Subgroup Analysis in Social Experiments: Measuring Program Impacts Based on Post-Treatment Choice

LAURA R. PECK

ABSTRACT

A fundamental question within the field of program evaluation is "Do social programs work?" Although experiments allow us to answer this question with certainty, they have some limitations. Experiments generate mean program impacts and even mean impacts by subgroup, but they often leave unexplored the impacts on subgroups determined by treatment use. This work proposes a methodology for analyzing the impacts of social programs on previously unexamined subgroups. Rather than using a single trait to define subgroups—which is currently the dominant method of subgroup analysis—the proposed approach estimates the impact of programs on subgroups identified by a post-treatment choice while still maintaining the integrity of the experimental research design. Analysis of data from the experimental evaluation of New York State's Child Assistance Program (CAP) provides an application of the proposed technique.

INTRODUCTION

One of the fundamental questions within the field of program evaluation is "Do social programs work?" To answer this question, many social scientists generally prefer experiments when possible, where individuals are randomly assigned to treatment and control groups so that the only difference between the two groups (other than random sampling error) is that treatment group members are offered a program and control group members are not. The outcomes of the control group members provide a counterfactual, that is, what would have happened in the absence of the treatment. This ideal counterfactual allows evaluators to net out the effects of historical and maturation trends, of selection bias, of regression artifacts and of any other plausible rival explanations of observed changes in outcomes. The difference between treatment group outcomes and control group outcomes is the *impact* of the program. As a

Laura Peck • School of Public Affairs, Arizona State University, P.O. Box 870603, Tempe, AZ 85287-0603, USA; Tel: (1) 480-727-7081; Fax: (1) 480-965-9248; E-mail: Laura.Peck@asu.edu.

American Journal of Evaluation, Vol. 24, No. 2, 2003, pp. 157–187. All rights of reproduction in any form reserved.
ISSN: 1098-2140 © 2003 by American Evaluation Association. Published by Elsevier Inc. All rights reserved.

result, experiments are considered by many as the "gold standard" of program evaluation and are the best way to know with confidence that an intervention *causes* observed changes in individual outcomes.

Problem Definition

Within treatment groups, however, there often exists a combination of individuals who participate in the entire program and those who do not (because they are "no-shows," "dropouts," "noncompliers," or "nonparticipants" of some sort). Impacts are estimated as the *average treatment effect* on the combination of individuals who took up and did not take up the entire program or portions of it. In other words, the average treatment effect measures the program impact on the combination of individuals who complied and did not comply with their treatment group status. Measuring the average treatment effect is also considered an *intention-to-treat* analysis. That is, the average treatment effect measures the effect of the intention to provide a treatment of some sort and not the effect of whether a treatment was actually received; it measures the effect of the offer of treatment.[1] Attention has been given to this issue within the field of program evaluation, and researchers have suggested ways to generate more accurate estimates of how programs impact individuals who actually receive the treatment in full. These efforts focus on measuring the effect of the treatment on the treated instead of the effect of the intention-to-treat.

Efforts to Deal with the Problem

Scholars have described this problem either directly or indirectly for several years, attributing various names to what is essentially the same concern. What Bloom (1984) calls "no-shows," for example, Heckman, Smith, and Taber (1998) call "dropouts," and others call "noncompliers," but essentially they discuss the same evaluation topic. Similarly, the "distributional impacts" under study by Friedlander and Robins (1997) and Imbens and Rubin (1997) are akin to the "heterogeneous impacts" under study by Heckman (1999) and Heckman, Smith, and Clements (1994). Likewise, Manski's "mixing problem" (1995, 1996, 1997) deals with the same central concern as work by Angrist, Imbens, and Rubin (1996), Eberwein, Ham, and LaLonde (1997), Frangakis and Rubin (2000), and Hirano, Imbens, Rubin, and Zhou (1999), that of heterogeneous treatment groups that include a mixture of individuals who received and did not receive the test treatment.

An early one of these efforts, Bloom's (1984) "no-show" correction, lays out the assumptions necessary to estimate the effect of a program specifically on those who participate in the treatment. By assuming that individuals who do not show up for the treatment are unaffected by it, Bloom's correction provides a way to estimate a program's impacts on its participants. Bloom's approach is essentially the instrumental variables (IV) technique that Angrist et al. (1996) employ in their work on identifying causal effects. In their example, Angrist, Imbens, and Rubin assume that the only way for draft status to have had an effect on one's later health status is through actual service in the military. Empirical results based on their own relaxation of the exclusion restriction show that this assumption may not be realistic. Related work by Heckman and his coauthors (1997, 1998, 1999) explores the same program evaluation question, that of the heterogeneity of impacts that accrue to those within a treatment group who do or do not engage in the treatment to which they are assigned.

Empirical applications of this theoretical work highlight the importance of solving this evaluation problem. One such example (Hirano et al., 1999) examines the effect of being assigned to a treatment group among those who comply with their treatment status. The authors discuss the health impacts on patients of doctors' being encouraged to administer a flu vaccine. They assume that noncompliers are unaffected by the treatment and thereby focus on increasing the precision of the estimated treatment impact on compliers. This approach is common among applications.

The need to assume no impact on those who do not take up the treatment ignores the possibility that this subset actually might be affected in some way by the offer or presence of the treatment. As Heckman, Smith, and Clements (1997) discuss, for example, there is value to potential participants in having the additional options that social programs confer. If the presence of a treatment motivates individuals to change their behavior in some way, then it is likely that even the outcomes of those who do not take up the treatment might change because of the treatment offer.

Such an arrangement suggests the need to measure impacts not only on program participants, as has been the focus of prior work, but also on nonparticipants. Frangakis and Rubin (2000) undertake such an analysis by comparing subjects who "*got* a common value *s* . . . of the post-treatment variable," or, in other words, by estimating the effect of a treatment on its compliers and on its noncompliers. These authors suggest that theirs is the sole work within the field of program evaluation that attempts to *measure impacts on noncompliers* (or nonparticipants), while the field in general remains concerned either with more accurate estimation of a treatment's impact on compliers (or participants) or with ways of measuring a treatment's heterogeneity of impacts.

Sidebar: a paragraph on terminology. As the prior section highlights, the terminology used to describe the issues related to this problem vary by author, making it useful to discuss briefly how this article uses terms. In line with common practice in the program evaluation literature, I use "control group" to refer to those randomly assigned out of the treatment being tested. I use "treatment group" to refer to those randomly assigned to receive the intervention being tested, although others within the field refer to these commonly as the "program group" or "experimental group" or, infrequently (and incorrectly), as "participants."[2] Within the treatment group, there are subgroups that can be thought of as "compliers" or "participants," that is, those who, once assigned to the treatment group comply with, enroll in or engage in the treatment being tested. In turn, those who do not comply with, enroll in, or engage in the treatment offered have been called "no-shows," "dropouts," "noncompliers" or "nonparticipants" or, sometimes, "defiers." I use the terms "participants" and "nonparticipants" in this work, but it should be noted that they refer to the same subsets as compliers and noncompliers. In brief, then, the terminology used throughout refers to the control group as the counterfactual, the treatment group as those randomly assigned to the treatment, and participants and nonparticipants, respectively, as those who, within the treatment group, comply or do not comply with the treatment.[3]

Research Questions

Because this is a methodological project, the research questions, rather than asking whether a program has particular impacts, ask whether and under what conditions it is possible to measure certain types of impacts.

- Primary Research Question: To what extent is it possible to measure program impacts (in social experiments) on subgroups identified by a post-treatment choice?
- Secondary Research Question: What are the methodological caveats of such an analysis and to what extent do they have implications for the interpretation of findings?

In brief, this project's focal research question relates to the possibility of conducting an experimental subgroup analysis of impacts for groups defined in terms of post-treatment defined traits.

Summary of Methodology

This article suggests two versions of an analytic approach that distinguish between and measure program impacts on certain treated subgroups. Specifically, the focus is on measuring program impacts among both participants and nonparticipants, but the method is applicable to any number of subgroups, as described later. The methodological approach draws primarily on the work of Angrist et al. (1996) and Bloom (1984) to extend instrumental variables techniques to multiple subgroups, rather than just the participant subgroup. Development of this method is motivated by practical problems of evaluation in practice.

The first stage of the process involves identifying which individuals within the treatment and control groups would be participants or nonparticipants had they been exposed to the offer of treatment. Once these subgroups are identified, using what is essentially a propensity score derived through a basic multivariate regression (e.g., a logit, probit, or linear probability model), subsequent analysis of impacts can take one of two forms. Figure 1 displays the relationship between Stage One and the two variations of Stage Two. The first analytic stage is the same, regardless whether one chooses to use a discrete or continuous subgroup indicator in the second stage.

The discrete subgroup analysis (described in the next section) basically divides the sample into groups defined by breakpoints in the propensity score: those with low propensities (e.g., <0.5) are considered to be likely nonparticipants, and those with high propensities (e.g., ≥ 0.5)

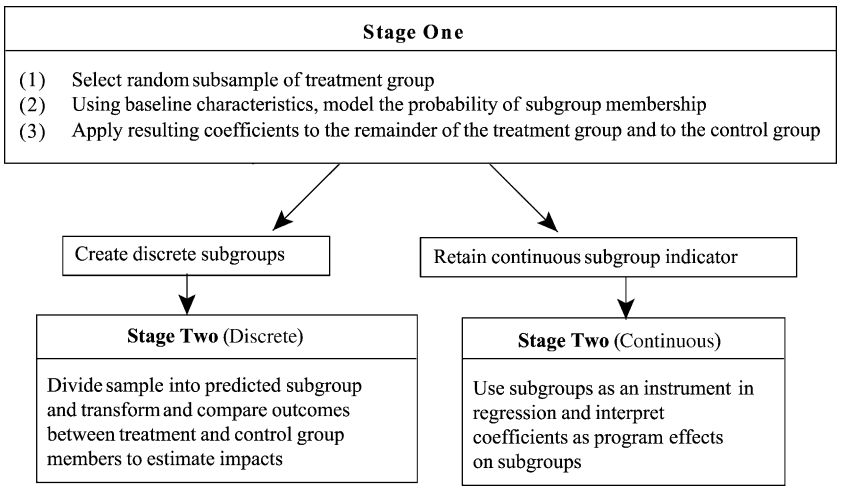

Figure 1. Flow chart of analytic process.

are considered to be likely participants. With a correction for miscategorization in the first stage, the second stage is basically a comparison of the outcomes among treatment and control group members that reveals the program's impact on the subgroups of actual nonparticipants and actual participants.

