ between the treatment and control groups (and in the examples mentioned by Boruch, Merlino, and Porter (2012), it apparently does not) and the turnover is typical of that which would occur in the absence of the experiments, the resulting impact estimates would seem to be valid measures of how well the tested intervention would perform if it was introduced on a permanent basis.

Changes to the environment that are external to experiments may be more damaging to experimental integrity. For example, when the New Jersey Income Maintenance Experiment was designed, AFDC in New Jersey did not cover two-parent households; but the program was unexpectedly modified to include such households as the experiment was being implemented, greatly attenuating the treatment difference between the treatment and control groups.

10. Stakeholder Resistance

Resistance by various stakeholders to certain aspects of experiments can force changes in the program being tested or the experimental design that weakens an experiment. Sometimes it may even result in the experiment being cancelled. Special efforts during the design phase on the part of those conducting the experiment may be required to overcome stakeholder resistance. Stakeholder resistance is especially likely in attempting to conduct experiments on already existing programs. For example, the National JTPA Study, which was a random assignment evaluation of training programs for the disadvantaged funded under the Job Training Partnership Act (JTPA), would have ideally randomly selected sites for participation in the study, but instead was limited to those self-selected sites that were willing to participate (Orr et al., 1996). More direly, in 2005 the U.S. Department of Labor planned and designed a random assignment evaluation of the Youth Offender Demonstration Project. The experiment

was to take place in six jurisdictions that were already operating the program. The random assignment design required that courts allow youths to be randomly assigned to one of three groups (two treatment groups and a control group). As it turned out, despite the efforts of the evaluation team, none of the courts or programs in the six jurisdictions was willing to accept the experimental design because of changes in the programs that would have resulted and because of various ethical, legal, and political issues that random assignment raised. As a result, the experiment was not conducted (Dunham, Wiegand, and Michalopoulos, 2008).

CONCLUSIONS

In this paper, we have examined 10 flaws that can occur in conducting social experiments, sometimes causing the experiment to fail. Table 1 lists the major lessons from this effort. The most important of the flaws are response bias resulting from attrition; a failure to adequately implement the treatment as designed; and too small a sample to detect impacts. The third of these flaws can result from insufficient marketing, too small an initial target group, disinterest on the part of the target group in participating (if the treatment is voluntary), or attrition.

The discussion of experimental flaws is in no way intended to discourage the use social experiments for evaluating social programs. It is often the best tool available. However, the discussion demonstrates that without due care and sufficient funding, experiments can face major obstacles. To a considerable extent, these can generally be minimized. For instance, implementation failures and too small a sample can usually be avoided with sufficient effort and planning and response bias can often be mitigated—for example, through increased follow-up efforts in conducting surveys.

Table 1 Summary of Lessons from Experimental Flaws			
Problem	**When It Occurs**	**Seriousness**	**Approaches for Addressing the Problem**
Too small a sample due to--			Power tests should be conducted prior to implementing the experiment to determine if this is likely.
Insufficient marketing	Test of voluntary treatment when pre-implementation planning is insufficient.	Potential failure to detect impacts.	Increase outreach and marketing effort.
Target group too small	Pre-implementation research insufficient.	Potential failure to detect impacts.	Possibly terminate experiment.
Disinterest in participating	Test of voluntary treatment.	Potential failure to detect impacts, but useful information still provided.	Possibly increase communication with target group.
Budgetary constraint	Pre-implementation planning insufficient.	Potential failure to detect impacts.	Consider not undertaking experiment.
Sample attrition	More likely when survey data are used.	Potential failure to detect impacts.	Increased effort at survey follow-up.
Response bias due to attrition	Much more likely when survey data are used.	Serious but not necessarily fatal.	When available, use baseline or administrative data to detect. Increase effort at survey follow-up. Conduct non-experimental analysis with statistical correction of selection bias.
Improper randomization	When those administering or subject to the treatment have some control or randomization, but sometimes done inadvertently.	Serious but not necessarily fatal.	When possible, compare treatment and control characteristics at baseline to detect. Conduct non-experimental analysis with statistical correction of selection bias.
Control cross-over	When treatment is attractive and administrators fail to prevent controls from receiving it.	Not too serious, if not too large.	Use implementation analysis to detect. Use the Orr cross-over correction.
Adverse publicity	When treatment is attractive and preventing controls from receiving it is controversial.	Can cause shut-down of experiment, but this is rare.	Improved public relation may help, but there may be no solution.
Failure in implementing treatment	Usually occurs with demonstration programs; budget inadequate or administers resistant to aspects of the treatment.	Depends on degree to which actual treatment deviates from planned treatment.	Use implementation analysis to detect. Possibly hold discussions with those implementing the treatment.
Inadequate communication of treatment	When treatment is complex and/or effort to explain treatment is insufficient. Sometimes difficult to avoid.	Not necessarily serious if lack of understanding in demonstration program and implemented program are similar.	Use implementation analysis to detect. Increase the effort at communication with the treatment group when communication has been inadequate.
Changes in the environment	Can occur at any time during experiment.	Can be serious to experimental integrity if external to the experiment.	Those running the experiment usually have little control over environmental changes
Stakeholder resistance	Especially likely when existing program being evaluated.	May cause experiment to be modified in unfavorable ways or not initiated.	Gain agreement to random assignment by key stakeholders as part of the design effort.

References

Angrist, J.D. and Lavy, V. (2002). *The effect of high school matriculation awards: Evidence from randomized trials*, Cambridge MA: National Bureau of Economic Research working paper 9389.

Banerjee, A., Duflo, E., and Glennerster, R. (2011). Is decentralized iron fortification a feasible option to fight anemia among the poorest? *In Explorations in the economics of aging*, David A. Wise (Ed.), Chicago: University of Chicago Press, 317 – 344.

