Analysis

SOCIAL EXPERIMENTATION:
EVALUATING PUBLIC PROGRAMS WITH EXPERIMENTAL METHODS **PART 6**

> *If the data were perfect, collected from well-designed randomized experiments, there would be hardly room for a separate field of econometrics.*
>
> — Zvi Griliches, President
> American Econometric Association,
> 1993–94

One of the great strengths of social experiments is the simplicity with which their results can be analyzed. In properly designed and implemented experiments, simple comparisons of the mean outcomes of the treatment and control groups provide unbiased estimates of the effects of the experimental intervention. Complex econometrics is not required to ensure the internal validity of the estimates.

This is not to say, however, that social experiments cannot provide a wide array of useful analyses. In this paper, the sixth in a series on the design, implementation and analysis of social experiments, we discuss the rich variety of ways in which experimental data can be analyzed to inform the policy process. Specifically, we discuss:

- The basic impact model;
- Estimating the time path of impacts;
- Estimating impacts on subgroups;
- Estimating and explaining variations in impact across sites;
- Dealing with potential biases; and,
- Analyzing the social benefits and costs of the program.

The Basic Impact Model

As noted above, the difference in mean outcomes between the treatment and control groups is an unbiased estimate of the impact of the experimental treatment. That estimate is, however, subject to sampling variability that arises from the random assignment of sample members to the two experimental groups. Outcomes depend not only on the experimental treatment, but also on the characteristics of individual sample members. In a test of a remedial education program, for example, students' grade point averages in the follow-up period will depend not only on whether they were assigned to the experimental program, but also on such individual factors as native ability, the quality of the student's previous education, and his or her home environment.

Random assignment guarantees that the treatment and control groups do not differ *systematically* in these characteristics, but it does not guarantee that they are *identical* in these dimensions. Therefore, random differences in the characteristics of the treatment and control groups, depending on the specific individuals assigned to each, contribute to the sampling variability of the impact estimate. Taking account of these differences will reduce this sampling variability and improve the power of the design—*i.e.*, it will allow us to detect somewhat smaller program effects than are detectable in a simple difference-in-means analysis.

The simplest way to "control for" or "hold constant" differences in the characteristics of the treatment and control groups is to use multivariate regression analysis to estimate a model of the form:

$$Y_i = a + \sum b_j X_{ji} + cT_i + e_i \tag{1}$$

where:

Y_i = the outcome measure of interest (*e.g.*, grade point average) for the ***i***th individual;

X_{ji} = a set of ***j*** individual background characteristics (*e.g.*, age, gender, past educational performance);

T_i = treatment status indicator (a dummy variable equal to 1 if the ***i***th individual is a member of the treatment group and 0 if he or she is a control); and,

e_i = a random error term.

In this model, the coefficient *c* estimates the impact of the program on this outcome, holding constant the **covariates** included in the set of personal characteristics X_{ji}. This estimate is called the **regression-adjusted impact estimate**. The set of coefficients b_j measures the effects of the various background characteristics on the outcome measure.[1]

The random error term, e_i, reflects the effects of all the individual characteristics and environmental factors not explicitly included in the model. The variance of the error term, $V(e_i)$, is related to the variance of the outcome Y_i according to the relationship:

$$V(e_i) = V(Y_i)(1 - R^2) \tag{2}$$

where R^2 is the proportion of the variance of the outcome *Y* explained by the regression equation.

The variance of the impact estimate *c* is proportional to $V(e_i)$:

$$\begin{aligned} V(c) &= V(e_i)\left(\frac{n}{n_t n_c}\right) \\ &= V(Y_i)(1 - R^2)\left(\frac{n}{n_t n_c}\right) \end{aligned} \tag{3}$$

where n, n_t, and n_c are the total sample size and the number of treatment and control observations, respectively. Controlling for more individual characteristics increases the R^2 of the regression, thereby reducing the variances of e_i and the impact estimate. This, in turn, reduces the minimum effect size that can be detected with a given sample.[2]

Suppose, for example, that in the example of the compensatory education program we are able to explain 20 percent of the variance of students' grade point averages in the follow-up period on the basis of student attributes not related to program participation. The variance of the regression-adjusted impact estimate will then be 20 percent smaller than the variance of the simple difference-in-means estimator.[3] This translates into an 11 percent reduction in the standard error of the estimate and, therefore, an 11 percent reduction in the minimum detectable effect. This is fairly typical of the gains in precision attainable by controlling for the baseline characteristics of the sample.

While an 11 percent reduction in minimum detectable effects may seem relatively small, it is equivalent to the gain in power associated with a 25 percent increase in sample size. Therefore, if it would be less costly to collect baseline data on characteristics of the sample than to increase the sample by 25 percent (as is nearly always the case), collecting baseline data on covariates would be a cost-effective investment of research funds.

In most applications, the single baseline characteristic that provides the greatest explanatory power, and therefore the greatest reduction in minimum detectable effects, is the preprogram value of the outcome variable itself. In the case of a compensatory education program, for instance, we would expect grade point average in the year prior to pro-

[1] These coefficients are generally not of interest in program evaluation, for two reasons. First, they reflect *natural variation* in the outcome, not program-induced effects. And second, because of correlations among the covariates, and between the covariates and variables not included in the impact model, as measures of the causal effects of these characteristics they are subject to a number of biases. The purpose of the covariates is simply to reduce the sampling variability of the impact estimates.

[2] See the fourth paper in this series for the derivation of the minimum detectable effect.

[3] This can be seen from equation 3. The difference-in-means estimator is equivalent to a regression model that includes only the treatment status indicator—*i.e.*, a model with no covariates. As shown in equation 3, under the null hypothesis of no impact such a regression would have an R^2 of zero and the variance of *c* would be $V(Y_i)(n/n_t n_c)$. Adding sufficient covariates to achieve an R^2 of .20 would reduce the variance of *c* to $.8V(Y_i)(n/n_t n_c)$.

gram entry to explain much of the variation across students in post-program grade point average. This is because the preprogram value incorporates the effects of all of the individual characteristics that influence grade point average; to the extent that these characteristics do not vary over time, differences in preprogram values of the outcome across the sample will be good predictors of post-program differences across the sample. In the National JTPA Study, for example, the earnings of adult sample members in the five calendar quarters prior to random assignment explained about 10 percent of the variance in their post-program earnings; addition of a wide variety of other demographic and socioeconomic characteristics only increased the explanatory power of the model by another 10 percentage points.[4]

It is important to bear in mind that only characteristics measured prior to random assignment may be used as covariates in the impact model. Characteristics measured after random assignment could, in principle, be affected by the experimental treatment. If so, controlling for them in the impact regression would capture part of the effect of the program in the coefficients of the covariates, thereby biasing the impact estimate c.

It is also important to note that this basic model estimates the *average* impact of the program on the *entire* treatment group. If program impacts are expected to vary within the treatment group, there may be interest in estimating the impacts on subgroups of the sample. In a later section of this paper, we discuss variants on the basic model that can be used to estimate impacts on subgroups in such cases.

Estimating the Time Path of Impacts

In some cases we would expect the impact of a program to vary systematically over time. Many programs involve a short-term investment of time and resources in order to achieve a longer-term goal—*e.g.*, training programs divert workers from the labor market while they are in training in order to increase their longer-term earnings. In such cases, we might not expect the program's effects to become evident until the participant has left the program. In other programs, we might expect the strongest effects to occur while the participant is in the program—*e.g.*, tutoring programs intended to raise students' school performance. In either type of program, we would be interested to see whether, and how long, program effects persist after the participant leaves the program.

Basic estimation approach

In the absence of other complications (which we will take up in a moment), tracing out the time path of program impacts is quite straightforward. One simply uses the basic model presented in the previous section to estimate impacts for each time period (*e.g.*, month, semester, year) after random assignment.[5] In evaluating a training program, for instance, one might use the participant's earnings in each month after random assignment, *seriatim*, as the dependent variable in the impact model.[6] This will generate a sequence of monthly impact estimates, one for each month of the follow-up period.

An important point to note here is that the data are aligned in terms of number of months after random assignment, not in terms of calendar time. In the training program example, the dependent variable in the first impact regression will be the earnings of each sample member one month after random assignment; in the second regression, it will be earnings in the second month after random assignment, etc. Since sample members will have been randomly assigned at different points in calendar time, each regression may contain observations from a number of different calendar months.

Exhibit 1 (see next page) shows the results of such an analysis, for one of the AFDC Homemaker–Home Health Aide Demonstrations. The two lines in the chart represent the earnings of the treatment and control groups in each month after random assignment. The vertical distance between the two lines in any given month equals the difference in earnings between the two groups in that month, or the estimated impact of the program.[7]

[4] Unpublished computations by Abt Associates.

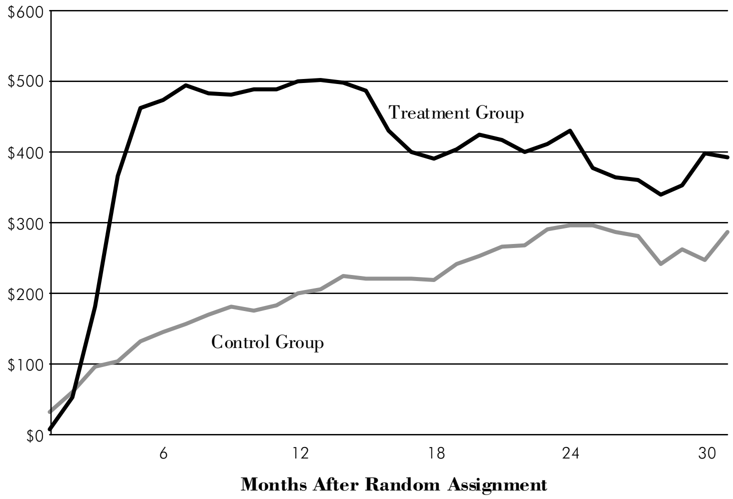
[5] An alternative approach that is sometimes used is to include each time period as a separate observation for each individual in a single regression, with separate treatment status indicators for each time period. Thus, for example, in an experiment with 1,000 sample members and a 24-month follow-up period, the impact model would contain 24 separate treatment status indicators, one for each month, and would be run on 24,000 person-month observations. While this approach provides some gains in power, the likelihood of serial correlation in the outcomes for each sample member greatly complicates the tests of statistical significance of the monthly impact estimates.

[6] Because only baseline variables can be used as covariates, and these are the same regardless of the month for which the outcome is analyzed, the right-hand side of the impact equation will be identical for all months of the follow-up period.

[7] The earnings shown in Exhibit 1 are regression-adjusted; *i.e.*, they are the predicted values from a model like equation 1. Therefore, the impact estimates shown in the exhibit control for the baseline characteristics of the sample. See Enns *et al.* (1987).