A second way to undertake such an analysis is by using the continuous subgroup indicator as an instrument (described in a subsequent section). In regression results, the coefficients associated with the indicator (instrument) itself and the indicator interacted with the treatment dummy are interpreted directly as the impact of the program on nonparticipants and participants. The analytic details, applications, and implications are the crux of the present paper.

Prior applications of a "regression-based approach" to creating subgroups (Kemple & Snipes, 2000) beg for more deliberate thinking about what these methods are, what they may offer over alternatives, and what assumptions are necessary to use them in practical applications. Hollister and Metcalf (1977), for example, examine earnings outcomes relative to what they *would have been* in the absence of the treatment according to a subgroup indicator that proxies family earnings. More recent examples include Fein et al. (1998), Kemple and Snipes (2000), and Peck (1999). These examine program impacts on individuals who were tracked into a job-ready subgroup (Fein et al., 1998), who took up the treatment offer of a generous financial incentive to work (Peck, 1999), and who would have been high school drop outs (Kemple & Snipes, 2000). These works, to a certain degree, address methodology, but their main focus is substantive.[4] They aim to explain the impacts observed in four social experiments. The present work focuses on the methodological approach and suggests necessary conditions to apply it.

A major point of motivation for this work is that individuals, once they enter a treatment, may follow any number of paths. They might not participate fully in the program offered, or they might participate in certain subsets of a multifaceted program. Because these choices take place after the point of random assignment in an experimental evaluation, this heterogeneity of treatment group experiences poses a problem to evaluators. But, *because* of random assignment, we know with confidence that any subgroup that exists within a treatment group must have a counterpart within the control group; this is the bonus of random assignment and is graphically depicted in Figure 2. In the common instance where individuals are offered the opportunity to participate in a new treatment, some will accept the offer and some will not, as shown in the left portion of Figure 2. The random assignment process provides assurance that a subset of the control group (right portion of Fig. 2) *would have* participated had they been extended the same offer. The problem is that we *know* ultimately who participates in the treatment group, but we can not easily identify that subgroup's counterparts in the control group. The valid comparison is always between the entire treatment group and entire control group or between subgroups of those identified by *baseline* characteristics. A comparison of members of the two treatment group subgroups identified above (participants and nonparticipants) clearly introduces selection bias. As a result, these comparisons are made only with extensive controls and caveats, and the classical experimental comparison is simply between the mean outcomes for those in the treatment group and the mean outcomes for those in the control group.

In turn, an impact estimated as the difference in the overall treatment and control group outcomes reflects the impact of the intention-to-treat and not actual receipt of the treatment. Although an ITT-measured impact can have policy relevance, there are other questions about the impact of the treatment on the treated that may be of interest. In addition, not only might the impact of the treatment on the treated be of interest but so too might be the impact of the treatment on the *un*treated and the difference in impacts between these two groups. To what

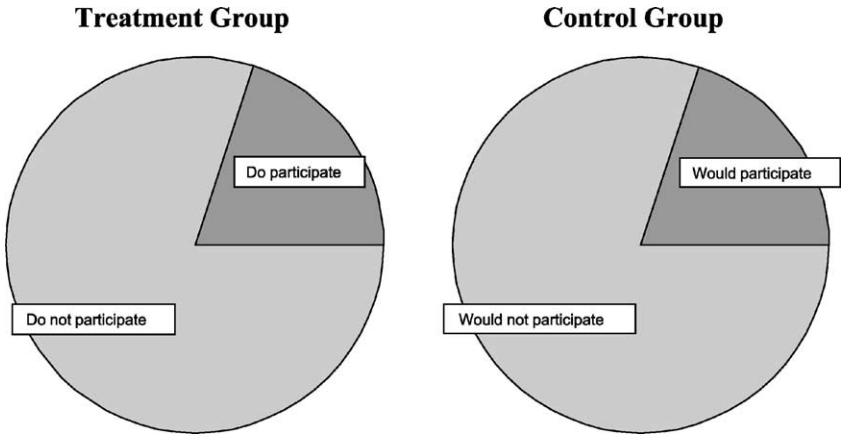

Figure 2. The bonus of random assignment: having comparable subgroups between treatment and control groups.

extent do programs influence the outcomes of individuals who engage in them, and how does the impact on these participants differ from the program's impacts on nonparticipants?

Participants and nonparticipants are two important subgroups, but the evaluation problem described here has broader implications. The evaluation problem—of there being heterogeneity within treatment groups in terms of their members' interaction with the treatment being tested—concerns not only whether or not individuals take up the treatment offer but also how social programs impact a variety of treatment group subsets. When programs being tested involve a variety of features, individuals in the treatment group may choose to engage in some elements but not others.

Welfare reform demonstrations in the 1990s involved time limits, work mandates, sanctions, extended transitional assistance, new restrictions on unwed minor parents and other features all at once. The traditional treatment-control group difference provides the impact of the combination of program features, but knowing whether participation in certain elements of the treatment causes variation in impacts across the treatment group is likely to be of interest to program evaluators and administrators. For example, it might be useful to know the effect of a time limit on individual outcomes separately from the effect of using transitional assistance, but the current state of the science prevents doing so. In addition, it can be useful to think about how certain kinds of individuals in the absence of a treatment—such as those with more or less risk or with a certain type of history (e.g., public assistance, earnings, work)—are more or less likely to benefit from program services.

For the most part, however, subgroup analyses in social experiments have used individual baseline traits, which are exogenous to the treatment, to define subgroups. Sometimes what Beecroft and Lee (2000) describe as "not readily identifiable subgroups" may be those with the greatest policy relevance, but clear approaches for how to identify these groups and analyze their impacts are not agreed upon within the field. Recent pilot tests of some interesting initiatives provide additional motivation to develop new methods. The problem of heterogeneous treatment groups exists in many evaluations and in turn limits the information one can gain from using traditional ITT-estimates of program impact. Measuring the impact on the treatment group as a whole may miss interesting impacts that accrue to a wide variety of

subgroups, in particular those defined by some post-treatment choice. The implications are important for the practice of program evaluation, for policy making that is based on findings from evaluation research, and for program administration where targeting may improve program effectiveness.

Outline

The remainder of this document details the two versions of a proposed analytic procedure that measures the impacts of social programs, evaluated through a classically designed experiment, on subgroups identified by a post-treatment choice. The next section focuses on measuring impacts on discretely-identified subgroups. It describes the two stages of the analytic process and the technical estimation requirements, as well as the assumptions necessary for empirical results to be credible. It then illustrates an application of the analysis with data from the experimental evaluation of the New York State Child Assistance Program (CAP). The section that follows proposes an alternative estimation process, a direct instrumental variable technique that uses predicted subgroup membership as the instrument. Although the first analytic stage is the same as that described in MEASURING IMPACTS ON DISCRETE SUBGROUPS, the second stage parametrically estimates program impacts on subgroups. Necessary assumptions are discussed, and the approach is then applied again to the CAP data. The final section concludes by revisiting the research questions, discussing other applications and possible extensions of this work, and discussing of the implications of this work for program design and evaluation.

MEASURING IMPACTS ON DISCRETE SUBGROUPS

The proposed method has two stages. The first stage involves estimating a model, using the baseline characteristics of a random subset of the treatment group for which a certain type of participation is known (e.g., they enrolled in CAP). The results of that model then predict the same type of participation for the remaining treatment group members and also for the control group members.

To avoid overfitting, a random subsample is used for modeling and then those observations are excluded from the subgroup analysis. Because the model will provide a better fit among the modeling subsample than it will among the rest of the sample, it is important to exclude this subsample from the Stage Two analysis to avoid introducing any bias.[5] In essence, this method identifies a subgroup that is defined by a mix of baseline traits,[6] and this subgroup is associated with the eventual path that individuals follow after entering the treatment group. By using *predicted* likelihood that individuals will be in a certain subgroup of treatment and control group members (based on exogenous baseline characteristics), an important type of selection bias is eliminated and the resulting subgroups are suitable for comparison in an experimental context.

The second stage of the analysis involves estimating impacts on the predicted subgroups, and, with the addition of certain assumptions, transforming the results of this analysis into estimates of impacts on actual subgroup members. There are two main ways to estimate these impacts. Very simply, this analysis compares the outcomes of those with predicted low (or high) scores in the treatment group to the outcomes of their counterparts in the control group. In a subsequent section, I describe the other way to operationalize the second stage of the analysis—and to estimate impacts on subgroups—which retains the continuous score as the subgroup indicator, using it as an instrument for subgroup membership in a straightforward regression model.

Stage One

A key feature of the discrete subgroup analysis described here is the ability first to distinguish between or among subgroups within a heterogeneous experimental treatment group. The simplest of subgroup analyses chooses a single individual characteristic, as measured at baseline, on which to divide the population. One might be interested in whether there is variation in program impacts among individuals with varying racial or ethnic backgrounds, for instance, or among men as compared to women, or among those with or without earnings at baseline. To do so, the treatment and control groups are segmented and impacts estimated for each subgroup.

As noted earlier, characteristics exist within the treatment group that are not easily defined within the control group and therefore pose a challenge for subgroup analysis. For example, whether a treatment group member engaged in the program is a condition that does not exist within the control group.[7] Because the control group is excluded from participating in the program, there is no obvious distinction between participants and nonparticipants among control group members; but random assignment assures that there are subsets of the control group that *would have* participated (or not) had they been in the treatment group (recall Fig. 2).

Because evaluations track the paths of individuals subject to the treatment, it is possible to identify what characteristics are associated with membership in certain treated subgroups. In turn, with information on these characteristics for both treatment and control group members, comparable subgroups can be identified. The process of identifying subgroups involves developing a model with regressors that predict who would be part of the subgroup. It then involves estimating the model's parameters from data for an existing, relevant sample. Next, it uses this model to generate a predicted score for other members of the treatment and control groups such that each individual's score identifies his or her likelihood of being part of the subgroup of interest. This predicted score is essentially a propensity score, a single number that represents a set of that individual's characteristics.