Birtwhistle, A., Barnes, D. and Looby, C. (1994). *Evaluation of supportive caseloading (1-2-1) in North Norfolk.* Sheffield, U.K.: Research and Evaluation, Employment Service, Tracking Study.

Boruch, R., Merlino, J., and Porter, A. (2012). Golfing in a hurricane: education system instability, randomized controlled trials, and children's achievement, unpublished paper.

Bruhn, M. and McKenzie, D. (2009). In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics*, 1, 200-232.

Cancian, M., Kaplan, T., and Rothe, I. (2000). *Wisconsin's self-sufficiency/pay for performance: Results and lessons from a social experiment.* Madison: University of Wisconsin-Madison, Institute for Research on Poverty.

Cook, T. D. and Wong, V. (2007). The Warrant for universal pre-K: Can several thin reeds make a strong policy boat? *Social Policy Review*, 21(3), 14-15.

Cook, P. J., O'Brien, M., Braga, A., and Ludwig, J. (2012). Lessons from a partially controlled field trial. *Journal of Experimental Criminology*, 8 (3), 271-287.

Davies, V. and Irving, P. (2000). *New Deal for Young People: Intensive Gateway Trailblazers. A final report to the Employment Services*. Birmingham, U.K.: ECOTEC Research and Consulting Limited.

Devine, J. A., Wright, J. D., and Brody, C. J. (1997). Evaluating an alcohol and drug treatment program for the homeless: An econometric approach, *Evaluation and Program Planning*, 20, 205-215.

Dunham, K., Wiegand, A., and Michalopoulos, C. (2008). *The effort to implement the Youth Offender Demonstration Project (YODP) impact evaluation: Lessons and implications for future research, final report*. Ottawa Ontario: Social Policy Research Associates.

Dynarski, M. and Wood, R. (1997). *Helping high-risk youths; results from the Alternative Schools Demonstration Program*. Princeton, NJ: Mathematica Policy Research.

Greenberg, D., Shroder, M., and Onstott, M. (1999). The social experiment market, *Journal of Economic Perspectives*, 13 (3), 157-172.

Grinstein-Weiss, M., Sherraden, M. W., Gale, W. G., Rohe, W., Schreiner, M., and Key, C. (2012). *Long-term follow-up of individual development accounts: Evidence from the ADD experiment*. Chapel Hill: The University of North Carolina.

Harrison, G. W. and List, J. A. (2004). Field experiments. *Journal of Economic Literature*. 42, 1009-1055.

Hendra, R. et al. (2011). *Breaking the low-pay, no-pay cycle: Final evidence from the UK employment retention and advancement (ERA) demonstration.* Sheffield: Department for Work and Pensions Research Report No 765.

Leiman, J. M. (1982). *The WIN Labs: A federal/local partnership in social research*. New York: MDRC.

List, J. A. (2011). Why economists should conduct field experiments and 14 tips for pulling one off. *The Journal of Economic Perspectives*. 25, 3-15.

Manski, C. (1990). Nonparametric bounds on treatment effects. *American Economic Review*. 80 (2), 319-323.

Maxfield, M., Castner, L., Maralani, V., and Vencill, M. (2003). *The Quantum Opportunity Program demonstration: Implementation findings*, Princeton, NJ: Mathematica Policy Research.

Newman, J., Pradham, M. and Rawlings, L. B. (2002). An impact evaluation of education, health, and water supply investments by the Bolivian Social Investment Fund. *The World Bank Economic Review*, 16 (2), 241-71.

Orr, L.L., Bloom, H.S., Bell, S.H., Doolittle, F., Lin, W., and Cave, G. (1996). *Does training for the disadvantaged work? Evidence from the National JTPA Study*. Washington, DC: Urban Institute Press.

Orr, L. (1999). *Social experiments: Evaluating public programs with experimental methods*, Thousand Oaks, CA: Sage Publications.

Orr, L. L., Bell, S., & Lam, K. (2007). *Long-term impacts of the new deal for disabled people: Final report.* Leeds: DWP Research Report No. 342.

Puma, M.J., Burstein, N.R., Merrell, K., and Silverstein, G. (1990). *Evaluation of the Food Stamp Employment and Training Program: Final Report*. Volume I: Washington, D.C.: U.S. Department of Agriculture, Food and Nutrition Service.

SRI International (1983). *Final report of the Seattle-Denver Income Maintenance Experiment*, Volume 1. Washington, D.C.: U.S. Department of Health and Human Services.

Turturro, C., Benda, B., and Turney, H. (1997). *Arkansas Welfare Waiver Demonstration Final Report*. Fayetteville: University of Arkansas.

U.S. Department of Health and Human Services (1992). *Demonstration Partnership Programs Project: Summary of final evaluation findings from FY1989.* Washington, D.C.: Office of Community Services, Monograph Series 100-89: Case Management Family Intervention Models.

U.S. Department of Health and Human Services (1994). Partnership for Hope, Whatcom County Opportunity Council, Bellingham, Washington. *Summary of final evaluation findings from FY 1990, Demonstration Partnership Project.* Washington, D.C.: Office of Community Services, Monograph Series 100-90: Case Management Family Intervention Models, 100-4-72—110-4-83.

Walker, R., Hoggart, L., and Hamilton, G. (2006). *Making random assignment happen: Evidence from the UK employment retention and advancement (ERA) demonstration*. London: Department for Work and Pensions Research Report 330.

Wilson, L. A., Stoker, R. P., & McGrath, D. (1999). Welfare bureaus as moral tutors: What do clients learn from paternalistic welfare reforms? *Social Science Quarterly*, 80 (3), pp. 473-486.