Total Monthly Earnings by Month After Random Assignment, Ohio Homemaker–Home Health Aide Demonstration

EXHIBIT 1



In the homemaker demonstrations, AFDC recipients received 4-6 weeks of training as homemaker–home health aides and were then guaranteed employment as aides for up to a year. As can be seen in the exhibit, there was little or no impact on earnings in the first several months after random assignment, as the trainees participated in the training and awaited placement in their subsidized jobs. Once in subsidized employment, trainees' earnings quickly increased, showing large gains over the control group throughout much of the follow-up period. Toward the end of the follow-up period, however, the impact of the program declined somewhat, for two reasons. First, as trainees left their subsidized jobs, some of them had trouble finding unsubsidized employment and their earnings declined somewhat. Second, throughout the post-random assignment period, the earnings of the control group steadily rose, as many of them found employment even without the help of the demonstration program.

Exhibit 6.1 reveals that the effect of the experimental program was still relatively large at the end of the follow-up period. This means that some of the impact on earnings almost certainly extended beyond the follow-up period, so that the evaluation did not capture the full effect of the program. This is important to know in comparing program benefits with program costs. In a program like the homemaker demonstrations, program costs are incurred at the outset, and therefore are fully captured by the evaluation. If some program benefits accrue after the end of the evaluation follow-up period, the evaluation will understate the net benefit of the program if it does not attempt to estimate these benefits. In a later section of this paper, we discuss ways to project program benefits beyond the evaluation follow-up period. For now, we simply note that it is important to determine whether there are likely to be benefits beyond the end of the evaluation follow-up period; estimation of the time path of impacts helps to do that.

Estimation approach when there are start-up effects

For many new programs, there is an initial period during which staff are learning their roles and responsibilities, and operational problems are being worked out. During this period, the program may not be as effective as it will be once these start-up problems have been resolved. If so, impact estimates that include this initial period will understate the steady-state impact of the program. Moreover, since the length of time the program has been in operation will be positively correlated with time since random assignment, this bias will be strongest for the estimated impacts in the early months after random assignment. This will create the appearance of a more pronounced change in impacts over time than would be the case in a mature program. In evaluations of experimental programs that have been set up specially for purposes of evaluation, then, it will be important to control for start-up effects.

This can be done relatively easily, by including in the impact regression an interaction between the treatment status indicator and the length of time the program has been in operation. One such regression specification is:

$$Y_i = a + bX_{ji} + cT_i + d\left(\frac{T_i}{t}\right) + e_i \tag{4}$$

where t is the length of time the program has been in operation at the point in the follow-up period for which impacts are being estimated. Note that t will vary across individuals, since different individuals came into the program at different times relative to the start-up of the program.

In this model, the quantity d/t measures the effect of start-up on the impact of the program; therefore, $(c+d/t)$ measures the *actual* effect of the program, including start-up effects, when it has been in operation for t periods. The variable t enters the regression in reciprocal form to reflect the fact that start-up effects, by definition, disappear over time. In this specification, as t increases, d/t asymptotically approaches zero.[8] Thus, c is an estimate of the impact of the program in steady state.

In addition to allowing estimation of the time path of impacts net of start-up effects, this formulation provides a direct test for start-up effects. A coefficient d that is significantly different from zero is evidence of start-up effects.

Impacts on Subgroups

Program impacts may vary across individuals, as well as over time, for a variety of reasons. The program may be better suited to some participants—for example, a training program that presumes a minimum level of knowledge of mathematics may not work well for poorly educated trainees. Alternatively, some trainees may not apply themselves as diligently as others, and may therefore not gain as much from the program. Or the variation in impacts may be the result of environmental factors that are completely independent of the characteristics of the program or its participants. A training program may have large impacts on earnings in a locality with a booming economy, where there is ample demand for its graduates, but little effect in a locality where unemployment is high and demand for labor slack.

Knowledge of such variations in impact across individuals or localities can help policymakers in two ways. First, *it can help them target the program on those individuals or areas where its impact will be greatest*. And, second, *it can pinpoint weaknesses in the program that need to be addressed by identifying those subgroups for whom the program is ineffective*.

Basic estimation approach

We can estimate impacts on any subgroup that can be defined on the basis of baseline characteristics—*i.e.*, on the basis of data collected prior to random assignment—simply by applying the basic model in equation 1 above to that subgroup of treatment and control group members. If, for example, we wish to know whether the program was more effective for men than for women, we can simply divide the sample into the two gender subgroups and run separate regressions on each.

As noted above, characteristics measured *after* random assignment may be affected by the experimental treatment; if they are, they will define noncomparable subgroups of the treatment and control groups. Suppose, for instance, that we wish to know the impact of a training program on participants who moved out of the area after they left the program. We could divide the treatment and control groups into movers and stayers, and estimate the regression-adjusted treatment–control differences in outcomes within each of these two subgroups. But suppose that participation in the program encouraged the more highly motivated treatment group members to move out of the area to seek a better job. Then among movers the treatment group will be, on average, more highly motivated than the control group and their outcomes will be better because of this selection factor, regardless of the effects of the program. Thus, the treatment–control difference in outcomes among movers will provide an upward-biased measure of the impact of the program. Among stayers, the opposite will be true: The treatment group will be less motivated than controls, on average, and the difference in outcomes between the two groups will understate the effects of the program.[9]

Fortunately, for the first purpose of subgroup analysis noted above—targeting of program services—*only* subgroups defined on the basis of baseline characteristics are relevant. This follows from the fact that program managers

[8] This specification assumes that the decay of start-up effects follows a very specific functional form. More complex specifications would allow the data to determine the time path of start-up effects more flexibly. We use this form here primarily for simplicity of exposition.

[9] Nonexperimental techniques are available to attempt to reduce or eliminate these biases. Discussion of these techniques is beyond the scope of this paper. Here, we only note that one can never be sure whether such techniques are successful in any particular application, because one can never be sure that the assumptions on which they are based hold in that application.

Impacts on Earnings of Adult Women, by AFDC Experience

EXHIBIT 2

Subgroup	*Estimated Impact*
Never on AFDC	$ 883
On AFDC less than 2 years	1,582
On AFDC more than 2 years	3,519*

*Estimated impact statistically different from zero at the 10 percent level.

Source: Orr, *et al.* (1996)

can obviously only target the program on subgroups that can be identified prior to program entry. Even for the second purpose noted above—identifying weaknesses in the program—subgroups defined on the basis of baseline characteristics provide a wide array of participant types for analysis.

Separate estimation of impacts for different subgroups provides unbiased estimates of those impacts and a test of statistical significance for each. That is, for each subgroup taken by itself, we can test whether the estimated impact *differs from zero* by more than could be expected by chance alone. Such tests do not, however, tell us directly whether the *difference in impacts* among subgroups is greater than could be expected on the basis of random sampling variability.

Consider, for example, the estimated impacts of JTPA on the earnings of subgroups of women on welfare, shown in Exhibit 2. In the exhibit, estimates that are significantly different from zero at the 10 percent level are designated by an asterisk. The fact that the estimated impact on the earnings of women who had received welfare for two years or more is significantly different from zero means that we can be confident that the program had a positive effect for this subgroup. We cannot have equal confidence that the program was effective for those who had been on welfare less than two years; the fact that this estimate is not statistically significantly different from zero at our chosen level of significance means that we cannot be confident that the program had a positive effect on their earnings, the positive estimate notwithstanding.

This does *not* mean, however, that we can be confident that the program had a *greater impact* on the earnings of longer-term welfare recipients than on those of shorter-term recipients, even though the latter estimate is much smaller and is not significantly different from zero. To compare two impact estimates, we must test whether the difference between them is significantly different from zero. That test depends on the sampling error of the difference between the two estimates, which in turn is a function of the sampling errors of the two estimates. We can compute the standard error of estimate of the difference between impact estimates for two independent subgroups as:[10]

5
$$\mathrm{SE}(\mathrm{I}_a - \mathrm{I}_b) = \sqrt{\mathrm{SE}_a^2 + \mathrm{SE}_b^2}$$

where SE_a and SE_b are the standard errors of the two impact estimates and $SE(I_a\text{-}I_b)$ is the standard error of the difference between them. A *t*-statistic to determine the statistical significance of the difference can be computed as:

6
$$\mathrm{t} = \frac{(\mathrm{I}_a - \mathrm{I}_b)}{\mathrm{SE}(\mathrm{I}_a - \mathrm{I}_b)}$$

i.e., the difference between the two impact estimates divided by the standard error of that difference.

When this test is applied to the estimated impacts in Exhibit 2, we find that the earnings gains of longer-term recipients were *not* significantly larger than those of shorter-term recipients. This type of finding is fairly common in subgroup analysis. One frequently finds that, while one can be confident of a positive impact for one subgroup and cannot be sure the program had any effect for another, one *cannot* say that the program had a larger effect on the former group than on the latter.

This reflects the fact that a more powerful design is required to distinguish between two estimates, both of which

[10] By "independent," we mean non-overlapping subgroups. If the two subgroups have some members in common, calculation of the standard error of the difference is more complex.

are measured with some sampling error, than to distinguish between a single estimate and a fixed point (zero). If comparisons between subgroups are likely to be important in the analysis, this should be taken into account in the design of the experiment, and a sufficiently large sample randomly assigned to provide adequate power to detect subgroup differences that are of practical importance.

Estimating subgroup impacts jointly

An alternative to estimating separate impact regressions for each subgroup is to estimate the impacts for the various subgroups jointly in a single regression equation of the form:

$$Y_i = a + bX_{ji} + \sum g_k Z_{ik} T_i + e_i \tag{7}$$

where Z_{ik} equals 1 if the ith sample member belongs to the kth subgroup and is 0 otherwise. This is just the basic impact model introduced earlier in this paper, with the single treatment status indicator T_i replaced by a set of subgroup dummies interacted with the treatment status indicator. In this model, the impact of the program on Y for the jth subgroup is given by g_k.

This approach has two advantages over estimation of separate equations for each subgroup. First, it usually provides more power because it uses the full sample to estimate the a and b coefficients.[11] Second, it allows one to test whether there are statistically significant differences in impact among the subgroups taken as a set (rather than between pairs of subgroups).

To test for significant differences among subgroups, one performs an **F-test** on the g_k coefficients. The F-statistic can be used to test the null hypothesis that the g_k coefficients are all equal. Detailed explanation of the F-test is beyond the scope of this paper; in essence, however, it tests whether adding the Z_{ik} variables to the impact model significantly improves its explanatory power. If the F-test rejects the null hypothesis, we can be confident that program impacts vary across the subgroups, although we cannot be sure which subgroups' impacts differ without conducting pairwise tests of significance.[12] An efficient analysis strategy, then, is to conduct F-tests on each of the sets of subgroups of interest (*e.g.*, subgroups defined by gender, those defined by ethnicity, etc.) and then, if desired, conduct pairwise significance tests within the sets where the F-test rejects equality of impacts.