If a particular post-random assignment treatment choice can be modeled as a function of baseline characteristics, then that choice can be compared across treatment and control groups in a subgroup analysis. "Post-random assignment treatment choice" refers to any action taken by a member of the treatment or control group after the point of random assignment. Such a choice may refer to participation (or not) in the program, it may refer to participation (or not) in a certain element of a multifaceted treatment, or it may refer to the intensity with which individuals and the treatment interact. These are all subgroups that can be modeled and then compared. The approach proposed here eliminates at least one problematic type of selection bias under certain conditions and provides subgroups that are comparable within an experimental context.[8] As stated earlier, the existence of random assignment provides two statistically identical groups with the single exception that one group received a treatment of some sort. Making predictions from baseline (pre-treatment) characteristics results in identification of subgroups similar to those in Figure 2.

Creating a subgroup indicator. Identifying the subgroup to which an individual belongs requires the following four straightforward steps:

- Select a random subsample of the treatment group;
- Using baseline characteristics, model the probability of subgroup membership within the subsample;

- Apply the resulting coefficients to the remainder of the treatment group and to the entire control group; and
- Decide whether and, if so, how to break the continuous subgroup score into a discrete indicator of predicted subgroup membership.

These steps provide each treatment and control group member with a score, or index, that reflects his or her likelihood of subgroup membership. More detail on these steps follows.

Step one. By selecting a random subsample from the treatment group (and then excluding it from the impact analysis), one is, in actuality, creating an external sample to estimate the subgroup selection model. This process prevents overfitting the data. Using a modeling subsample allows an external set of individuals, with the same characteristics as the rest of the treatment and control groups, to provide information on the post-random assignment treatment choice (e.g., to participate or not). If the whole sample were used for modeling, then the model might offer a better fit among the treatment (modeling) group than it would among the control group, resulting in some unknown amount and direction of bias.[9]

Of course this step's selection of a random sample to use for modeling and exclude from the subsequent analysis is contingent on sample availability. With a sufficiently large sample, discarding a subset will have little effect; but with a smaller starting sample size, one might consider using bootstrapping methods, using repeated samples with replacement, to generate the predictive model's coefficients in the following step.

Step two. Any analytic approach that helps classify individuals is relevant to this process. For example, logit, probit and linear probability models are possibilities as are discriminant analysis or latent class models. Social experiments tend to collect rich data at baseline, and these data are the source for the predictive model that might take any of these suggested forms.

Step three. The next step involves applying the coefficients, generated from the model created in step two, to the remainder of the treatment group and to the control group. The result of this process is that each individual will have a score (or propensity, or probability, ranging from zero to one) that reflects his or her likely subgroup membership. The predicted subgroup indicator is used for defining which treatment and control group members to compare to one another.[10]

Step four. The continuous indicator of subgroup membership provides useful information about the likelihood that an individual would be part of the subgroup. In some instances, however, it might be preferable to convert the continuous indicator into a dichotomous or polytomous (categorical) indicator. Examining the distribution of scores may help assess where the logical breakpoints are. To dichotomize the score, 0.5 seems a logical breakpoint. Another possibility is to select as the breakpoint the score that maximizes correct placement of individuals (in the treatment group) into the subgroup of interest.

Creating a categorical indicator of subgroup membership might involve selecting breakpoints associated with the distribution of scores. In such an instance, the bottom 25th percentile would represent those with a low chance or risk of subgroup membership, the middle 50 percent those with a medium chance or risk, and the 75th percentile and above those with a high chance or risk of subgroup membership. These choices should be made based on knowledge of the program and population being studied. For example, if the program being

studied had participants and nonparticipants, then clearly two is the appropriate number of subgroups.[11]

Assessing the quality of the subgroup indicator. Although it is not possible to judge the accuracy of the model in placing control group members in correct subgroups, it is possible to do so among treatment group members because one knows their actual subgroup membership. Unlike a standard regression for continuous outcomes, statistical models for discrete outcomes are not well assessed by an R^2 (or pseudo-R^2) statistic. Instead, correct placement of individuals into subgroups serves as a reasonable proxy for judging the model's predictive ability. To the extent that the model discriminates among the subgroups, the model is providing useful information for the subgroup analysis. In contrast, poor predictive ability in Stage One affects the external validity of estimates generated in Stage Two.

Stage Two

For discrete predicted subgroups—either binary or categorical—the impact estimation procedure compares the treatment and control group outcomes for individuals within each subgroup. Assuming a dichotomous subgroup indicator (e.g., predicted nonparticipants or predicted participants), the comparisons would be as follows:

$$\Delta_{\hat{n}} = \bar{Y}_{t\hat{n}} - \bar{Y}_{c\hat{n}}$$

$$\Delta_{\hat{p}} = \bar{Y}_{t\hat{p}} - \bar{Y}_{c\hat{p}}$$

where Δ, refers to the impact, or change in outcomes; $\bar{Y}$, is the average outcome of interest; the subscripts t and c refer to treatment and control group members, respectively; and the subscripts $\hat{n}$ and $\hat{P}$ refer to predicted nonparticipants and predicted participants, respectively. Increasing the number of subgroup categories of interest increases the number of treatment-control comparisons to be made.[12]

These impact estimates, Δ, as described above, represent the treatment-control differences in outcomes among *predicted* subgroups, and it is these *predicted* subgroups that are comparable between treatment and control groups. But ultimately we are interested in the impacts of subgroup membership on *actual* subgroup members. In order to convert these results from predicted to actual, straightforward algebra and one of two assumptions are needed. In brief, the estimated impact on actual subgroup members is a weighted sum of the predicted impact, where the weight is associated with the proportion of individuals correctly placed in the subgroup.

To explain, begin by considering graphically the placement of actual nonparticipants and actual participants into predicted nonparticipant and predicted participant cells, as in Figure 3. For simplicity, assume that we segmented each of the treatment and control groups into two subgroups, predicted nonparticipants and predicted participants, based on their having a propensity score of <0.5 and ≤ 0.5, respectively.[13] Within the treatment group's subgroup of predicted nonparticipants, for instance, a certain proportion will be actual nonparticipants. Likewise, within the treatment group's subgroup of predicted participants, a certain proportion of the group will be actual participants. Because the treatment and control groups are statistically the same, this means that had we the information on control group members' actual participation status, it would reflect the same distribution of correct placement. Let π represent the proportion of the predicted subgroups that is comprised of members of its corresponding actual subgroup.

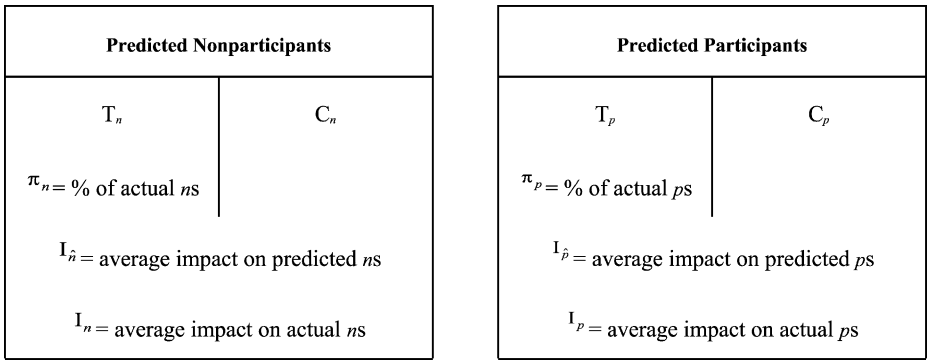

Figure 3. Notation for placement of actual nonparticipants and actual participants into predicted nonparticipant and predicted participant categories and resulting impacts.

Additional notational definitions are necessary and noted in Figure 3. Two impact-related estimates are of interest. The first is the impact on predicted subgroup members. These are $I_{\hat{n}}$ and $I_{\hat{p}}$, respectively the impacts on predicted nonparticipants and predicted participants. The second is the impact on actual subgroup members. These are I_n and I_p, respectively the impacts on actual nonparticipants and actual participants. Although the πs are generated with knowledge on just the treatment group members, one must make the assumption that the proportion would apply to control group members as well. Having random placement into treatment and control groups means that this assumption is met to an increasingly good approximation as sample size increases.

The process for converting impacts on predicted subgroups to impacts on actual subgroups involves solving for I_n and I_p from the following two equations:

$$I_{\hat{n}} = \pi_n I_n + (1 - \pi_n) I_p \tag{1}$$

$$I_{\hat{p}} = \pi_p I_p + (1 - \pi_p) I_n \tag{2}$$

Equation (1) is the impact on predicted nonparticipants, and Equation (2) is the impact on predicted participants. Basically, Equation (1) states that the impact on predicted nonparticipants is a weighted sum of the impacts on actual participants and actual nonparticipants, where the weights represent the proportion of individuals correctly identified as nonparticipants. Likewise, Equation (2) reveals that the impact on predicted participants is a weighted sum of the impacts on actual participants and actual nonparticipants, where the weights represent the proportion of individuals correctly identified as participants.

The numbers $I_{\hat{n}}$ and $I_{\hat{p}}$ are easy to compute from existing data: they are the measured impacts on predicted nonparticipant and predicted participant subgroups, respectively. But I_n and I_p—the impact on actual nonparticipants and the impact on actual participants—are the unknowns in these equations that are of ultimate interest. Equations (1) and (2) are written as they are because they contain both the elements that one can easily generate ($I_{\hat{n}}$, $I_{\hat{p}}$, π_n and π_p) and the elements that are desirable (I_n and I_p); knowing the impacts on *actual* subgroups is the goal of this analysis.

