FLAWED SOCIAL EXPERIMENTS

David Greenberg and Burt Barnow

University of Maryland, Baltimore County

5/29/2012

FLAWED SOCIAL EXPERIMENTS

Almost all social experiments 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. There is obviously a continuum between a minor flaw and a fatal failure. In this paper, we examine serious experimental flaws, but not all of these 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 failure, 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, and users of the experimental findings must exercise great caution.

What causes serious flaws in social experiments? We consider eight, somewhat overlapping, potential sources and illustrate each with examples from previous experiments.[1]

1. *Few show up*. This problem occurs when some potential participants in the program being tested are available for random assignment, but many less than planned for in the research design, resulting in an insufficient sample for hypothesis testing. 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 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,[2] 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

[1] Many, but not all, of these examples are taken from David Greenberg and Mark Shroder (2004), *The Digest of Social Experiments*, third edition, Washington, D.C.: Urban Institute Press.
[2] The Working Tax Credit Program is similar to the Earned Income Tax Credit in the U.S.

their disappointment if they were randomly assigned to the control group and thus were ineligible for the payments. Partially as a result, very few working single mothers volunteered to be randomly assigned. As a consequence, the 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 and was almost entirely located in one of the demonstration's six sites.[3] 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 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 inadequate, and its slowness in getting the program it was testing underway (Greenberg and Shroder, 2004, pp. 35-36). 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 (Greenberg and Shroder, 2004, pp. 78-79). Finally, the Gary (Indiana) Income Maintenance Experiment successfully tested a negative income tax program, but the intention was to test two other treatments as well: a social service access worker (a personal ombudsman) and the expansion of day care services in one neighborhood in Gary. Both services were undersubscribed and subsequently discontinued. The Gary experiment experience at least provided useful information about likely interest in such services (Greenberg and Shroder, 2004, pp. 201-202).

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 alone. One reason this may not occur is because of improper randomization. This most often happens when those administering the treatment or those who are

[3] See Robert Walker, Lesley Hoggart, and Gayle Hamilton (2006), *Making Random Assignment Happen: Evidence from the UK Employment Retention and Advancement (ERA) Demonstration*, London: Department for Work and Pensions Research Report 330.

suppose 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 at the time of random assignment with those of the control group.

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, well under one-third of those assigned to the treatment groups were actually randomized. Consequently, the analysis was conducted using non-experimental selection bias correction techniques. Ironically, these corrections actually increased the estimated impact of the treatment, implying that the staff selected those less likely to benefit from the tested treatment for the treatment group (Greenberg and Shroder, 2004, pp. 336-338).

Another experiment, Wisconsin's Self-Sufficiency First/Pay for Performance Program (SSF/PFP), resulted in a probable non-comparability of the treatment and control groups. 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 only required to participate in 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

[4] For good discussions of properly randomly assigning individuals, see Larry Orr (1999), *Social Experiments: Evaluating Public Programs with Experimental Methods*, Thousand Oaks, CA: Sage Publications, and Annex 2 of Stephen Morris et al. (2002), *Designing a Demonstration Project: An Employment Retention and Advancement Demonstration for Great Britain*, London: Cabinet Office, Government Chief Social Researcher's Office.

the non-comparability problem was minimal for active AFDC recipients, but not for AFDC applicants. As they explain,[5]

> "…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.

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 (Greenberg and Shroder, 2004, pp. 445-446). The Harbinger Mental Health Project provides a 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 (Greenberg and Shroder, 2004, pp. 378-379).

3. *Control cross-over*. Sometimes called "control contamination," this occurs when some members of the control group receive the treatment being tested that they are suppose 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, Larry Orr (1999)[6] has suggested the following simple correction that can be used when the proportion of the control group that crossed over is known to the evaluators:

---

[5] Maria Cancian, Thomas Kaplan, and Ingrid Rothe (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. The evaluators were hired after the experiment was completed. The firm originally employed to evaluate the experiment was terminated.

[6] Larry L. Orr (1999), *Social Experiments: Evaluating Public Programs with Experimental Methods*, Thousand Oaks, CA: Sage Publications.

$I_c = pI_u/(p\text{-}c)$,

where I_c is the corrected impact estimate, I_u is the uncorrected estimate, c is the proportion of the control group that crossed over, and p is the proportion of the treatment group that participated in the treatment. Under this formulation, it is assumed that those who cross-over receive the same treatment as those in the treatment group who participated in the treatment.

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 suppose to attend; however, the evaluators corrected for the resulting cross-over in estimating the impacts of the alternative schools (Greenberg and Shroder, 2004, pp. 267-269). 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 suppose to receive these funds. The evaluation took this cross-over into account, although it used a different method than the one suggested by Orr (1999) (Greenberg and Shroder, 2004, pp. 434-436). However, the evaluation of Denmark's Job Training Demonstration did not correct for cross-over, although almost one quarter of the control group received job training that they were supposed to be denied, (Greenberg and Shroder, 2004, pp. 423-424).