Interpreting the results of multiple tests

There may be policy interest in a large number of participant subgroups and, in fact, many studies estimate impacts for large numbers of subgroups. While this is perfectly legitimate, in doing so one must be cognizant of the implications of performing large numbers of significance tests for the interpretation of the individual estimates.

Suppose, for example, that one estimates impacts for 20 subgroups and performs a *t*-test on each at the 10 percent significance level. At that significance level, *each* test has a 10 percent chance of rejecting the null hypothesis of no effect when it is in fact true—*i.e.*, of producing a false positive result. Therefore, even if there were no true program impacts in any of the subgroups, in 20 tests one would expect, on average, 2 positive test results due to sampling error alone. This has several implications for the interpretation of the test results.

First, if the proportion of impact estimates that are significantly different from zero is close to the significance level (in this case, 10 percent, or 2 of 20 estimates), there is a strong possibility that they represent false positive tests. One cannot be sure, of course, that this is the case, because there is no assurance that the *actual* number of false positive tests will equal the *expected* number in any specific sample.

Second, even if the number of estimates that are significantly different from zero is larger than the number that would be expected by chance alone, one must bear in mind that some of these are likely to be false positive results. Suppose, for example, that the null hypothesis is rejected in 6 of 20 tests at the 10 percent significance level. On average, we would expect approximately 2 of those 6, or one third of the statistically significant estimates, to be false positives.[13] Again, one cannot know the actual number of false positives in any given set of tests; we do know that, on average, some will be present.

[11] The combined model assumes that the functional form of the relationship between the X variables and Y is the same in all subgroups. If this assumption is incorrect, the combined model may not provide more power than separate regressions.

[12] These can also be done with F-tests.

[13] The expected number of false positives in this case will be somewhat less than 2, for two reasons. First, the expected number of false positives is 10 percent of the subgroups *for which the true impact is zero*. The 6 subgroups with estimated impacts that are significantly different from zero presumably include some for which the true impact is nonzero. Second, the 14 subgroups for which the estimated impact is *not* significantly different from zero presumably include some false negatives—*i.e.*, subgroups for which the true impact is nonzero—further reducing the base upon which the number of expected false positives should be calculated. The number of false negatives will depend on the power of the design; a weak design is less likely to detect nonzero impacts.

Third, and perhaps most importantly, we cannot know *which* of the test results are likely to be erroneous. We only know that, among those subgroups with impact estimates that are significantly different from zero, the probability that the null hypothesis has been rejected incorrectly is higher than would be indicated by the significance level of the individual test.

The risk of false positive test results when large numbers of tests are conducted suggests that experimenters should exercise restraint in the number of subgroups for which they estimate impacts and caution in the interpretation of the test results. In particular, we would caution against "fishing" for subgroup impacts. Frequently, when the estimated impacts for the sample as a whole are not significantly different from zero, researchers begin estimating impacts for one subgroup after another in search of positive impacts. If enough subgroup impacts are estimated, some estimates that are significantly different from zero will almost certainly be found; as the foregoing discussion makes clear, however, these may simply reflect false positive test results.

One way to reduce the danger of false positive test results for a given number of subgroup estimates is to reduce the significance level of the test—*i.e.*, to impose a more stringent test. Suppose, for instance, that the significance tests in the example above were conducted at the 5 percent level rather than the 10 percent level. At this significance level, we would expect only one false positive if there were no true nonzero impacts in any of the 20 subgroups.

Unfortunately, there is no simple rule for deciding on the reduction in significance level required to offset the performance of multiple tests. That is a complex issue that depends on, among other things, the number of tests, the number of true nonzero impacts among the subgroups analyzed, and the power of the design.

Multiple tests: an illustrative example

To illustrate the principles just discussed, Exhibits 3 and 4 present the results of a set of subgroup analyses performed in the National JTPA Study. Exhibit 3 shows the estimated impacts of JTPA on the earnings of a number of subgroups of adult men, along with two types of significance test: *t*-tests of the difference of the estimated impact from zero in each subgroup, and F-tests of the difference in impacts within each set of subgroups defined by a single baseline characteristic (*e.g.*, three different ethnic groups). Estimates that are significantly different from zero at the 10 percent level on the basis of the *t*-test are indicated by asterisks beside the impact estimate; significant differences *among* subgroups are indicated by an asterisk in the row labeled "F-Test, Difference Among Subgroups." ("n.s." in this row indicates that the difference in impacts among subgroups was not significantly different from zero at the 10 percent level.)

The first row of the exhibit shows that the estimated impact of JTPA on the overall sample was a $978 increase in earnings. This estimate was significantly different from zero at the 10 percent significance level (as indicated by the asterisk). The next panel shows estimaed impacts for three ethnic subgroups. Even though the estimated impacts for two of these subgroups are larger than the estimated impact on the overall sample—one of them substantially so—none of the subgroup estimates are significantly different from zero. This reflects the smaller sample sizes, and therefore reduced power, available at the subgroup level. And while the estimates vary substantially across subgroups, the F-test shows that they do not differ significantly at the 10 percent level.

Our interpretation of these results is as follows. We can be reasonably confident that JTPA had a positive effect on the earnings of adult men; our best estimate of this effect is an earnings gain of $978 (see the first row of Exhibit 3). But when we consider any individual ethnic group—say, white men—we cannot be sure that the program had a positive effect. Our *best estimate* of the program's impact on the earnings of white men is a $931 gain; but the chance that there was no true effect in this subgroup is greater than 10 percent. And while our best estimate of the program's impact on the earnings of Hispanic men is significantly different from zero and is more than twice as large as the estimated impact on white men's earnings, we cannot confidently rule out the possibility that the program had the same impact on all three ethnic groups.

When we break the sample down by welfare status (next panel of the exhibit) a similar pattern emerges. Setting aside the issue of multiple tests (to which we return momentarily), the test results here indicate that the program increased the earnings of men who were not receiving welfare at baseline; our best estimate of this gain is $1,529. We cannot be sure that men who were receiving welfare at baseline were helped by the program, although our best estimate for this subgroup is an earnings gain of $305. And, despite the substantial difference between these two estimates, we cannot be confident that the former group gained more than the latter (as indicated by the F-test); the difference in estimates could simply reflect sampling error.

The other panels of the exhibit display similar patterns. The only breakdown for which a significant difference in impacts among subgroups emerges is that based on house-

Estimated JTPA Impacts on Earnings: Subgroups of Adult Men

EXHIBIT 3

Subgroup defined by:	*Estimated Impact*
Full Sample	978*
Ethnicity	
White, non-Hispanic	707
Black, non-Hispanic	931
Hispanic	1,784*
F-test, difference among subgroups	n.s.
Welfare status	
Receiving cash welfare	305
No cash welfare	1,529*
F-test, difference among subgroups	n.s.
Education	
High school diploma or GED certificate	931
No high school diploma or GED certificate	1,353*
F-test, difference among subgroups	n.s.
Recent work experience	
Worked less than 13 weeks in past 12 months	735
Worked 13 weeks or more in past 12 months	1,140*
F-test, difference among subgroups	n.s.
Work history	
Never employed	-2,104
Earned < $4/hour in last job	245
Earned $4/hour or more in last job	1,647*
F-test, difference among subgroups	n.s.
Household composition	
No spouse present	248
Spouse present	2,759*
F-test, difference among subgroups	*
Family income in past 12 months	
$6,000 or less	733
More than $6,000	1,556*
F-test, difference among subgroups	n.s.
Age at random assignment	
22 - 29	1,221
30 - 54	1,151
F-test, difference among subgroups	n.s.

* Estimate significantly different from zero at the 10 percent significance level
n.s. Estimates not significantly different at the 10 percent level
Source: Orr et al. (1996).

Estimated JTPA Impacts on Earnings: Subgroups of Adult Women

EXHIBIT 4

Subgroup defined by:	*Estimated Impact*
Full Sample	1,837*
Ethnicity	
White, non-Hispanic	1,973***
Black, non-Hispanic	1,927
Hispanic	467
F-test, difference among subgroups	n.s.
Welfare status	
Receiving cash welfare[a]	2,359***
No cash welfare	1,634**
F-test, difference among subgroups	n.s.
Education	
No high school diploma or GED certificate	1,499
High school diploma or GED certificate	1,753***
F-test, difference among subgroups	n.s.
Recent work experience	
Worked less than 13 weeks in past 12 months	2,100***
Worked 13 weeks or more in past 12 months	1,029
F-test, difference among subgroups	n.s.
Work history	
Never employed	1,270
Earned < $4/hour in last job	1,437
Earned $4/hour or more in last job	2,540
F-test, difference among subgroups	n.s.
AFDC history	
Never AFDC case head	883
AFDC case hear less than two years	1,582
AFDC case head two years or more	3,519***
F-test, difference among subgroups	n.s.
JTPA required for welfare, food stamps, or WIN program[b]	
Yes	2,190
No	1,560***
F-test, difference among subgroups	n.s.
Household composition	
No spouse or own child present no spouse present	920
Own child under age 4, no spouse, present	2,519**
Own child, none under 4, no spouse present	598
Spouse present, with or without own child	2,617**
F-test, difference among subgroups	n.s.
Family income in past 12 months	
$6,000 or less	1,199*
More than $6,000	2,448***
F-test, difference among subgroups	n.s.
Age at random assignment	
22 - 29	1,746**
30 - 54	2,020***
> 54	833
F-test, difference among subgroups	n.s.

* Estimate significantly different from zero at the 10 percent significance level
n.s. Estimates not significantly different at the 10 percent level
Source: Orr et al. (1996).

hold composition. Here, the F-test indicates that men with spouses present experienced significantly greater earnings gains than those with no spouses. This conclusion cannot be drawn with certainty, however. This is the only one of eight breakdowns tested to show a significant difference among subgroups; at the 10 percent level of significance, we would expect about one of eight F-tests to reject the null hypothesis of no effect by chance alone. Thus, there is a good chance that this is a false positive test result.

We are on somewhat firmer ground in accepting the conclusion that several of the individual subgroups of men experienced positive program impacts. Overall, the exhibit shows estimates for 18 different subgroups. If there were no true impacts at the subgroup level, we would expect about two of the estimates to be significantly different from zero at the 10 percent level by chance alone; in fact, four are. Thus, it seems likely that at least one or two of the impact estimates that are significantly different from zero represent real effects. But, among those estimates, we have no way to distinguish those that represent real effects from those that are false positive test results.

In the end, then, we must conclude that, although we can be confident that JTPA had a positive overall effect on the earnings of adult men, this experiment was not sufficiently powerful to identify the specific subgroups of adult men who benefited most from the program—or even to be sure which ones benefited at all. This example illustrates the limits of subgroup analysis. The National JTPA Study sample included over 5,000 adult men. But even this large sample was not sufficient to allow precise estimation of impacts within subgroups. This is in part the result of the high variance of the outcome being analyzed (earnings) and in part a reflection of the relatively small impacts of the program in this demographic group.