A simple system of equations with two unknowns, as this one, can be solved with straightforward algebra. This involves taking one of the equations, rearranging it to solve for either In I_n and I_p and then taking the resulting equation and substituting it into the other

equation, rearranging and solving for the other of I_n and I_p. The result of this rearrangement follows:

$$I_n = \frac{\pi_p I_{\hat{n}} - (1 - \pi_n) I_{\hat{p}}}{\pi_n + \pi_p - 1} \tag{3}$$

$$I_p = \frac{\pi_n I_{\hat{p}} - (1 - \pi_p) I_{\hat{n}}}{\pi_n + \pi_p - 1} \tag{4}$$

Equations (3) and (4) show that the impact of the program on actual subgroup members is a weighted sum of the program impacts on the two predicted subgroups, where the weights come from the proportion of correctly placed actual subgroup members within the predicted subgroups.

Just as Equations (1) and (2) compute the impacts of predicted nonparticipants and predicted participants in terms of the impacts on actual nonparticipants and actual participants, Equations (3) and (4) compute the impacts of actual subgroup membership in terms of predicted subgroup membership. Equations (1) and (2) must be rearranged to produce Equations (3) and (4) so that the unknowns can be solved through the known elements.[14]

In sum, this process shows how to compute the impact of the program on actual nonparticipants and actual participants, or, in other words, the effect of the treatment on the untreated and on the treated, provided the assumption that follows holds. It also allows the estimation of impacts on nonparticipants, which has particular salience when one has reason to believe that the program may influence even those who do not participate. Furthermore, this strategy for estimating impacts on multiple subgroups extends to groups defined in any way, not solely as nonparticipants and participants in a program. The ability of this approach to estimate impacts on multiple subgroups, including both nonparticipants and participants, is its value added over prior work.

Assumptions. In order to benefit from the additional information that the proposed method offers, an underlying assumption about the relationship between predicted and actual subgroup membership must hold true. But first, two sources of potential selection bias require discussion. The first type of selection bias has to do with internal validity, the second with external validity. The discussion of Figure 2 highlighted one source of that bias—comparing subgroup members to nonmembers within a treatment group. This type of selection bias poses no problem for the current analysis because the approach proposes comparing subgroups between treatment and control groups, where the subgroups are determined exclusively through the use of baseline data (exogenous to the treatment). So, the comparison of *predicted* subgroups across treatment and control groups generate results that are internally valid; the estimated impact is in fact the *impact* on the subgroup. But how might we describe what that subgroup represents? It is a subset of individuals *predicted* to follow some post-treatment choice, but to the extent that they would not make that choice in actuality, this limits the analysis's external validity.

This second source of selection bias is potentially more problematic and pertains to whether the *predicted* subgroups are at all like the *actual* subgroups. If the predicted and actual subgroups correspond perfectly then there is no selection bias. If they correspond to a high degree then there is very little selection bias. But, if the predicted subgroup members are quite different from the actual subgroup members, then the results of the analysis are challenged. The analysis may show that the *impact* on some *predicted* subgroup is X, but if we do not know

what being in that predicted subgroup means then the results are of little use. In order to deal with the problems associated with this second variation of selection bias (and therefore for the analytic approach to achieve its desired results), one of two assumptions must be made.

The first assumption is that impacts of actual subgroup membership are constant across predicted subgroups, and the second is that the mean impact of actual subgroup membership is uncorrelated with the likelihood of predicted subgroup membership (or the characteristics used to predict it). With application to nonparticipant and participant subgroups, the first assumption is that the impacts of nonparticipation and of participation are constant. If this is the case, then Equations (1) and (2) clearly imply Equations (3) and (4). That is, the actual elements in Equations (1) and (2) must be identical, and if they are then rearranging the elements to create Equations (3) and (4) is possible. The second assumption is that the mean impact of nonparticipation or of participation is uncorrelated with predicted nonparticipation or predicted participation, respectively. Each of these assumptions is discussed in turn below.

Most stringent assumption. The most stringent assumption necessary in order to undertake this analysis is as follows:

> *The impact of nonparticipation is the same for all nonparticipants, and the impact of participation is the same for all participants.*

In other words, impacts are *constant* within each actual subgroup. This assumption makes possible the conversion of impacts on predicted groups to impacts on actual groups through the weighting scheme described above. Assuming constant impacts of actual subgroup membership within each predicted subgroup allows estimation of impacts on any number of subgroups.

To a certain degree, the necessity of having to make this assumption may seem at odds with the reason for undertaking this analysis in the first place. Heterogeneity of treatment-subject interaction and the related potential heterogeneity in program impacts on treated subgroups is, in part, what motivates this work. Although it may be necessary, it is certainly not desirable to make the assumption of constant impacts given real world data (that probably rarely exhibit this feature).

Weaker assumption. The analysis is still possible with a less stringent assumption:

> *The mean impact of nonparticipation does not depend on predicted nonparticipation* $(1-\hat{P})$*, and the mean impact of participation does not depend on predicted participation* $\hat{P}$*.*

Rather than requiring constant impacts, this assumption allows for the possibility of varying impacts within subgroups. Instead, it requires that the impact is the same *on average* across the actual subgroups, regardless of first-stage placement, or that subgroup impacts be attributable to the treatment received by the subgroup and not to the characteristics of the individuals who comprise the subgroup.[15] This is more likely to be true when there is high predictability from the initial prediction model, resulting in less reclassification between *predicted* and *actual* subgroups.

In other words, this assumption requires that the impact of subgroup membership not be correlated with the likelihood of subgroup membership. This assumption allows for the possibility of heterogeneous impacts but requires the restriction that the *likelihood* of participation (or nonparticipation) does not affect impacts. If people participate for purely random reasons, then the assumption holds. Evidence supporting the viability of this assumption comes from Michalopolous and Schwartz's (2000) exhaustive subgroup analysis of the programs making

up the National Evaluation of Welfare to Work Strategies (NEWWS). They report that a variety of impacts accrue to a variety of individuals and that programs with many features work with diverse target populations. If this suggests a low correlation between specific traits and the presence of program impacts, then a sensible extension is to conclude an almost-random association between individual characteristics and program impacts. Nevertheless, substantial evidence exists to suggest that selection into programs, based on unobservable characteristics, exists and therefore an imperfect placement of individuals into predicted subgroups means that some bias will influence generalization of findings.

Application to the Child Assistance Program

This section applies the two stages of the discrete subgroup analysis described above to the data from New York State's Child Assistance Program.

Application of Stage One. In 1988, the New York State Department of Social Services began to pilot test the Child Assistance Program, which underwent a rigorous experimental evaluation that followed 4,287 families in three counties for five years (Hamilton et al., 1996). CAP is a welfare reform initiative that changes the incentive and support structure of traditional welfare in an attempt to increase self-sufficiency among single parents otherwise reliant on the state for support. Designed with explicit incentives to motivate recipients to take steps toward financial self-sufficiency, CAP's main work incentive is its low tax rate on earnings where benefits are reduced gradually until recipients' income exceeds the federally-defined poverty line. Additional CAP features that encourage responsibility include requiring families to secure court-ordered child support from absent parents, giving families cash rather than food stamp coupons, and eliminating AFDC's limit on the amount of assets a family can have and remain eligible. CAP also attempts to remove welfare stigma by conveying a non-welfare image with separate, professional-looking office space and case management. Although CAP case workers assist active CAP participants, they also encourage AFDC recipients to seek the jobs or support orders needed to make CAP worthwhile for them (Hargreaves, 1992). This program is relevant to the current discussion because members of the treatment group who were offered the opportunity to enroll in CAP may have changed their behavior in response to the offer regardless of whether they ever enrolled in CAP. As a result, it is not realistic to assume that only CAP enrollees within the treatment group had changed outcomes.

The first stage of the analysis models what factors predict whether someone enrolls in CAP. With CAP participation (ever enrolled over the five-year follow-up period) as the dependent variable, explanatory variables include the grantee's demographic characteristics, the household's needs and resources, the characteristics of the absent parents and status of child support, the grantee's welfare history, and indicators for the county.[16] Table 1 shows the results of this model, which was estimated on a random sample of one-fifth of the treatment group.

Next, these coefficients are applied to the remaining four-fifths of the treatment group and to the control group to create predicted participation scores. Because both the actual participation status and the predicted participation scores are known for the treatment group, it is possible to examine how well the model performs in correctly identifying CAP participants and nonparticipants. Shown in Figure 4, each set of bars totals 100 percent, indicating the percentage of actual nonparticipants and participants that comprise each quantile of predicted participants. Where predicted participation score exceeds 0.5, the proportion of actual

TABLE 1.
Logistic Model of CAP Participation

Variable	*Priors*	*Coefficient*	*Odds Ratio*
Grantee's demographic characteristics			
Divorced	?	−0.08	0.92
Widowed, separated, or currently married	?	−0.50[a]	0.61
Black	?	−0.02	0.98
Hispanic	?	−0.19	0.83
Less than age 25	?	0.07	1.07
Age 35 and older	?	−0.33	0.72
High School Graduate	+	0.37[a]	1.45
Household's needs and resources			
Multiple parents adults present	–	−0.35	0.71
Two or more children present	?	0.66[a]	1.94
Child(ren) under age three present	–	0.08	1.08
Grantee has earnings	+	0.70[a]	2.02
Amount of grantee's earnings ($100/month)	+	0.10[a]	1.00
Grantee's longest job lasted at least one year	+	0.37[a]	1.45
Characteristics of absent parents and child support status			
Number of distinct absent parents	?	0.01	1.01
Number of children with support orders	+	0.18	1.20
Number of children lacking support orders	–	−0.35[a]	0.70
Absent parent(s) of children without support order living in state	+	0.30	1.35
Absent parent(s) of children without support orders ever in contact with custodial parent	+	0.51[a]	1.67
Absent parent(s) of children without support orders ever married to custodial parent	+	0.57[a]	1.76
Grantee's welfare history			
On AFDC a total of over two years	–	−0.14	0.87
Demonstration sites			
Niagara County	+	0.40[a]	1.49
Suffolk County	–	−0.50[a]	0.61
Intercept		−2.56[a]	
N		424	
Association of predicted probabilities and observed responses			
Concordant		75.8%	
Discordant		23.9%	
Tied		0.3%	

[a] Standard error < coefficient.

participants in the predicted group exceeds the proportion of nonparticipants in the group.[17] Although there is some overlap between actual and predicted participation status, there is a clear distinction between actual participants' and actual nonparticipants' predicted participation scores. Figure 4 shows that actual participants are more likely (than actual nonparticipants)