4. *Adverse publicity resulting in canceling an experiment*. While rare, this has happened. For example, the New Deal for Disabled Persons (NDDP), which was a voluntary welfare-to-work program for incapacity claimants in the United Kingdom, was originally supposed to be tested as a random assignment experiment. Although the effectiveness of the program was unproven, the program was nonetheless introduced nationally, but without the planned experimental evaluation. The experiment was largely terminated 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.[7] Due to adverse publicity, the random assignment Matriculation Awards Demonstration in Israeli high schools was suspended after one year of a planned three (Greenberg and Shroder, 2004, pp. 406-407)].

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

[7] Larry L. Orr, Stephen Bell, and Ken Lam (2007), *Long-Term Impacts of the New Deal for Disabled People: Final Report*, Leeds: DWP Research Report No. 342.

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, implementation was so weak in one site in the first experiment that it was ultimately dropped. In the later experiment, no site implemented the education or the community service components of the tested program as prescribed (Greenberg and Shroder, 2004, pp. 270-271 and 282-284).

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, 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 (Greenberg and Shroder, 2004, pp. 428-429).

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 suppose 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 also did not result (Greenberg and Shroder, 2004, pp. 335-336). 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 suppose 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. For example, there was difficulty in securing attendance in the mandatory course (Greenberg and Shroder, 2004, pp. 452-453).

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 way, 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 if the same lack of understanding would exist were the tested program was actually adopted.

   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).[8] 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 they could use to determine 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 and immunization requirements for the children of AFDC recipients, provides another illustration of when a failure to understand the treatment does not necessary affect the validity of an experimental test of a policy. 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).[9] Knowledge among those who had been sanctioned through a grant

[8] SRI International (1983), *Final Report of the Seattle-Denver Income Maintenance Experiment*, Volume 1, Washington, D.C.: U.S. Department of Health and Human Service.
[9] Laura A. Wilson, Robert P. Stoker, and Dennis McGrath (1999), "Welfare Bureaus as Moral Tutors: What Do Clients Learn from Paternalistic Welfare Reforms?" *Social Science Quarterly*, 80 (3), pp. 473-486.

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. Yet, the ultimate compliance with the programs requirements was over 90 percent, which meant that almost all the AFDC recipients facing the mandate were able to document to their caseworkers that they were in compliance with it (Wilson, Stoker, and McGrath, 1999, p. 484). Thus, even without knowledge of the mandated requirements, they understood the documentation they had to provide to continue receiving benefits.

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 ten 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 ten experimental sites believed they were subject to the cap, although they were not (Greenberg and Shroder, 2004, pp. 124-125).

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)[10] found that 18.6 percent had samples of fewer than 200,[11] 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 few hundred, tend to be underpowered. Consequently, even if the treatments produce true impacts, the estimated impacts are likely to be statistically insignificant.[12]

[10] David Greenberg, Mark Shroder, and Matthew Onstott (1999), "The Social Experiment Market," *Journal of Economic Perspectives*, 13 (3), pp. 157-172. The authors believed that their database included most of the social experiments completed between 1962 and 1996.

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

[12] For example, evaluators of the Tulsa Individual Development Account demonstration conducted a 10-year 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. 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.

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.[13] 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. 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 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. Indeed, a few small experiments that have resulted in large impacts have been influential. For example, the evaluation of the Perry Preschool Program was based on a program group of 58 children and a control group of 65 children (Greenberg and Shroder, 2004, pp. 229-231), but as a consequence of its large, favorable, and often statistically significant impacts it has been influential in promoting early-intervention programs targeted at young children. Although the small sample is somewhat troubling,[14] the experiment was not underpowered due to its exceptionally large estimated impacts.

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 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

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

[14] 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 than members of the control sample.

mobility among the sample population, and the large number of homeless persons who were randomly assigned (Greenberg and Shroder, 2004, pp. 213-214).

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 compensation system. These records do not include individuals who move out of the state in which the experiment was conducted. However, attrition is usually smaller 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 (Greenberg and Shroder, 2004, pp. 213-214). 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 (Greenberg and Shroder, 2004, pp. 99-100).

The other problem with sample attrition, and often the more serious one, is that it is unlikely to be random. In particular, 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.

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 tested 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 larger for the survey respondents than for the full sample, strongly suggesting the presence of response bias.[15] Fortunately, because many of the

[15] Richard Hendra 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.

key outcomes 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 some 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 ten years after random assignment. Rather remarkably, 855 of the 1,103 individuals originally randomly assigned were 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.

This paper examined eight major flaws in implementing social experiments. Table X lists the major lessons from this effort.

Table X 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	Much 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.