A much clearer set of subgroup results was obtained for adult women in the National JTPA Study.[14] As shown in Exhibit 4, in 15 of the 26 subgroups for which impacts on earnings were analyzed, the estimated impacts were significantly different from zero. Thus, we can be reasonably confident that the program had positive effects for these subgroups. Our ability to detect these effects is partly due to the somewhat larger sample size for adult women (about 6,000) and partly the result of somewhat larger program effects among adult women than among adult men (an estimated overall effect of $1,176 for women, as compared with $978 for men). Even for adult women, however, the experiment was not powerful enough to detect differences among subgroup impacts of the size that actually occurred, as indicated by lack of significance of the F-tests of differences among subgroups in Exhibit 4.

Explaining Variation in Impacts Across Sites

One set of subgroups in which there is often great interest is the samples of program participants in different sites. In multisite experiments, the impact of the program may vary across sites for a variety of reasons. The most obvious is that the program may be implemented differently—either deliberately or inadvertently—in the different sites. But impacts may also vary because of differences in the participant population that allow those in some sites to benefit more from the program or because the local environment (*e.g.*, the labor market or the educational system) is more conducive to positive impacts in some sites. Distinguishing among these causes of variation—indeed, even determining whether there *is* meaningful variation across sites—is not a simple task.

Determining whether there is meaningful cross-site variation

Even in a perfectly designed and implemented experiment, the impact estimate for any given set of sample members will differ from the true impact of the program because of random sampling error. Thus, even if all sites were identical in terms of the experimental program, the program participants, and the local environment, we would expect them to yield different impact estimates due to sampling error alone. The first question that must be addressed with respect to site-specific impacts, then, is whether the estimates differ by more than could be expected on the basis of chance alone.

The treatment group members within each site can be viewed as a subgroup of the overall treatment group. Thus, the subgroup estimation techniques and significance tests discussed in the previous section can be applied directly to the analysis of site-specific impacts. That is, one can estimate site-specific impacts with an equation of the form given by equation 7 and use an F-test to determine whether those estimates differ significantly. Only if the F-test rejects the null hypothesis that the true impact is the same in all sites does the experiment provide evidence of variation in impact across sites.

[14] Neither overall estimated impacts nor any of the estimates for subgroups were significantly different from zero among male or female youth. (See Orr *et al.*, 1996.)

Site-specific Impacts on Earnings, by Target Group: National JTPA Study

EXHIBIT 5

Site	*Adult Women*	*Adult Men*	*Female Youth*	*Male Youth*
1	$2,628	$5,310*	$3,372*	$9,473*
2	2,308*	4,338	2,320	5,464
3	2,095	3,908	1,404	1,918
4	1,786*	2,533	1,222	1,414
5	1,190	2,335	649	1,192
6	1,181	2,197	556	1,090
7	1,109	1,655	244	973
8	1,069	1,540	117	119
9	884	1,212	–432	–204
10	787	721	–1,064	–1,298
11	309	710	–1,298	–2,206
12	–438	630	–1,471	–2,876
13	–1,108	–484	–2,179	–3,029
14	–1,369	–1,083	–2,355	–4,147
15	–1,410	–2,412	–3,821*	–5,836
16	–2,033	–2,637	—	—
F-test, difference among sites	n.s.	n.s.	n.s.	n.s.

Notes: Sites were ranked separately for each target group, in order of size of estimated impact. Therefore, listings for different target groups in the same row do not necessarily refer to the same site. No youth were assigned in one site.

* Significantly different from zero at the 10 percent level.

n.s.: Differences among sites not statistically different from zero at the 10 percent level.

Source: Orr et al. (1996)

Exhibit 5 shows an illustrative set of site-specific estimates, taken from the National JTPA Study. The exhibit shows estimated impacts for four major demographic groups, for each of the 16 sites in the study.[15] As can be seen, even though the estimates vary widely across sites within each demographic group, in no case are they significantly different from one another (see F-test results in last row of exhibit). As is always the case when a test of significance fails to reject the null hypothesis, this does not mean that the impact of the program *doesn't* vary across sites; it simply means that the experimental design was not powerful enough to detect whatever variation there is.

This example illustrates the difficulty of confidently estimating differences in impact across sites. Site-specific estimates are much less precise than estimates based on the overall sample because of the much smaller samples available at the site level. In this case, even with overall sample sizes as large as 6,000, the samples available in individual sites are too small (relative to the variance of the outcome of interest) to provide sufficient power to confidently identify whatever differences in impact exist across sites.

[15] Note that, to facilitate comparison of the estimates across sites, the sites are ordered in descending size of estimated impact *within each demographic group*. Thus, the estimates in any given row of the exhibit may represent *different sites* for each of the four demographic groups. Impact estimates are given for youth in only 15 sites because no youth were randomly assigned in one site.

Explaining cross-site differences in impacts

Even when the experiment finds that impacts differ significantly across sites, that finding is not, by itself, very helpful to policymakers. To be useful for policy, the analysis must explain *why* program effects were different in different sites. If the differences are due to differences in the local environment or participant population—factors that are outside the policymakers' control—analysis of site-specific impacts may be helpful in predicting the impact of implementing the program in other localities, but it will not help design a more effective program.[16] Only if the differences are due to differences in the implementation of the program across sites will knowledge of cross-site differences be useful for program design. In that case, policymakers can adopt those program features that maximize program impact.

Analysis of site-level impacts may well be the weakest area of the current practice of program evaluation, both experimental and nonexperimental. All too often, the analysis of variations in impact across sites proceeds as follows. The investigators perform significance tests on estimates of site-specific impacts, to identify those that are significantly different from zero. Those sites are then *assumed* to have larger impacts than those with estimated impacts that are not significantly different from zero—without benefit of the appropriate significance test. Finally, the researchers attribute this "difference" in impacts to whatever difference in site characteristics seems most salient to them—usually some feature of the program being evaluated—without testing for other potential causes of variation in impact.

This approach is subject to two potential errors. First, as noted in the previous section, there may be no real difference in impacts across sites, even if some site-specific estimates are significantly different from zero and others are not. Second, even if there are real differences, they may be the result of any number of differences in the local site environment or participant characteristics, rather than differences in the program being evaluated. Only through careful statistical analysis can we be confident that (a) there are real differences in impacts, and (b) we know why they occurred.

A more rigorous approach to analysis of site-level impacts attempts to test formally whether impacts vary with program characteristics, *holding constant characteristics of the local environment and program participants.* We do this by estimating an impact model of the form:

$$\boxed{8}\quad \mathrm{Y_i = a + b_jX_{ji} + cT_i + \sum d_jX_{ji}T_i + \sum f_kS_{ki} + \sum h_kS_{ki}T_i + \sum p_mP_{mi} + \sum q_mP_{mi}T_i + e_i}$$

where:

S_{ki} = a set of k characteristics of the site in which sample member i lives (*e.g.*, the local unemployment rate, average wages, etc.);

P_{mi} = a set of m characteristics of the experimental program in the site in which sample member i lives (*e.g.*, type or duration of training, availability of support services, etc.).

As before, the j variables X_{ji} measure the personal characteristics of sample member i and T_i indicates whether the sample member is a treatment or control group member.

Inclusion of the interactions of X_{ji} and S_{ki} with T_i allow the estimated impacts to vary with individual and site characteristics, thereby capturing that portion of the cross-site variation in impacts explained by these factors rather than by variations in the experimental program. The interaction of P_{mi} with treatment status allows us to test whether impacts vary with program characteristics, *holding constant individual and site characteristics*. The coefficients q_m measure this variation and the significance tests on these coefficients provide a measure of the confidence we can have that program impacts truly vary with each program characteristic—again, holding constant individual and site characteristics. (X_{ji}, S_{ki}, and P_{mi}, are also included in the impact equation without interactions, to capture any variation in the outcome with these characteristics that are common to both the treatment and control group members.)

Suppose, for example, that we have found (by estimating an impact model like equation 7, with sites as the subgroups) that the impact of an experimental training program varies across sites. On the basis of our knowledge of the experimental programs in the various sites, we might hypothesize that the more successful programs had larger impacts because they provided more intensive training or because they provided better support services for the trainees. But, without further tests, we could not rule out the alternative explanations that the participants in the successful sites were better able to take advantage of the training because of their personal characteristics or be-

[16] Differences in impact due to differences in the participant population could, in principle, be useful to policymakers in deciding how to target the program. In most cases, however, this question can be addressed more directly through analysis of subgroups defined on the basis of participant characteristics. Only in programs that can be targeted on specific geographic areas would knowledge of variation in impact due to local environmental characteristics be useful in the targeting decision.

cause the local labor market was particularly favorable for graduates of this type of training program.[17]

To test these hypotheses, we would estimate a regression equation of the form shown in equation 8. The X_{ji} variables would include such personal characteristics as education and work experience that might influence the individual's ability to benefit from training. The S_{ki} variables would include such site characteristics as the unemployment rate and the rate of growth of employment, which might affect the returns to training in the local labor market. Finally, the P_{mi} variables would include programmatic features that might influence the program's effectiveness. These could include the length of the training course, the percent of the training staff with recent private sector experience in the skills being taught, or the amount spent on supportive services per trainee.

If the coefficient of one or more of the P_{mi} variables in this model is significantly different from zero, then (subject to the qualifications discussed in the next section), we can be reasonably sure that variation in that program feature affects program impacts. Policymakers can use this information to improve the effectiveness of the program.

Limitations on our ability to explain cross-site variations in impact

The model described in the previous section is subject to several important qualifications.

First, it is important to note that *the estimate of variation in impact with program characteristics provided by this model is essentially nonexperimental*. The variation in program characteristics we observe is *natural* variation, and may therefore be correlated with other, unobserved, factors that affect the outcomes of interest. If so, the variation in impacts across sites may be due to these unobserved factors, rather than to the program. That is, our estimates of the influence of program features on impacts are potentially subject to selection bias. We cannot, therefore, have the same confidence in them that we have in the overall impact estimates, which are fully experimental. For this reason, if testing alternative program features is an important objective of the research, the experimenter should consider designs of the type discussed in an earlier paper in this series, in which multiple program variants are implemented in the same site (to hold site effects constant) and participants are randomly assigned to alternative variants (to eliminate systematic differences in participant characteristics across treatments).

The second important limitation of this model is that *the number of program features that can be tested is limited by the number of sites*. If we were to include in the impact model as many program features as there are sites, the model would explain the variation in impact across sites *perfectly*—not because we have found the real explanation for cross-site differences, but because we have provided a *different* explanation for the level of impacts in each site. In effect, the number of observations available to estimate the effect of any variable that does not vary within a site is equal to the number of sites. And it is a general proposition in econometrics that one must have more observations than the number of coefficients to be estimated. Therefore, the maximum number of site-level variables—*i.e.*, both site and program characteristics—that can be included in the equation is one less than the number of sites.