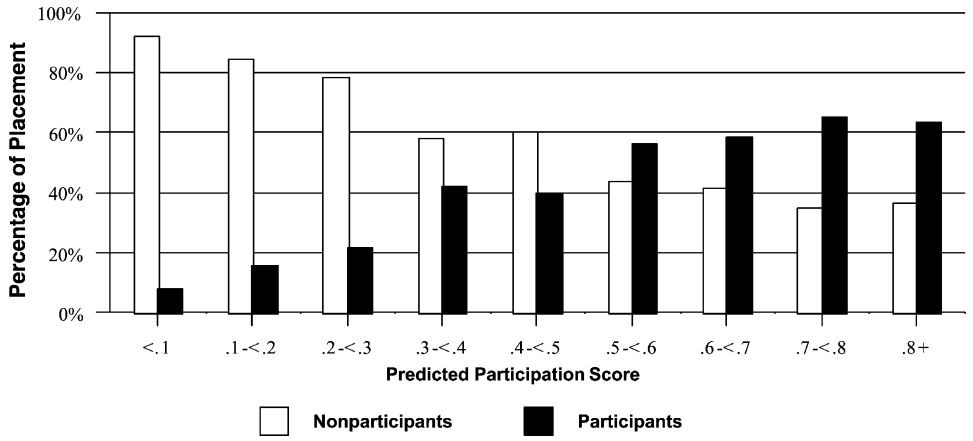

Figure 4. Comparison of actual and predicted participation status among treatment group members: percent of correct placement.

to have higher predicted participation scores, and actual nonparticipants are more likely (than actual participants) to have low predicted participation scores.

The observation that the groups are not entirely distinct may raise concerns about the external validity of such a subgroup analysis, described earlier as the external validity type of selection bias. Use of the predicted subgroups is a means to an end; it serves to eliminate the internal validity kind of selection bias, thereby segmenting the treatment and control groups into relevant subsets. To the extent that the first-stage model creates *any* distinction between groups, the model is providing valuable information that allows subsequent subgroup analysis. And if impacts on actual subgroup members are the same on average across the predicted subgroups, then the conversion of impacts from predicted to actual is viable, though making conclusions about *actual* rather than *predicted* subgroup members has its conditions.

Application of Stage Two. After computing the predicted participation scores for both treatment and control group members, the next step in the process is to compare program impacts among those with similar scores. This involves first converting the score into a dichotomous indicator and comparing the treatment and control group outcomes for those in each subgroup, as shown in Table 2.

These results reflect the impacts on the *predicted* subgroups. In order to generate the impacts on actual participants and nonparticipants, it is necessary to apply information, from Stage One, on the correct placement of actual individuals into the predicted subgroups. Using 0.5 as the cut-off, Table 2 shows the correct placement of actual nonparticipants and participants within predicted nonparticipant and participant categories, respectively.

The numbers in Table 3 can be explained from the vantage point of correct predicted subgroup placement: Predicted nonparticipants are comprised of 78.5 percent actual nonparticipants and 21.5 percent actual participants, and the predicted participant subgroup is comprised of 38.7 percent actual nonparticipants and 61.3 percent actual participants. These are the numbers needed to create the weights used to convert impacts on predicted subgroup members to impacts on actual subgroup members. Specifically, what is needed is the percent of actual participants within each predicted group (21.5 percent and 61.3 percent,

TABLE 2.
Earnings Outcomes and Impacts on Predicted Nonparticipants and Predicted Participants (Unadjusted)

	Predicted Nonparticipants ($\hat{P}$ <.5)				*Predicted Participants ($\hat{P}$ <.5)*			
	Treatment	*Control*	*Impact*	*% Change*	*Treatment*	*Control*	*Impact*	*% Change*
Year 1	$2,706	$2,450	$256***	10.5	$6,239	$5,374	$865***	16.1
Year 2	3,502	3,119	383***	12.3	6,501	6,150	352*	5.7
Year 3	4,006	3,574	432***	12.1	7,476	7,547	−72	−1.0
Year 4	4,871	4,067	804***	19.8	8,333	8,337	−4	−0.0
Year 5	5,399	4,713	686***	14.6	9,304	8,930	374	−4.2
Overall	$20,435	$17,898	$2,557***	14.3	$37,829	$36,297	$1,532	4.2

Significance levels: (*) $p < .10$, (**) $p < .05$, (***) $p < .01$.

TABLE 3.
Correct Placement of Dichotomous Subgroups

	Actual Nonparticipants	*Actual Participants*	*Total*
Predicted Participants	1,121	307	1,428
	78.5%	21.5%	
Predicted Nonparticipants	120	190	310
	38.7%	61.3%	
	1,241	497	1,739

because one minus these numbers is used in the computations to generate the correct set of weights).

In order to back out the impacts on actual subgroup members from the impacts estimated for predicted subgroup members (shown in Table 2), the proportions that describe correctly placed actual subgroup members (within predicted subgroups) are the appropriate ones to consider. The weights described in the prior section can be computed as shown in Table 4,

TABLE 4.
Computation of Weights from Percentages of Correct Placements

	Proportions				*Weights*
Definition	π_n	π_p	*Numerator*	*Denominator*	
$W_{n\hat{n}} = \dfrac{\pi_p}{\pi_n + \pi_p - 1}$	.785	.613	.613	.398	1.54
$W_{n\hat{p}} = \dfrac{(1-\pi_n)}{\pi_n + \pi_p - 1}$	.785	.613	.215	.398	0.54
$W_{p\hat{p}} = \dfrac{\pi_n}{\pi_n + \pi_p - 1}$	.785	.613	.785	.398	1.97
$W_{p\hat{n}} = \dfrac{(1-\pi_p)}{\pi_n + \pi_p - 1}$	.785	.613	.385	.398	0.97

TABLE 5.
Earnings Impacts on Actual Nonparticipants and Participants

	Impact on Actual Nonparticipants	*Impact on Actual Participants*
Year 1	−$73	$1,457***
Year 2	400**	322
Year 3	704***	−561
Year 4	1,241***	−790
Year 5	855***	71
Overall	$3,111***	$535

Significance levels: (*) $p < .10$, (**) $p < .05$, (***) $p < .01$.

where the left column reports the formulas defined in Appendix A, the next columns identify the appropriate numbers from Table 3, and the right column combines these to show the computed weights. The application of these weights to the program impacts by predicted subgroup (reported in Table 2) results in values, shown in Table 5, that are the program's impacts on actual nonparticipants and actual participants. For example, recall from Table 2 that the year one impact on predicted nonparticipants is $256, and the year one impact on predicted participants is $865. Therefore, the year one impact on actual nonparticipants is $(W_{n\hat{n}} \times 256) + (W_{n\hat{P}} \times 865)$ and on actual participants is $(W_{p\hat{n}} \times 865) + (W_{p\hat{P}} \times 256)$. These computations reflect an application of the relative proportions of actual subgroup members that appear in the predicted subgroups to the predicted subgroups' impacts. They also reflect the identifying assumptions by showing that the impacts on predicted subgroup members are the same across the computations. The assumption of constant or equal mean impacts is necessary to allow these computations (standard errors are computed using the computational scheme described in Appendix A).

These results show that CAP impacts the earnings of both nonparticipants and participants, and that the timing of impacts differs between the two subgroups. Although participants are responsible for the impacts in year one, it is nonparticipants who generate positive earnings impacts in later years.

A possible explanation for the participant impacts is that CAP motivated its participants to increase their work effort sooner than they would have without the offer. Over time, however, participants in the control group caught up to their treatment group counterparts. This type of finding is not uncommon in evaluations of job training programs, where earnings gains have been shown to be difficult to sustain over time. Accepting the offer of CAP had the effect of increasing short-term earnings.

A possible explanation for the impacts on nonparticipants is consistent with the earlier discussion about how the CAP offer worked. That is, at any give time, CAP's treatment group was comprised of AFDC recipients, some of whom were making efforts to become eligible for CAP, and of CAP recipients. Because CAP case manager outreach targeted individuals while they were on AFDC, before they became CAP-eligible, it makes sense that there would be impacts in general among the CAP nonparticipant group. Some treatment group members made major efforts to become eligible for CAP, and this is reflected in the fact that the program shows positive impacts on the subgroup of nonparticipants. Although the argument for positive subgroup impacts at all among nonparticipants is clear, the explanation for these impacts' being greater in the later years is somewhat less clear. In brief, CAP's service provision appears to have raised the earnings of treatment group members who did not enroll in CAP.

Discussion of assumptions in practice. How realistic are the strong and weak assumptions discussed earlier in the application of this analytic approach to CAP? Although it is difficult to imagine a scenario in which the stronger assumption holds, it may be reasonable to assume that the second assumption holds. First, the problem that motivated this inquiry is that of treatment group and impact heterogeneity. To make the assumption that dividing the experimental sample into just two groups eliminates all the possibilities of variation is naive. That is, even within the two subgroups examined here, it is plausible that impacts are not constant, but they may be the same on average.

Nevertheless, is may be reasonable to make the assertion that it is the program and not individual characteristics that generate the observed variation in program impacts. The weaker assumption can be thought to hold when we consider the fact that there is variation in the type of individual that falls into the subgroups analyzed here. That is, the subgroup of nonparticipants is diverse, as is the subgroup of participants. As a result, the measured subgroup impacts are likely the result of variations in experience with the program instead of variation in personal traits. Stated another way, the expected mean value of the impacts on actual nonparticipants and actual participants is the same regardless of predicted subgroup membership. Predicted subgroup membership is the vehicle through which it is possible to estimate impacts on actual subgroups.

The assumption may not hold if placement of predicted subgroup members into actual subgroups is imperfect, as in the CAP example. Such imperfect placement (61 percent of participants and 79 percent of nonparticipants are correctly placed) has implications for the results' external validity. Some influence of unobservable characteristics means that generalizeability to actual subgroups is limited by the analysis.

MEASURING IMPACTS USING CONTINUOUS SUBGROUP INDICATORS

One important reason to use the approach described in the previous section is that it creates predicted subgroups that are easy to comprehend (i.e., someone is either in a predicted subgroup or not); as such, it has face validity. A continuous score, however, contains more information than does its dichotomized counterpart. In order to retain the information inherent in the continuous score, this section considers using it as an instrument for subgroup membership in a simple multivariate regression. The particular analysis discussed here extends instrumental variables estimation in a new direction to help provide information about the impacts of social programs on subgroups of policy interest. After discussing some general uses and properties of instrumental variables, the section describes the regression equation and the instruments used to generate subgroup impact estimates. The section then illustrates again with CAP.

Stage One Revisited

Recall from the previous section that Stage One models the probability of being in an experimental subgroup. This involves first selecting a random subsample of the treatment group and then using baseline characteristics to model the probability of subgroup membership within this subsample. Next, applying the resulting coefficients to the remainder of the treatment group and to the entire control group generates a score for each individual. This score reflects each individual's likelihood of being in the subgroup. The Stage One process is the same whether one chooses to use the dichotomous or continuous measure as a subgroup indicator in Stage

Two.[18] In brief, Stage One produces a single, continuous (propensity) score that represents a group of individual characteristics that is associated with an individual's ultimate path through the program. This section describes how to use the continuous score as an instrument to estimate program effects on certain subgroups.

Stage Two

Instrumental variables analysis is used to solve one of three types of problems: omitted variable bias, measurement error in an independent variable (the so-called errors-in-variables problem), or reciprocal causality (or simultaneity bias). Of these three problems, the current analysis faces the first, omitted variable bias. Specifically, the variable—an indicator of participation status or any subgroup membership—that one would like to include in the analysis is missing. Without this variable, one uses an instrument that serves as a proxy for the omitted variable. The IV solution to the problem describe here, then, is to include an instrument in place of the omitted variable. In this particular application, the instrument used is the predicted subgroup scores that were estimated in a first-stage regression. Because subgroup status is unknown for those in the control group (remember, they were not extended the offer to participate), it is necessary to estimate the following model where predicted subgroup membership is a proxy for actual subgroup membership:

$$Y_i = \alpha_0 + \alpha_1 \hat{P} + \beta_0 T_i + \beta_1 T_i \hat{P}_i + \varepsilon_i$$

where Y is the outcome of interest; T is a binary variable equal to 1 for treatment group members and 0 for control group members; $\hat{P}$ is the predicted probability of being in the subgroup, or the instrument for participation; and is the subscript i indexes individuals.

To illustrate how to interpret the coefficients in this model, consider the instance in which the predicted participation (the instrument for participation) is zero. The elements remaining in the equation are α_0 and $\beta_0 T$. That is, when $\hat{P} = 0$, the value of α_0 is the mean *outcome* for those in the control group (with participation scores of zero), and the value of β_0 is the *impact* of being in the subgroup among those in the treatment group. In contrast, assuming that the predicted participation score equals one, the elements remaining in the equation can be rewritten as: $Y = (\alpha_0 + \alpha_1) + (\beta_0 + \beta_1)T$. This shows that the outcome for control group members in the subgroup (when $\hat{P} = 1$) is the sum of the coefficients α_0 and α_1 and that the estimated impact on treatment group members in the subgroup is the sum of β_0 and β_1.

That the subgroup indicator would be either zero or one is unrealistic, though imagining it as such aids in explaining how to interpret the coefficients. Predicted scores may approach zero or one, but they are more likely to fall in between. Nevertheless, because $\hat{P}$ is an *instrument* for the effect of nonparticipation, and because $\hat{P}$ interacted with the treatment dummy is an instrument for the incremental effect of participation, the coefficients on these terms can be interpreted *directly* as the program's impacts on these subgroups.

Assumptions. What is necessary to undertake this analysis is that the conditions of instrumental variable estimation be met. Specifically, the instrument of choice should be uncorrelated either with the error terms or with the other explanatory variables. The necessary assumption is that $\hat{P}$ is a good (exogenous) instrument for participation. As stated earlier:

> *The mean impact of nonparticipation does not depend on predicted nonparticipation* $(1-\hat{P})$, *and the mean impact of participation does not depend on predicted participation* $(\hat{P})$.

The weaker assumption asserted in the context of the discrete version of the analysis is essentially the same for the continuous (IV) version of the analysis. The language used to describe the necessary assumption in the continuous version of the subgroup analysis derives simply from basic IV estimation. Basically, because the characteristics used to create the instrument in the first place are completely exogenous to treatment and control group status, the instrument is therefore appropriate; but if there are interaction effects (variation in impacts by baseline characteristics) then the instrument becomes weaker.

Determining statistical significance. In any two stage procedure where a first-stage-generated predicted value is used as a covariate in the second stage, it is necessary to compute the standard error accordingly. The use of an instrument introduces an additional source of error into the model, and this error must be accounted for in order to compute correctly the standard error of the parameter estimates. Standard statistical packages, such as SAS, have appropriate commands for implementing this process in computing two-stage or IV parameter estimates.[19]

With the models described above, standard regression output will provide tests of significance for each coefficient (e.g., β_0). As a result, determining whether the impact measured for a particular subgroup is statistically different from zero is straightforward. It is likely, however, that the analyst would be interested not only in knowing whether the sum of the coefficients (e.g., $\beta_0 + \beta_1$, the impact on the participant subgroup) is statistically significant but also in knowing whether the impacts measured for each subgroup are statistically different from each other. That is, does the program cause important variation in impacts across treated subgroups? Additional computation is necessary in order to answer this question.[20]

Application to the Child Assistance Program

The results from the first analytic stage that were presented in the previous section need not be repeated. The second analytic stage differs when using the continuous subgroup indicator as an instrument (compared to using the discrete subgroup indicator). These second-stage results from analysis of the CAP data are presented below.

Application of Stage Two. Starting with the participation scores identified in Stage One, impact estimation involves using these scores in a regression to generate impacts on nonparticipants and participants. The score alone and the score interacted with treatment status are included in the model as instruments. As discussed earlier, the coefficients on these variables are interpreted directly as the effect of the program on nonparticipants and on participants.[21]

As Table 6 shows, the earnings impacts that CAP achieved varied by year and by subgroup: nonparticipants experience no impacts at first but then positive impacts in later years, whereas participants experience the reverse, positive impacts at first and none in later years.

Findings from the discrete subgroup analysis and the continuous subgroup analysis tell roughly the same story: Participant impacts are larger in year one and then decrease, whereas nonparticipant impacts are larger (and significant) in the later years of follow-up. As with the discrete version of the subgroup analysis, these findings support that even CAP nonparticipants are influenced by the treatment's offer. This suggests that enrollment in CAP is only one route to improved outcomes; the extension of CAP-preparation services through case management, for example, increased the earnings of the treatment group as a whole regardless of program take-up, with impacts on nonparticipants being the more substantial of the two subgroups.

TABLE 6.
Earnings Outcomes and Impacts on Nonparticipants and Participants, Based on Coefficients from Instrumental Variables Regressions

	Nonparticipants				*Participants*			
	Treatment	*Control*	*Impact*	*% Change*	*Treatment*	*Control*	*Impact*	*% Change*
Year 1	$986	$825	$161	19.6	$8,392	$7,428	$964	13.0
Year 2	1,832	1,489	343	23.0	8,067	7,534	533	7.1
Year 3	2,180	1,407	772	54.9	9,360	9,977	−616	−6.2
Year 4	2,896	1,684	1,212**	72.0	10,325	10,916	−591	−5.4
Year 5	3,377	2,314	1,063	45.0	10,612	10,937	−325	−3.0
Overall	$11,224	$7,718	$3,506*	45.4	$46,731	$46,644	$88	0.2

Significance levels: (*) $p < .10$, (**) $p < .05$, (***) $p < .01$. The hypothesis test that the subgroup impacts are equal to each other is accepted; subgroup impacts are not statistically significantly different from each other.

Main differences between the two sets of results have to do with the absolute size of the impacts and the likelihood of statistical significance. Measured impacts on nonparticipants are about the same in the continuous version of the estimation and the discrete version ($3,506 and $3,111, respectively). Similarly, the IV-estimated impacts on participants are about the same size as those estimated through the discrete subgroup analysis ($88 and $535, respectively, both of which are indistinguishable from zero). The IV-estimated results have greater standard errors and are therefore less often statistically significant compared to the discrete analysis's results. This relatively large difference in the size of standard errors may result from the fact that the continuous version of the method takes into consideration first-stage error in computing its standard errors, whereas the discrete version of the method does not; as a result, the standard errors are larger when first-stage error is accounted for in the analysis.

Another difference between the two versions' results has to do with the cross-group allocation of impacts. Overall, the treatment-control difference is $2,527, and the IV-estimation restricted the coefficients such that this overall impact is allocated as appropriate across the subgroups. The discrete version of the analysis does not incorporate this feature, and the overall impact (not by subgroup) that it actually allocates is $2,374, which is about 94 percent of the known total impact.

Discussion of assumptions in practice. In order for the results of the IV-estimation to be useful, the instrument used must be a good one. That is, the variables used to predict subgroup status in the first stage should be unrelated both to program impacts and to the unobserved components of impact. If the subgroup indicator is predicted by variables that also are associated with program impacts, then the IV-estimated model may be mis-estimating the unique contribution of subgroup membership to explaining the outcome variable. It is important, then, to search explicitly for predictors of subgroup membership that are not also predictors of impact.

Because the process of random assignment was implemented correctly—there are only random differences in background characteristics between treatment and control group members—all of the variables used to predict subgroup membership are exogenous in terms of treatment receipt and therefore also program impact. Nevertheless, there may be some

interaction effects that suggest that the treatment is more effective for individuals with certain types of characteristics. To the extent that any of these variables help generate the instrument, the instrument is weaker than it would be if there were no interaction effects.

In the application of this analysis to the CAP data, some predictor variables clearly are unrelated to the outcome earnings, but some of the variables used to predict participation may also be associated with earnings. Whether these variables explain earnings *levels* does not matter because their effects are netted out when computing the program's *impacts*.[22] One would expect that those things that explain earnings and child support order behavior are predictors of CAP participation. If the program offer affects the relationship between prior earnings and post-program earnings, then there is some element of the instrument that will be contaminated. An example of ideal subgroup membership predictor variables for CAP are whether a family has one or more child support orders in place or the characteristics of the absent parent (that might make it harder or easier for a single-parent to secure a support order). These characteristics are associated with higher CAP participation but are unrelated to earnings levels. Only if some of the predictor variables have interaction effects with the program treatment do they compromise the instrument. Unfortunately, it is not possible to test whether each of the subgroup prediction variables has an impact on the dependent measure of earnings, because any influence they might have on the outcome would be confounded with participation or nonparticipation in the program.