This is often a severe limitation. Most experimental programs are implemented in a relatively small number of sites—it is unusual for evaluations to have more than ten sites. In contrast, the experimental sites and programs can differ in literally hundreds of ways. Thus, the analyst is faced with the problem of choosing a small number of site and program characteristics to be tested from among the large number that could potentially affect program impacts. It is important to note that the limit on the number of variables that can be tested includes *both* site characteristics and program features. Since it is generally important to control for at least a few site characteristics, the number of program features that can be tested will usually be quite small.

Nor can one increase the number of program features analyzed by estimating the impact model repeatedly, testing different sets of site-level characteristics *seriatim*. Such "fishing" invalidates the significance tests associated with the impact estimates. If enough characteristics are tested, one can be sure of finding some effects that are significantly different from zero by chance alone.[18] (See the discussion of interpreting the results of multiple significance tests earlier in this paper.)

[17] It should be noted that the participants with the largest impacts need not be those who would do the best in the absence of the experimental program—*e.g.*, the best-educated or highest-skilled workers, or those living in sites with the greatest demand for labor. It could be that the program can't improve the prospects of well-educated, high-skilled workers, or that workers living in areas with booming economies could get just as good a job without the program. Thus, simply comparing the individual and site characteristics of the successful and unsuccessful sites cannot tell us whether these factors are responsible for the observed differences in impacts across sites; only by formally estimating the variation in impact with these characteristics can we answer that question.

[18] Fishing may, however, be a good way to generate hypotheses for testing in*future* experiments, so long as it is recognized that the results obtained from the current data are no more than suggestive.

Moreover, the power of the impact model to detect the effects of site-level variables depends on the number of **degrees of freedom** associated with these variables. The degrees of freedom for site-level variables is equal to the number of sites minus the number of site-level variables in the model. The smaller the degrees of freedom (*i.e.*, the larger the number of site-level variables), the less power the model will have to detect variation in impact with program characteristics.

Given these stringent limitations on the number of site-level variables that can be tested rigorously, it is perhaps understandable that, as noted in the previous section, many analysts simply *assume* that differences in site-specific impacts are caused by whatever differences in the program seem most salient to the researcher. In an experiment with only three or four sites, it is generally relatively easy to find some program characteristic that correlates with the differences in impacts across sites without any formal analysis. The problem is that there may be *many* site and program characteristics that correlate with these differences and, with a small number of sites, there is no way to choose from among them the real cause of variations in impact.

The correct conclusion to be drawn in such cases is that the experiment is simply not capable of providing an explanation for the variation in impacts across sites. And the lesson to be drawn for experimental design is that, if policymakers are interested in the effects of program variants, either the experimental program should be implemented in a large number of sites or, preferrably, the experiment should be explicitly designed to compare alternative program designs within the same locality.

Dealing With Potential Biases

Properly implemented and analyzed, experiments provide unbiased estimates of the impact of the experimental program. But like all other forms of research, in practice experiments are subject to a variety of imperfections that require attention at the analysis stage. In this section, we discuss several of the most common problems that can create bias in the experimental estimates: control group members who receive the experimental treatment ("cross-overs"); differences in random assignment ratios across sites or over time; missing follow-up data; and inferring the effects of a permanent program from a limited duration experiment. The discussion of these problems here is necessarily brief, intended more as an introduction to the issues involved than a comprehensive treatment of the problem and its solution.

Cross-overs

As noted in a previous paper in this series, the experimental estimates will be biased if controls receive an amount or type of service that they would not have received in the absence of the experiment. Control group contamination that takes the form of nonexperimental services similar to the experimental treatment is virtually impossible to detect because, in general, we do not know what level of these services controls would have received in the absence of the experiment—that is, after all, the purpose of the control group. We can, however, measure—and in some cases correct for—control group contamination in the form of controls receiving the experimental treatment. Such controls are termed **cross-overs**.

Cross-overs can occur for a number of reasons, depending on the institutional context of the experiment. A simple example is the case of an individual who applies to a program and is assigned to the control group, then later reapplies. If the experiment does not adequately monitor for applications by controls, the individual may be randomly assigned again or, if random assignment has ended and the program is ongoing, he or she may simply be admitted to the program.[19] After the fact, it is usually possible to detect such cross-overs by matching the experiment's random assignment records against the program's administrative records.

Some analysts simply include cross-overs in the treatment group or exclude them from the analysis altogether. Either of these approaches is likely to destroy the comparability of the treatment and control groups, leading to biased impact estimates. Fortunately, under at least some circumstances it is possible to correct for the influence of cross-overs on the impact estimates derived from the entire experimental sample, with cross-overs included in the control group.

We can correct for the effect of cross-overs on the experimental impact estimates *if we can assume that the program had the same effect on cross-overs that it would have had if they had been assigned to the treatment group.*[20] Under that assumption, the outcomes of the cross-overs can be expected to be the same (on average) as those of a corresponding set of "cross-over-like" individuals in the treatment group. While we cannot identify this latter group, we know that they exist because under random assignment

[19] See the fifth paper in this series for a discussion of the steps that can and should be taken in implementing experiments to protect the integrity of random assignment.

[20] The following derivation is based on Bloom (1986).

every subgroup of the control group has a matching subgroup in the treatment group. And since the average outcomes of the cross-overs and the cross-over-like subgroup of the treatment group are, under this assumption, the same, the estimated program impact on cross-over-like individuals is zero.

We can express the estimated impact on the overall treatment group, I, as a weighted average of the impact on cross-over-like individuals, I_c, and the impact on all other treatment group members, I_o:

$$I = cI_c + (1-c)I_o\,, \tag{9}$$

where c is the proportion of the control group that crossed over. Since, as noted above, the estimated impact on cross-over-like individuals, I_c, is zero, equation 9 reduces to:

$$I = (1-c)I_o \tag{10}$$

which can be solved for I_o:

$$I_o = \frac{I}{(1-c)} \tag{11}$$

Thus, a simple adjustment—similar to the no-show adjustment discussed in an earlier paper—is available to remove the effect of cross-overs from the impact estimates. Under the maintained assumption, I_o is an unbiased estimate of the impact of the experimental program on non-cross-over-like individuals. And like the no-show adjustment, the cross-over adjustment requires no assumption about the nature of the cross-overs. In particular, one need not assume that they are similar to the rest of the control group.

As noted above, however, one does have to assume that the program had the same effect on cross-overs as it would have had if they had been assigned to the treatment group (and therefore the same effect as it had on cross-over-like members of the treatment group). This is a relatively strong assumption, which will not always be satisfied. Consider, for example, the case of a compensatory education program that lasts for a year. At mid-year, through an administrative oversight, some control students are transferred into the classroom receiving experimental services. These students are clearly cross-overs, but they do not receive the full experimental treatment; therefore, the experimental program probably does not have the same effect on their outcomes it would have had if they had been assigned to the treatment group and had been in the experimental classroom from the beginning of the year. In such cases of partial treatment, the best that can be done is a sensitivity analysis of the effect of cross-overs on the overall impact estimate. Such an analysis is conducted by assuming alternative values of the impact on cross-over-like individuals, relative to the rest of the treatment group. Suppose, for example, that we assume that the effect on cross-overs was two-thirds the effect on non-cross-over-like individuals. Substituting $.67I_o$ for I_c in equation 9 yields:

$$\begin{aligned} I &= c(.67 \cdot I_o) + (1 - c)I_o \\ &= (1 - .33 \cdot c)I_o \end{aligned} \tag{12}$$

and:

$$I_o = \frac{I}{(1 - .33 \cdot c)} \tag{13}$$

Substitution of other values, ranging from no effect to effects equal to the impact on the rest of the treatment group, will trace out the range of possible effects on non-cross-over-like individuals.

It is important to note that the cross-over-adjusted estimate of impact applies *only* to non-cross-over-like sample members, not to the entire population randomly assigned. This is an unfortunate but unavoidable limitation of the results—because we did not observe cross-over-like individuals in a true control condition, there is no way to estimate program impacts on their outcomes. This limitation highlights the importance of taking the steps described in Chapter 5 to minimize the likelihood that cross-overs will occur.

Variations in the random assignment ratio

For operational reasons, it sometimes becomes necessary to change the random assignment ratio part way through the experiment. For example, in the National JTPA Study some sites encountered great difficulty recruiting a sufficient number of youths both to fill the available program slots and to provide for a control group. In order to secure the continued participation of those sites in the evaluation, the researchers temporarily changed the random assignment ratio from 1 control for every 2 treatment group members to 1 control for every 3 or 6 treatment group members (depending on the site). This reduced the number of youth these sites were required to recruit while still allowing them to fill all program slots. However, it also injected a potential bias into the sample.

To see this, consider a simple experiment in which 100 sample members are to be assigned to the treatment group in each of two time periods. Suppose that in the first period one control is assigned for every treatment group member, while in the second period one control is assigned for every two treatment group members. The resulting sample distribution will be:

	Time Period 1	***Time Period 2***
Treatment group	100	100
Control group	100	50

In the resulting sample, half of the treatment group is assigned in each period, whereas two-thirds of the control group is assigned in the first period and one-third in the second. Thus, the treatment and control groups are not well-matched in terms of time of assignment. If outcomes differ systematically over time (*e.g.*, if there are time trends in the outcomes), this confounding of treatment and time of assignment will bias the experimental estimates. A similar bias would occur if different random assignment ratios were used in different sites and outcomes differed systematically across sites.

There are several ways to deal with this potential bias. The simplest is to randomly remove 50 treatment group members from the sample in the second period, in order to equalize the treatment–control ratio in the two periods.[21] (It is important that the sample members to be excluded be selected randomly, in order to preserve the match between the treatment and control groups within the second period.) While this approach restores the match between the overall treatment and control groups, it is inefficient in that it throws away information on the individuals excluded from the sample.[22]

An analytic approach that uses all the available data to estimate unbiased impact estimates when random assignment ratios vary is to estimate the impact of the program within each random assignment "stratum" (*i.e.*, within each subsample assigned under the same random assignment ratio) and then compute the impact on the overall treatment group as the weighted average of the stratum-specific impacts. Since the treatment and control groups within each stratum are well-matched, this yields an unbiased impact estimate.

A number of different weighting schemes can be used to obtain the overall impact estimate. For example, weighting the estimated impact in each stratum by the proportion of the total sample in that stratum will yield an unbiased estimate of program impact for the population represented by the total sample. Alternatively, under the assumption that the impacts are the same for all strata, the minimum variance estimate of the overall impact is produced by using weights that are inversely proportional to the variances of the stratum-specific estimates.