Another observation that stems from the analysis's application to CAP is the difference in results between the discrete and continuous approaches. This difference suggests that the underlying assumptions may not be fully satisfied. One would expect that with greater correct placement of predicted subgroup members into their actual subgroup, the difference in results between the two approaches would diminish. Additional work in this area is warranted in order to understand better not only the tradeoffs between the discrete and continuous analytic approaches but also the extent to which the underlying assumptions hold.

Although the assumptions necessary to undertake this analysis are reasonable in some instances, it is also important to consider what might happen when the assumptions do not hold. If, for example, impacts are larger on actual participants with low propensity scores (making them predicted nonparticipants) than they are on actual participants with high propensity scores, then the estimated impacts on participants would be understated and the estimated impacts on nonparticipants would be overstated. Similarly, if impacts are positively correlated with estimated propensity to participate, then this factor is confounded with actual participation, suggesting that the impacts of participation would be overstated and the impacts of nonparticipation would be are understated. The reverse if true is impacts are negatively correlated with the estimated propensity to participate. If the bias were in the same direction for both actual subgroups' impacts, then the difference in impacts between the groups is likely to be estimated as greater than it should be. In brief, the consequences of the assumption of constant impacts or the assumption of mean independence *not* holding are potentially great. Nevertheless, if the assumptions seem realistic given program and data knowledge, then the method suggested here will produce results that may provide useful information about a program's subgroup impacts.

Summary

In the evaluation of social experiments, instruments have been used in a specific way. They have been created from random assignment status and been used to estimate the impact

of program participation. For example, if a particular program element is the only way for treatment group members to have achieved a certain outcome, then placement in the treatment group is an ideal instrument for measuring the effect of that program element. Although, if there are competing pathways to achieve program outcomes (and impacts), a single instrument is insufficient (Bloom, Hill, & Riccio, 2001).

In lieu of using the random assignment status itself as the instrument, as other applications of instrumental variables in experiments have done, the approach described here uses selected baseline characteristics to generate an instrument. Because these characteristics are exogenous to receipt of the treatment, they provide a way to generate an instrument that captures the effect of a post-random-assignment treatment choice (such as whether to participate or not). If the variables that predict subgroup membership are not also associated with the program's impact, then the resulting instrument is ideal. The resulting instrumental variables estimates are more valid but less precise than would be OLS estimates of program impacts. This occurs because a limited portion of the total variation in subgroup membership is explained by the instrument. Nevertheless, the regression's resulting coefficients are interpreted directly as impact of the program on members and nonmembers of the subgroup.

DISCUSSION, CONCLUSION AND FUTURE DIRECTIONS

This study is motivated by problems of program evaluation practice. Evaluations' estimates of mean impacts do not provide potentially useful detail on subgroup impacts. Further, some subgroups—particularly those identified by a post-treatment choice—have received limited examination in prior analyses even though they might be policy-relevant.

Summary of Process and Findings

Because this is a methodological project, the research questions it aims to answer, rather than being about particular program impacts, are about whether and under what conditions it is possible to measure certain types of impacts.

By proposing and developing two variants of a new evaluation method, this article aims to analyze more deliberately the impacts of social programs on subgroups. The methodological approach is a two-stage process in which predicted subgroup membership is identified in a first stage and then used as a covariate in a second stage to estimate program impacts by subgroup, by using either a discrete or a continuous subgroup indicator. As the second and third sections of this article demonstrate, it is possible to measure a program's subgroup impacts in one of two, straightforward ways by capitalizing on the predictive ability of individuals' pre-program traits.

Applying both variations of this analytic method to data from the evaluation of New York State's Child Assistance Program yields findings that are both substantive (about CAP) and methodological (about the analytic process). Learning about the impacts that accrue to members of treated subgroups, which are identified by a post-treatment choice, provides more information than assessing average impact alone. In the case of CAP, subgroup impacts differ from each other and over time: program participants showed positive earnings impacts immediately but not later in the follow-up period, and the program's offer impacted nonparticipants' earnings as well. Information from CAP's process evaluation provides corroborating support for results gained through this subgroup analysis.

This project's substantive finding about CAP's impacts fits well into the debate about subgroup impacts in welfare employment programs. Early subgroup analyses (e.g., Boudett & Friedlander, 1997; Friedlander, 1993) found that program impacts are greater for those people on the margins; that is, programs are more effective for the least advantaged subgroups. More recently however, evaluators have found that programs that offer a wide variety of services are effective for a wide variety of individuals (Michalopolous & Schwartz, 2000). That those with varied demographic backgrounds benefit from varied welfare reforms is a finding that strengthens this analysis's necessary underlying assumption.

Other Applications and Future Directions

Although applying this method to the CAP data highlights how this method can help identify impacts among participants and nonparticipants in a treatment, the broader application is to any kind of subgroup, including those that engage in only a certain portion of a treatment or who engage in the treatment with varying levels of intensity. Understanding how programs affect individuals conditional on some post-treatment choice or condition is important, but it is an analytic problem that the program evaluation field has not yet solved. The method that I propose uses a treatment group condition and compares outcomes to those of counterparts identified in the control group. But, subgroups need not be defined by the characteristics of the treatment group. Another starting point might be to use baseline characteristics within the control group to create predicted subgroups within the treatment group that *would have* been distinctly different in the absence of the treatment. This approach allows varied comparisons of outcomes but follows the same analytic technique.

Hollister and Metcalf (1977) and Kemple and Snipes (2000) used a set of individual traits from the control group to predict behavior or status in the absence of the treatment and apply those conditions to the treated. In the evaluation of the Career Academies program, for instance, Kemple and Snipes model the likelihood of dropping out of school in the absence of the program and then predict comparable subgroups among the treated. Using CAP as another example, one might predict long-term welfare use in the absence of the program and examine how the program offer did or did not impact the outcomes of the comparable subset of the treatment group.

In the case of multifaceted programs, the subgroups of interest may be more complex. For example, in a welfare reform demonstration that involves time limits, sanctions for program noncompliance and extended medical and child care assistance, various subsets of the treatment group are likely to be influenced in varying ways by each of these program features. If it is possible to model the probability of being punished by welfare program sanctions, then one subgroup analysis might compare the outcomes of those sanctioned with the outcomes of those not sanctioned. Similarly, if it is possible to model the probability of taking up transitional child care, then it would be possible to estimate the impact of that program feature separately from others. This type of subgroup analysis can apply to any intervention that includes more than one feature, and it improves evaluators' ability to describe what in particular about a given intervention is responsible for its impacts.

A clear extension of this work calls for analyzing existing (and forthcoming) experimental data with attention to new kinds of subgroups. Certainly we can learn a lot about how programs achieve their impacts (or lack of impacts) by examining how treated subsets are affected. Although the examples offered here have focused on welfare and employment related programs, any program evaluated experimentally can benefit from added analyses of discrete subgroups.

For example, experiments in the realms of education, housing, criminal justice, public health, and psychology can provide rich data, the analysis of which can inform public policy making and program design. In addition, more detailed analysis of the CAP data—including using income, in addition to earnings, as an outcome measure—would be useful.

Another extension of this work can involve undertaking more sophisticated ways of identifying subgroups, including latent class modeling and cluster analysis. Latent class modeling, for example, intends to classify individuals by some unobservable characteristic. Nagin and Land's (1993) work in the field of criminology has been seminal in explaining the types of criminal careers that exist among heterogeneous criminal populations. Without their analysis and subsequent work in the field, criminal theory would misclassify individuals as similar who actually show very different behavioral patterns. Latent class modeling is relevant to experimental subgroup analysis in that it provides another way to identify subgroups, the number of which might not be obvious.

Cluster analysis is another method for grouping common observations and may help reveal groups of individuals with similar behavior or patterns within experimental treatment groups. Groups identified by a focus on baseline characteristics can be compared to similar control subgroups. The work of Yoshikawa, Rosman, and Hsueh (2001) using cluster analysis with experimental treatment group data is one example that reveals the possibilities for extending the cluster approach to experiments in general. Clearly the better able one is to identify subgroups of interest within treatment and control groups, the more interesting and relevant will be the results achieved by this project's proposed method. Regardless of the exact method used to identify subgroups, subgroup membership can still be used as an instrument in the second analytic stage in order to estimate impacts on members and nonmembers of the subgroups.

Implications for Program Design and Evaluation

Perhaps a defining characteristic of the employment and welfare demonstrations evaluated through experiments is that they have generated modest impacts at best. A possible explanation for such minimally meaningful impacts and the associated disappointment with social programs' ability to achieve change is the lack of serious examination of impacts across heterogeneous populations. In instances where measured impacts are indistinguishable from zero, for example, the policy implication suggests terminating the seemingly unsuccessful program. A recent examination of an employment and training demonstration reveals that the program's average treatment effect (of zero) failed to tell the story of what happened to the treatment group: those "who completed key program components may have benefited from them, while those who dropped out prematurely experienced only the opportunity cost of their participation" (Bos, 1995, p. 98). The information that some treated individuals may be positively affected while others are negatively affected is much more useful than the average treatment effect. Moreover, as Heckman et al. (1997) describe, society might be willing to support programs that have little or even no overall impact if they have the right kinds of impacts on certain subsets of individuals. Without further examination of how impacts accrue among treated subgroups, this information is not readily available. In order to make the choice about whether to terminate or redesign a seemingly ineffective program, full information about how the program impacts varying subsets of the target population is necessary.