Stratum-specific impact estimates can, of course, be obtained by estimating separate regressions for each stratum. Alternatively, under the assumption that the effects of the covariates are the same in all strata, all of the stratum-specific impact estimates can be obtained from a single regression of the form:

$$\boxed{14}\quad \mathrm{Y_i} = \mathrm{a} + \sum_{\mathrm{j=1}}^{\mathrm{j=J}} \mathrm{bX_{ji}} + \sum_{\mathrm{k=2}}^{\mathrm{k=K-1}} \mathrm{c_k S_{ki}} + \sum_{\mathrm{k=1}}^{\mathrm{k=K}} \mathrm{d_k S_{ki} T_i} + \mathrm{e_i}$$

where Y_i, X_{ji}, and T_i are, respectively, the outcome of interest, the personal characteristics, and the treatment status of the ith individual, and S_{ij} is a dummy variable that equals 1 if the *i*th individual is in the *j*th random assignment ratio stratum, and zero otherwise.[23]

The estimated impact in the *j*th stratum is d_j. The estimated overall impact is the weighted average of these coefficients, as discussed above. The null hypothesis that the overall impact is zero is tested by computing the F-test for the weighted average of the estimated d_j.

Survey nonresponse

Perhaps the most common departure from the ideal in real-world social experiments is the loss of follow-up data due to survey nonresponse. A typical response rate in a follow-up survey is 70–80 percent. If the experiment relies entirely on survey data to measure outcomes (as most do), this means that outcomes cannot be measured for 20–30 percent of the sample.

[21] Alternatively, one could randomly remove 50 controls from the period 1 sample.

[22] When the number of observations that would be removed under this approach is small relative to the overall sample, it may be worth the loss of information to avoid the added complexity involved in the approach described below. This was, in fact, the approach taken in the National JTPA Study, where 473 treatment group members were randomly excluded from an initial sample of 20,601 individuals (Orr *et al.*, 1996).

[23] Note that one of the stratum dummies must be omitted from the covariates if there is a constant term, but that all of the stratum dummies are included in the set of interactions with treatment.

If nonrespondents were a random subset of the overall sample, this loss of data would not be great cause for concern. It would reduce the precision of the estimates, because of the reduction in sample size, but the sample for whom data are available would still be representative of the overall sample randomly assigned and the treatment and control groups would still be well-matched. Thus, the experimental impact estimates would still be unbiased estimates of the program's effects on the population randomly assigned.

Unfortunately, there are usually good reasons to suspect that survey nonrespondents are systematically different from respondents. For example, surveys are less likely to be able to track people who have moved during the follow-up period than those who remain at the same address. In telephone surveys, nonrespondents are more likely to lack telephones or to have unlisted numbers. And people who are not employed are easier to locate and interview than those who work several jobs and are seldom home.[24]

Even if the subsample for whom follow-up data are available is not a representative subset of the population randomly assigned, the treatment–control difference in outcomes estimated on the basis of these data may still be an unbiased estimate of the impact of the program *on this subset* (although not of the impact on the overall population randomly assigned). This will be the case if the nonrespondents in the treatment group do not differ systematically from the nonrespondents in the control group, leaving respondent subgroups that are still well-matched.

In the worst case, nonrespondents in the treatment group differ systematically from those in the control group, resulting in a mismatch between the two groups of respondents. This will be the case when the experimental treatment influences the probability that the sample member will respond to the follow-up survey. If, for example, the experimental program encourages treatment group members to move or increases their employment rate, nonresponse is likely to be higher, and the kinds of individual who respond are likely to be different, in the treatment group than in the control group. In these cases, the experimental impact estimates may well be biased because the subgroups of the treatment and control group for whom outcome data are available are not well-matched.

Some relatively simple diagnostics are available to determine whether survey nonresponse is creating any of the problems discussed above. Some initial indications can be obtained from the survey response rate itself. A low overall response rate (say, less than 70 percent) should be viewed as a danger signal. Any substantial difference in response rates between the treatment and control groups is also a warning that the respondents in the two groups may be systematically different.

A direct test for differences between the respondent and nonrespondent groups can be obtained by comparing the baseline characteristics of the two groups. Tests of statistical significance (t-tests, F-tests, or chi-square tests, depending on the nature of the characteristic) can be applied to the difference in each baseline characteristic to determine whether the two groups differ by more than one would expect on the basis of sampling error alone. If the number of differences that are significantly different from zero exceeds the number that would be expected by chance, the subgroup for whom follow-up data are available should be regarded as materially different from the population that was randomly assigned. This means that the experimental estimates, even if unbiased, may apply to a population that is somewhat different from the one of interest for policy.

Similarly, one can test for differences in baseline characteristics between the respondents in the treatment group and those in the control group. These tests will provide an indication of the degree to which the treatment and control groups are mismatched and, as a result, will provide biased impact estimates, even for the respondent population.

In some cases, it is possible to test directly for differences in impacts between the respondent and nonrespondent groups. This is the case when follow-up data on some outcomes are available from administrative records that cover the entire experimental sample. These data can be used to derive separate impact estimates for those who responded to the follow-up survey and those who did not. If these two estimates are similar, one can be somewhat less concerned about response bias; to the extent that they are significantly different, one's concern is heightened. Of course, the fact that the two groups have similar impacts on the outcomes measured with administrative data does not guarantee that they will have similar impacts on those outcomes that are measured only with follow-up survey data—for which we can never know the impacts on survey nonrespondents. But the comparison is at least suggestive—especially if the two outcomes are closely related.

In the AFDC Homemaker–Home Health Aide Demonstrations, for example, earnings over the follow-up period were measured in a telephone survey with a response rate of 66 percent. Data on welfare benefits from state administrative records were available for the entire sample, however, including the survey nonrespondents. These data were used

[24] In most household surveys, the overwhelming majority of the nonresponse is attributable to failure to contact the sample member, rather than to refusal to be interviewed.

to derive separate estimates of program impact on welfare benefits for survey respondents and nonrespondents. These estimates showed that in six of the seven demonstration states survey respondents experienced larger reductions in welfare benefits as a result of the experimental program than nonrespondents. This suggests that the program-induced earnings gains of respondents were probably also larger than those of nonrespondents, since welfare benefits vary inversely with earnings. Thus, while this analysis did not allow direct measurement of the response bias in estimated earnings effects, it did suggest the likely direction of that bias.[25]

None of the tests described above provide conclusive evidence of bias or lack thereof. Differences in baseline characteristics need not necessarily lead to biased impact estimates, unless the outcomes are sensitive to the characteristics on which the groups differ. In any case, it is possible to control for differences in *measured* characteristics by including these characteristics as covariates in the impact regression.[26] Conversely, the fact that respondents in the treatment and control groups do not differ in their measured characteristics does not guarantee that the impact estimates will be unbiased; the two groups may still differ in *unmeasured* characteristics. Even when administrative data are available to test directly for response bias in the impact estimates for some outcomes, there is no guarantee that the results of those tests are applicable to the outcomes for which only survey data are available.

The experimenter with incomplete follow-up data is in much the same position as the nonexperimental analyst who wants to know if a nonexperimental comparison group is well-matched to the program group. While one can conduct a number of tests that give one more or less confidence in the impact estimates, in the end one can never know how well the two groups are matched. Perhaps the most important difference between these two situations is that the experimenter at least knows that the two groups started out well-matched.

The remedies for any mismatch between the two experimental groups are also essentially the same as those available to adjust for differences between a nonexperimental comparison group and the program group. Discussion of the large, complex literature on these econometric techniques far exceeds the scope of this paper. Here we note only that there is little consensus on which, if any, of these techniques are adequate to address the problem. This discouraging conclusion argues strongly for taking every step possible to minimize nonresponse in experimental follow-up surveys.[27]

Duration bias: inferring responses to permanent programs from temporary experiments

Experiments are run for a limited time period—usually one to five years—defined in advance. In many—perhaps most—cases, the temporary nature of the experimental program does not affect the experience of the program participant. Participants in an experimental training program, for example, receive the entire experimental treatment in a few weeks or months and exit the program. Whether the program continues to enroll other participants does not affect their experience or the impacts of the training on their subsequent outcomes.

In certain situations, however, the response to a temporary program could be quite different from the response that could be expected if the same intervention were adopted on a permanent basis. Consider, for example, an intervention like the insurance plans provided by the Health Insurance Experiment. These plans decreased the net price of health care to covered families for the period of time they were enrolled in the experiment.[28] This meant that health care was “on sale” for that period of time, creating an incentive to accelerate the purchase of services that they would normally have made later. Obviously, the timing of many health care expenditures is not discretionary. But the consumer has a good deal of latitude in the scheduling of some services, such as dental and psychiatric care, eyeglasses, preventive services, and elective surgery. If the experimental subjects used this latitude to purchase services during the experimental period, while their price was low, rather than later, their observed consumption during

[25] See Enns *et al.* (1987). By assuming that the relationship between earnings gains and welfare benefit reductions (*i.e.*, the “benefit reduction rate”) was the same for respondents and nonrespondents, the analysts were able to derive an estimate of the bias. These estimates suggested that in five of the seven demonstration states, the bias was less than 20 percent. In the other two states, however, the estimated bias was substantially larger than the experimental effect.

[26] Specifically, they should be included both as main effects and as interaction terms with the treatment indicator, to allow treatment to vary with differences in these characteristics. The overall impact can then be evaluated on the basis of the estimated coefficients, setting these variables equal to their sample means.

[27] See the fifth paper in this series for a discussion of ways to reduce follow-up survey nonresponse.

[28] The change in net price to the family depended on the provisions of the specific plan to which they were assigned, as compared with those of the insurance they would otherwise have had. In some cases, the net price to the family was actually higher under the experimental plan. The point here does not depend on the direction of the change, but on the fact that it was a *temporary* change.

the experimental period would overstate that which could be expected under a permanent program with the same insurance provisions.[29] Such an effect is termed **duration bias**.

The most direct way to deal with this potential bias is to design variation in the length of the intervention into the experiment. In the Health Insurance Experiment, for example, a random subset of families was assigned plans that lasted 5 years, rather than the standard 3 years. The responses of families receiving the 3-year treatment were then compared with those of families receiving the 5-year treatment. Since these two sets of families differed only in treatment duration, any significant difference in response between the two sets would be evidence of duration bias in the estimates of the impact of the 3-year treatment.

Of course, even the longer experimental treatment was still a temporary intervention, and may itself have been subject to duration bias. The only way to be sure that duration bias has been eliminated would be to implement a treatment of such long duration that it is, for all practical purposes, permanent. To our knowledge, this has only been done once. In the Seattle-Denver Income Maintenance Experiment, a random subsample of about 200 families were enrolled in negative income tax plans that were intended to last for 20 years. The responses of families assigned to the 20-year plans were compared with those of families assigned to 3- and 5-year plans to test for duration bias in the latter two groups, which constituted the bulk of the experimental sample.

Unfortunately, maintaining experimental treatments for such a long period of time is very expensive. Moreover, a long-term subsample is only useful for research purposes during the initial period when its responses can be compared with those of the main sample. For these reasons, the 20-year sample in the Seattle–Denver Experiment was terminated after approximately 8 years, with the families given two years advance notice, during which time they received fixed monthly payments to help them readjust to the absence of income support. While early termination of the sample was clearly optimal from a research standpoint, it raised difficult ethical questions about the experiment's obligations to the families.

If variation in duration has not been incorporated into the design of the experiment, duration bias is much more difficult to detect. In some instances, an indication of the existence of duration bias can be obtained by examining the time path of impacts over the experimental period. In the case of the Health Insurance Experiment, for example, one would expect the incentive (and opportunity) to shift expenditures into the experimental period to grow as the family nears the end of its enrollment in the experimental plan. Thus, rising expenditures toward the end of the enrollment period would be an indication of duration bias. This evidence would be stronger if such rising trends in expenditures were most pronounced among the types of health care that are subject to significant consumer discretion.

In the context of the early income maintenance experiments, Metcalf (1974) proposed a method of estimating duration bias that turned on analysis of changes in the family's savings and consumption behavior over time. Arrow (1973) has also suggested a method of estimating duration bias using follow-up data collected after the end of the experiment. It is not clear, however, how reliable—or generally applicable—these approaches are.

Estimating the Benefits and Costs of an Experimental Program

Up to this point, the discussion in these papers has focused on estimating the impacts of an experimental program on its participants. By "impact", we have generally meant the effect of the program on one or more behavioral outcomes that represent the objectives of the program. In evaluating a training program, for example, we focus on its impacts on participants' earnings; in evaluating a remedial education program, we focus on the program's effects on student performance.

A finding that the program had its intended effects is not, however, sufficient to justify adoption (or continuation) of the program. To determine whether the program is worthwhile, it is necessary to compare those effects—*i.e.*, the program's **benefits**—with the resources given up to produce them, as well as any adverse impacts—*i.e.*, the program's **costs**. In most cases, it is also important to examine the **distributional consequences** of the program—*i.e.*, who bears the costs and who reaps the benefits.

Cost–benefit analysis has given rise to a voluminous literature; we will not attempt even to summarize this complex methodology here.[30] Rather, we discuss the relationship between program benefits and costs and the experimental

[29] See Newhouse (1993) for a more complete discussion of these issues in the Health Insurance Experiment. See Metcalf (1974) for a detailed theoretical and empirical analysis of the corresponding incentives in the income maintenance experiments.

[30] For an excellent text on the subject, see Boardman *et al.* (1996).

impact estimates. We then illustrate that relationship by presenting a conceptual framework for the cost–benefit analysis of a specific experimental evaluation and discussing the measurement of the benefits and costs involved in that evaluation.

Measuring benefits and costs in an experiment

A comprehensive cost–benefit analysis of the experimental program requires an exhaustive measurement of all of the impacts of the program (both beneficial and adverse), as well as the resources required to produce them. As we have seen, experiments are ideally suited to performing the first step of this process—the treatment–control difference in outcomes is an unbiased estimate of the impact of the program on *any* outcome. Thus, we need only anticipate and collect data on all of the outcomes affected by the program to obtain the impact estimates required for cost–benefit analysis.

Perhaps less obviously, *the resources required to produce those impacts are also appropriately measured by the difference in resources consumed between the treatment and control groups*. This is the increase in resource consumption that would occur if the experimental program were adopted on an ongoing basis.

The treatment–control difference in resource consumption may differ from the budgetary cost of the program for several reasons. First, *if the experimental program displaces some nonexperimental services*, then the net cost to society is the resources required to produce the experimental services less the resources freed up by the displacement of nonexperimental services. Suppose, for example, that the costs of services per participant are as follows:

	Treatment Group	*Control Group*	*Difference*
Experimental	$1,000	$ 0	$1,000
Nonexperimental	200	400	–200
Total	$1,200	$400	$800

If the experimental program pays the full cost of the experimental services, then the budgetary cost of the program is $1,000 per participant. The net cost to society of the experimental program, however, is only $800 per participant—the treatment–control difference in total service costs. This net cost is the sum of the total cost of the experimental program ($1,000 per participant) and the savings that result from displacement of nonexperimental services ($200 per participant).

The second reason that the net cost of the experimental program may differ from its budgetary cost is that *the experimental program may not pay the full cost of the services it provides*. For example, training programs often refer participants to basic education programs funded by other sources or to community colleges that are subsidized by local taxpayers. The training program may reimburse these organizations for part of the cost of the services they provide—*e.g.*, it may pay the community college tuition. The *social cost* of the services provided, however, is measured by the full cost of the services, not just the part paid by the training program.

We can think of the portion of the cost of services to the treatment group that is not borne by the experimental program as nonexperimental services. (This interpretation is perhaps most natural in a case where the experimental program refers participants to another service provider, but it also applies to cases where the experimental program is subsidized by other funding sources.) In this case, then, the experimental program *increases* the consumption of nonexperimental services by its participants, rather than displacing nonexperimental services. In the cost framework presented above, this case might look as follows:

	Treatment Group	*Control Group*	*Difference*
Experimental	$1,000	$ 0	$1,000
Nonexperimental	800	400	400
Total	$1,800	$400	$400

Because of referrals to other service providers or subsidies to the experimental program, in this case the treatment group actually consumes more nonexperimental services than the control group ($800 vs. $400 per participant). As a result, the treatment–control difference in total service costs ($1,400) exceeds the budgetary cost of the experimental program ($1,000).

Critics of social experiments frequently complain that they understate the full effects of the program, because controls receive similar services from nonexperimental sources. This argument has some validity if the evaluator looks only at the impacts of the program. In a benefit–cost analysis, however, by taking account of the effects of the experimental program on the costs of nonexperimental services, we automatically adjust the cost side of the benefit–cost analysis to conform to the treatment–control service differential that produced the impacts (and therefore the benefits) measured by the experiment.

In the case where the experimental program *displaces* nonexperimental services, the treatment–control service

Conceptual Framework for Cost-Benefit Analysis of an Employment and Training Program

EXHIBIT 6

	Costs (–) and Benefits (+) from the Perspective of:		
	Participants	***Rest of Society***	***Society***
COSTS			
Operational costs of the program	0	–	–
Forgone leisure and home production	–	0	–
BENEFITS			
Earnings gains	+	0	+
Reduced costs of nonexperimental services	0	+	+
TRANSFERS			
Reduced welfare benefits	+	–	0
Wage subsidies	–	+	0
NET BENEFITS	+/–	+/–	+/–

differential is *less than* the full amount of services provided by the program. The estimated impacts are therefore presumably smaller than would be produced by the full amount of services provided by the experimental program. On the cost side, however, we take account of this displacement; as a result, our measure of social costs is correspondingly smaller than the full budgetary cost of the experimental program. Similarly, when the experimental program causes *increased* use of nonexperimental services, the treatment–control service differential and the measured social cost of the program are *greater than* the full amount of the experimental services provided and their budgetary cost. Thus, there is a strong argument for *always* conducting a complete cost–benefit analysis when either the treatment or control group receives some nonexperimental services.

Illustrative example: Benefits and costs of a training program

Exhibit 6 shows a social accounting framework for the cost–benefit analysis of a training program for low-income workers. Each row of the exhibit is a benefit or cost of the program, expressed in dollars per participant. The columns of the exhibit indicate the group within society to whom the benefit or cost accrues—the participants, the rest of society, or society as a whole. Since particpants and the rest of society together constitute society, the values in society's column are simply the sum of the values in the first two columns. In this exhibit, positive values indicate program benefits; program costs are denoted by negative values. The **net benefit** to each group is the algebraic sum of the program's benefits and costs to that group—*i.e.*, the sum of the values in the column corresponding to that group.

The principal costs of the program are shown in the first panel of the exhibit. *Operational costs* include staff salaries, rent, utilities, and all of the other administrative and overhead costs necessary to run the experimental program. They also include anything the program pays to other training providers to serve participants in the experimental program. The operational costs of the program are borne by the rest of society (*i.e.*, nonparticipants) and are also a cost to society as a whole. This entry corresponds to the treatment–control difference in the cost of experimental services in the tables presented in the previous section. (Effects on the cost of nonexperimental services are considered separately, below.)

The major cost borne by program participants is *forgone leisure and home production* as a result of time spent in training and any additional time spent working. We can measure the time participants spend in work or training relatively accurately, through program records and follow-up surveys. The time cost to participants is the treatment–control difference in total hours spent in these activities. Placing a monetary value on this time loss is more difficult, both conceptually and empirically. Theoretically, this value is measured by the area under the participant's labor supply curve.[31] In practice, it is often

[31] See Boardman *et al.* (1996), Chapter 9, for a derivation of the theoretically correct measurement of this value.

approximated by assigning the participant's wage rate (if known) or the minimum wage to each forgone hour.

The principal benefit of an effective training program is participant *earnings gains*. Earnings can be measured through follow-up surveys or the employer-reported earnings records maintained by Unemployment Insurance agencies in each state. The earnings gain attributable to participation in the experimental program is measured by the treatment–control difference in earnings. The principal problem in measuring earnings gains is that, although these gains may continue to accrue throughout the participant's entire working life, the typical evaluation follow-up period lasts only 1-3 years. Thus, the evaluation must either find some way to project these earnings gains beyond the follow-up period or run the risk of understating program benefits—perhaps substantially so.[32]

One way to project earnings impacts beyond the evaluation follow-up period is to estimate the rate of "decay" of the impact on earnings over the follow-up period and then project that rate of decay over the remainder of the participant's working life. If, for example, earnings gains are falling by 10 percent per year over the evaluation follow-up period, one could project that the impact observed at the end of the follow-up period would continue to decay at that rate.

Unfortunately, in many evaluations the follow-up period is not long enough to establish a reliable trend in the impact estimates—especially since impacts on earnings are likely to rise over the first part of the follow-up period and only begin to decay toward the end of the follow-up period. Because of this problem, some evaluations have simply used the decay rates estimated in similar studies that have followed the experimental samples for long periods.[33]

Another approach is to perform a "sensitivity analysis" by projecting earnings gains under different rates of decay, in order to examine the sensitivity of the resulting estimates of net benefits to different assumptions about long-term earnings impacts. If, for example, net benefits are positive under all plausible values of the decay rate, for policy purposes it does not really matter what decay rate is used.

Whatever method of projecting future earnings gains is used, it is essential that they be "discounted" to reflect that fact that a dollar of benefit in the future is not worth a dollar of cost in the present. More generally, because the costs of most social programs are incurred at the outset, whereas their benefits tend to be distributed over long periods of time, costs and benefits must be calculated in "present discounted value" terms in order to be comparable. Choice of an appropriate discount rate is a complex issue that is beyond the scope of this paper.[34]

In addition to the benefits to participants, in cases where the experimental training displaces similar nonexperimental services, the rest of society enjoys the benefit of *reduced costs of nonexperimental services*. (If the experimental program increased nonexperimental services, this entry would be a cost.) As indicated in the tables in the previous section, this benefit is measured by the treatment–control difference in the cost of nonexperimental services received by the sample.

The cost of nonexperimental services is one of the most difficult cost–benefit components to measure, because these services are often provided by a large number of agencies not directly involved in the experiment. Identifying and gaining the cooperation of these agencies can be an insuperable task. Sometimes acceptable cost data can be obtained from secondary sources—*e.g.*, other studies that focused on these organizations. In the National JTPA Study, for example, the evaluators were able to obtain data from federal statistical agencies on the cost of education and training at public schools. To collect data on private schools, however, it was necessary to conduct a telephone survey.[35]

Some program impacts produce benefits to one segment of society that are exactly offset by costs to another segment of society, so that the resulting net social benefit is zero. These are known as **transfers**. For example, if increased employment among program participants leads to reduced welfare payments, that represents a cost to participants equal to the loss of welfare benefits and an exactly offsetting benefit to the rest of society, in the form of reduced taxes. Similarly, if a training program subsidizes wages in private employment, the benefit to participants is exactly offset by an equal tax cost to nonparticipants. Transfers are measured by the treatment–control difference in the outcome in question (*e.g.*, welfare benefits or wage subsidies). Generally, this is measured for participants and the offsetting cost or benefit to the rest of society is imputed to be exactly equal (but opposite in sign) to the impact on participants.

[32] For a more complete discussion of the issues involved in projecting earnings gains, see Boardman *et al.* (1996), Chapter 9.

[33] In the area of employment and training programs, for example, several studies have followed experimental samples for five years or more (see Bell *et al.*, 1995; Friedlander and Burtless, 1995; U.S. General Accounting Office, 1996; and Couch, 1992).

[34] See Boardman (1996), Chapters 4 and 5, for a detailed discussion of discounting future benefits.

[35] See Orr *et al.* (1996), Appendices A and B.

As noted earlier, the net benefit of the program to any given social group is the algebraic sum of all costs and benefits to that group. The final row of Exhibit 6 shows this sum for each group. As with all costs and benefits, the net benefit of the program to society as a whole is just the sum of the net benefits to participants and to the rest of society.

While programs with positive net benefits to society as a whole are generally viewed as worthwhile, it is important to recognize that computation of net social benefits involves a very important assumption: It assumes that a dollar of benefit to one member of society just offsets a dollar of cost to another member of society, regardless of who the two individuals are. A typical cost–benefit finding is that net benefits to participants are positive, while net benefits to the rest of society (who bear the cost of the program) are negative. Under the fundamental assumption underlying the computation of net social benefits, such a program is socially worthwhile so long as the gains enjoyed by the participants exceed the costs borne by the rest of society. Some would question that assumption, arguing that in order to justify taking resources away from the rest of society to support the program, the gains to participants should substantially exceed the costs to nonparticipants. At a minimum, the analyst has a responsibilty to show the distributional consequences of the program (as is done in Exhibit 6) so that policymakers can form their own views of its social desirability.

Nonmonetary benefits and costs

In this illustrative example, we were able to place a dollar value on each of the major costs and benefits of the program, thereby allowing computation of net benefits to each segment of society in monetary terms. This is not always possible. Some program impacts simply cannot be expressed in monetary terms. For example, some youth programs are intended to promote good citizenship, develop leadership traits, and change youths' attitudes about work and education.

Even when important program impacts cannot be valued in monetary terms, they can often be measured experimentally in *nonmonetary* terms. For example, sociologists and social psychologists have developed scales that can be used to measure a number of different attitudes. The treatment–control difference in the mean score on such a scale at follow-up is a measure of the program's impact on that attitude.[36] Similarly, even if we cannot place a dollar value on illnesses prevented or lives saved by a social program, we can at least measure the program's effects as the treatment–control difference in those outcomes, measured in natural units (*e.g.*, sick days, lives lost).

When some program costs or benefits are measured in nonmonetary terms, policymakers must assess these costs and benefits along with net monetary benefits. In the simplest case, when both nonmonetary benefits and net monetary benefits are positive, the conclusion is straightforward: The program is worthwhile, no matter what value we place on the nonmonetary benefits. Positive nonmonetary benefits simply reinforce positive net monetary benefits. Conversely, if both nonmonetary benefits and net monetary benefits are negative, the program is not socially worthwhile.

When nonmonetary benefits are positive and net monetary benefits are negative, policymakers face a tradeoff—they must decide how much society is willing to pay to secure the nonmonetary benefits of the program. If, in their judgment, the estimated nonmonetary benefits of the program are worth more than its estimated net monetary cost, then the program is socially worthwhile; if not, it is not. Similarly, if nonmonetary benefits are negative and net monetary benefits are positive, policymakers must decide whether the former outweigh the latter.

In the AFDC Homemaker–Home Health Aide Demonstrations, for example, provision of home care services to elderly and disabled individuals entailed substantial net monetary costs. Although it was originally hoped that the operational costs of providing home care would be offset by reduced use of hospital and nursing home care, these monetary benefits largely failed to materialize. The program did, however, have statistically significant positive impacts on a number of measures of client well-being, such as orientation, ability to communicate, number of activities of daily living in which the client was able to function independently, and self-reported health status. The clients also overwhelmingly said that they enjoyed the companionship and assistance of the aides. In presenting the results of the evaluation, the researchers characterized the net monetary costs of the demonstration program as the price of obtaining these nonmonetary benefits.[37]

It is important to recognize that the value judgments involved in making such tradeoffs are the province of policymakers, who have been elected or appointed to intrepret society's preferences in such matters. They are not technical decisions that can be made by the analyst.

[36] Such impact estimates are, of course, only as valid as the scales on which they are based.

[37] See Orr and Visher (1987).

Costs and benefits that cannot be measured

Some costs and benefits cannot be meaningfully measured at all. For example, an important social benefit of programs that successfully place welfare recipients in private sector jobs is the satisfaction that taxpayers derive from knowing that such individuals are productively employed, rather than being dependent on public assistance.[38]

If such costs or benefits are likely to be important program effects, it is important to take note of them in the analysis and, if possible, to indicate their likely direction, even if they cannot be quantified. This serves both to document the limitations of the study and to remind policymakers to take such potential effects into account in assessing the overall costs and benefits of the program.

❧

References

Arrow, Kenneth J. 1973. "Welfare Analysis of Changes in Health Coinsurance Rates." Research Report R-1281-OEO. Santa Monica, CA: The Rand Corporation.

Bell, Stephen H., Larry L. Orr, John D. Blomquist, and Glen G. Cain. 1995. *Program Applicants as a Comparison Group in Evaluating Training Programs.* Kalamazoo, Michigan: W.E. Upjohn Institute for Employment Research.

Bloom, Howard. 1986. "Accounting for Cross-Overs", in *Data Collection and Analysis of JTPA Evaluation Experiments: Technical Proposal* (unpublished). Cambridge, MA: Abt Associates Inc.

Boardman, Anthony E., David H. Greenberg, Aidan R. Vining, and David L. Weimer. 1996. *Cost–Benefit Analysis: Concepts and Practice.* Upper Saddle River, NJ: Prentice–Hall, Inc.

Couch, Kenneth. 1992. "Long-Term Effects of the National Supported Work Experiment, and Parametric and Nonparametric Tests of Model Specification and the Estimation of Treatment Effects." Unpublished Ph.D. dissertation, Department of Economics, University of Wisconsin–Madison.

Enns, John H., Stephen H. Bell, and Kathleen L. Flanagan. 1987. *AFDC Homemaker–Home Health Aide Demonstrations: Trainee Employment and Earnings.* Bethesda, Md.: Abt Associates.

Friedlander, Daniel, and Gary Burtless. 1995. *Five Years After: The Long-Term Effects of Welfare-to-Work Programs.* New York: Russell Sage Foundation.

Metcalf, Charles E. 1974. "Predicting the Effects of Permanent Programs from a Limited Duration Experiment," *Journal of Human Resources, Vol. IX, No. 4, pp. 530-555.*

Newhouse, Joseph P. 1993. *Free for All? Lessons from the RAND Health Insurance Experiment.* Cambridge, Mass.: Harvard University Press.

Orr, Larry L., Howard S. Bloom, Stephen H. Bell, Fred Doolittle, Winston Lin, and George Cave. 1996. *Does Job Training for the Disadvantaged Work? Evidence from the National JTPA Study.* Washington, D.C.: Urban Institute Press.

Orr, Larry L., and Mary G. Visher. 1987. *AFDC Homemaker–Home Health Aide Demonstrations: Client Health and Related Outcomes.* Washington, D.C.: Abt Associates Inc.

U.S. General Accounting Office. 1996. *Job Training Partnership Act: Long-Term Earnings and Employment Outcomes.* GAO/HEHS-96-40. Washington, D.C. March.

[38] A similar benefit accrues to the (former) welfare recipients themselves. However, this benefit could arguably be measured in nonmonetary terms using an attitudinal scale. This approach is not available for measuring taxpayer satisfaction, because taxpayers would not derive the same satisfaction from a small demonstration project that they would from a large, ongoing program.

Social Experimentation: Evaluating Public Programs with Experimental Methods

PART 6

Analysis

Larry L. Orr

October, 1997

DEPARTMENT OF HEALTH & HUMAN SERVICES · USA

OFFICE OF THE ASSISTANT SECRETARY
FOR PLANNING AND EVALUATION

Preface

This paper is part of a series on the design and implementation of social experiments. These papers are intended to provide a relatively non-technical synthesis of the fundamental principles of the evaluation of public programs using experimental methods, for both those who design and conduct social experiments and those who use the results of experimental studies. A complete listing of the papers in this series is provided below.

The author is indebted to Wendell Primus and David Ellwood, former Deputy Assistant Secretary and former Assistant Secretary for Planning and Evaluation, respectively, of the U.S. Department of Health and Human Services, for the financial support needed to produce these papers. I also wish to thank Howard Rolston, Steve Bell, and the participants in a series of seminars for the staff of the Office of the Assistant Secretary for Planning and Evaluation for their insightful comments and suggestions on earlier drafts of several of these papers. Erik Beecroft and Lu Nguyen provided computational assistance in developing some of the exhibits.

PAPERS IN THE SOCIAL EXPERIMENTATION SERIES

Part 1: Background and Rationale
Part 2: Basic Concepts and Principles
Part 3: Alternative Random Assignment Models
Part 4: Sample Design
Part 5: Implementation and Data Collection
Part 6: Analysis
Part 7: Social Experimentation and the Policy Process

The author is Chief Economist, Abt Associates, Inc.   These papers were written while he was a visiting scholar in the Office of the Assistant Secretary for Planning and Evaluation, on leave from Abt.