This research generates suggestions for the next round of social experiments. If estimating program impacts based on a post-treatment choice will continue to be of interest, then it will be important to collect baseline data that improves evaluators' ability to predict

subgroup membership along these lines (thereby reducing challenges to the analysis's necessary assumptions). The baseline characteristics used to predict CAP membership, for example, are observable characteristics measured at baseline, but what if there is selection on unobservable characteristics that influences program participation? Commonly considered "unobservables," characteristics such as motivation level, propensity to engage in certain activities, or personal preference for certain kinds of services are important predictors of subgroup membership at least in the application to CAP. Evaluators could consider developing a battery of questions that captures "motivation" or "propensity to engage" or "preference for . . . ," which would be useful in identifying relevant subgroups in future analyses.

This research also suggests increased use of multi-stage random assignment processes. Rather than using the method that I propose, the design of an experiment can provide information on intention-to-treat effects separately from the effect of the treatment on the treated. A multi-stage random assignment research design might proceed as follows. People first would be assigned to receive the offer or not. Then, upon expressing intent to enroll in the program, people would be randomly assigned to receive program services or not. This design would allow us to know the impact of a program's offer separate from the impact of participating in the program itself. Although this design is attractive in theory, it is not commonly used to evaluate the difference between impacts of program offers and actual receipt of services. Any number of stages of random assignment might take place, and each would allow an experimental comparison of subgroups of individuals who follow varying treatment paths. Because multi-stage random assignment can be complex in practice, the method I propose and describe can be useful for answering questions about evident in many existing interventions, and new analyses can help program administrators, evaluators and analysts learn more about how social programs work.

ACKNOWLEDGMENTS

I am grateful to the following people for their useful contributions to this work: Howard S. Bloom, Jan Blustein, William Greene, Erik Beecroft, Hans Bos, Robinson Hollister, Tod Mijanovich, Robert Yaffe, Steve Baer, Brad Snyder, four anonymous reviewers and Melvin M. Mark, editor of the *American Journal of Evaluation*. Early support for this project was provided by the Daniel B. McGillis Professional Development and Dissemination Grant program, under the guidance of Chris Hamilton, Steve Kennedy, and Larry Orr at Abt Associates.

APPENDIX A: DETERMINING STATISTICAL SIGNIFICANCE

This appendix provides the framework for computing and understanding the standard errors that correspond to the impact estimates derived in the test. Standard errors of those estimates are, of course, necessary to know whether the estimates are statistically significant. To derive them, consider some information from Equations (3) and (4) in MEASURING IMPACTS ON DISCRETE SUBGROUPS; in particular, recall the combination of probabilities associated with each of the components ($I_{\hat{n}}$ and $I_{\hat{p}}$). For simplicity, let us rename the weights as follows:

$$W_{n\hat{n}} = \frac{\pi_p}{\pi_n + \pi_p - 1} \qquad W_{n\hat{P}} = \frac{(1 - \pi_n)}{\pi_n + \pi_p - 1}$$

$$W_{p\hat{p}} = \frac{\pi_n}{\pi_n + \pi_p - 1} \qquad W_{p\hat{n}} = \frac{(1 - \pi_p)}{\pi_n + \pi_p - 1}$$

The first subscript of the weight indicates which of Equations (3) and (4) the weight belongs to, and the second subscript indicates the term within that equation. This notation shows that, for example, $W_{n\hat{n}}$ is the weight to be applied to the $I_{\hat{n}}$ component of Equation (3) that solves for I_n, and that $W_{n\hat{P}}$ is the weight to be applied to the $I_{\hat{p}}$ component of the same equation. In turn, we can rewrite Equations (3) and (4) as follows, to reflect that they generate the estimated impacts on actual nonparticipant and actual participant subgroups:

$$\hat{I}_n = W_{n\hat{n}}\hat{I}_{\hat{n}} - W_{n\hat{p}}\hat{I}_{\hat{p}} \tag{A.1}$$

$$\hat{I}_p = W_{p\hat{p}}\hat{I}_{\hat{p}} - W_{p\hat{n}}\hat{I}_{\hat{n}} \tag{A.2}$$

Because the predicted subgroups represent independent samples, the following represents the variance:

$$\text{var}(\hat{I}_n) = (W_{n\hat{n}})^2\text{var}(\hat{I}_{\hat{n}}) + (W_{n\hat{p}})^2\text{var}(\hat{I}_{\hat{p}}) \tag{A.3}$$

$$\text{var}(\hat{I}_p) = (W_{p\hat{p}})^2\text{var}(\hat{I}_{\hat{p}}) + (W_{p\hat{n}})^2\text{var}(\hat{I}_{\hat{n}}) \tag{A.4}$$

Given Stage One results, the square root of the variance of each estimate (for $\hat{I}_n$ and $\hat{I}_p$) is its standard error. This calculation of the standard error incorporates information about the correct-placement rate of actual subgroup members into predicted subgroups by virtue of using π_n and π_p.

NOTES

1. The ITT-impact estimate may be preferred for understanding the likely results of program replication when program participation would remain voluntary.

2. Note, however, that sometimes the "treatment" is actually an *offer* to engage in a new program, but those extended that offer are, in their entirety, still called the treatment group, even if some did not take up the offer.

3. Although this discussion has focused on terms relevant primarily to the treatment group (and subgroups thereof), other analyses are concerned with "cross-overs" from the treatment group to the control group. This is also called "control group contamination" and is less central to the analysis presented here. It should be noted that the term "complier" is also used sometimes to refer to those who comply with their treatment or control group *status*—that is those who are not cross-overs—and not just those who take up the treatment offer.

4. Fein et al. (1998) and Kemple and Snipes (2000) relegate the methodological discussion to an appendix, with focus in the main text on evaluation findings. Similarly, Hollister and Metcalf (1977) and Peck (1999) discuss briefly their methods and also focus discussion on the substantive results.

5. In their similar analysis, Kemple and Snipes (2000) use the entire group for modeling and also for comparison. They acknowledge the potential bias that this introduces and attempt to quantify it.

6. This subgroup indicator, a single variable that represents a set of individual characteristics, is technically a propensity score, although my practical use of it differs from how others have used propensity scores. Rather than create better comparison groups for nonexperimental analysis (e.g., Dehejia & Wahba, 1999), I use the propensity score to identify subgroups within the treatment and control groups in experimental data.

7. Likewise, the opposite might be true: one might want to examine the outcomes for those in the treatment group based on what would have occurred in the absence of the treatment.

8. The classic approach for dealing with problems of selection bias is Heckman's two-stage selection correction procedure (Heckman & Robb, 1985a, 1985b), which accounts for the non-random selection that characterizes the enrollment of individuals into training programs in particular.

9. Although I have described the use of a subset of the experimental sample here, it would also be possible to use a separate sample altogether. Such a sample could come, for example, from a distinct yet similar experiment or sample, or could be drawn from within the general population.

10. If instead one were to use the actual subgroup indicator in the treatment group and compare that to the predicted subgroup indicator in the control group, bias, in some unknown quantity and direction, would be introduced. The comparison of *predicted* participants in the treatment group, for example, to *predicted* (would-be) participants in the control group assures that the two groups are alike and absent of the kind of selection bias that would affect the analysis's internal validity.

11. If there is no a priori expectation about the number of subgroups (e.g., those based on the amount of treatment received) that exist within a treatment group, more sophisticated analysis, such as latent class modeling or cluster analysis, may help reveal underlying subgroups among treated cases.

12. As the number of subgroups increases, the analysis approaches that of the continuously-identified subgroup analysis, which is described in the next section.

13. As I hope will become clear later, the cut-off point that places individuals into predicted nonparticipant and predicted participant cells becomes irrelevant when we apply the weighting scheme and make the assumption of constant impacts or of mean independence.

14. If there were perfect first-stage classification (that is π_n and π_p both equal one), then the equations would simply reduce to $I_{\hat{n}} = I_n$ and $I_{\hat{p}} = I_p$.

15. A proof, developed by Howard S. Bloom, that this weaker assumption allows the estimation of unbiased estimators appears in Peck (2002).

16. The model was estimated with and without the site indicators, and the results were essentially the same.

17. Given that just 16 percent of the treatment group ever enrolled in CAP, one might argue that predicting that 100 percent of the treatment group were nonparticipants would achieve a higher rate of correct subgroup placement. That is, 84 percent of those predicted to be nonparticipants, under this rudimentary approach, were actually nonparticipants. Since 84 percent correct is higher than the correct-placement rates reported here, it might appear that this classification scheme be preferred to the regression-based approach. The problem, however, is that under the predict-that-everyone-is-a-nonparticipant approach, 100 percent of those who ever enrolled in CAP would be misidentified as nonparticipants. Since participants are one of the two main subgroups in this analysis, it is unacceptable not to identify any of them.

18. Given that using the continuous subgroup indicator places this analysis squarely within an instrumental variables framework, it is worth noting that there are similarities between my first-stage analysis and what Angrist and Krueger (1994) call "split sample instrumental variables" (SSIV).

19. For example, The SAS System's SYSLIN procedure either will compute the instrument in its processes or will allow a previously-estimated variable (with error) to be accounted for appropriately in the error computations.

20. An *F*-test, which can be requested as additional output in any standard statistical software, will document whether certain combinations of coefficients are statistically different from zero.

21. It should be noted that The SAS System's SYSLIN procedure allows elements of the model to be restricted. To make use of the information available, this analysis restricts the coefficients on the treatment indicator (T) and the interaction of the treatment indicator with predicted participation ($\hat{P}$) to reflect the overall impact of the program. That is, the simple difference in outcomes between the entire treatment and control groups over the five years is $2,527, which implies that the weighted impact of the subgroups should equal the same amount.

22. This is the main benefit of having an experimental design, that events or phenomena that affect the treatment group also affect the control group.

REFERENCES

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association, 91*, 444–455.

Angrist, J. D., & Krueger, A. B. (1994). *Split sample instrumental variables*. Cambridge, MA: National Bureau of Economic Research, Technical Working Paper #150.

Beecroft, E., & Lee, W. S. (2000). *Looking beyond mean impacts to see who gains and who loses with time-limited welfare: Evidence from the Indiana welfare reform evaluation*. Paper presented at the Annual Meeting of the National Association for Welfare Research and Statistics, Scottsdale, AZ.

Bloom, H. S. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review, 8*, 225–246.

Bloom, H. S., Hill, C. J., & Riccio, J. (2001). *Modeling the performance of welfare-to-work programs: The effects of program management and services, economic environment, and client characteristics*. New York, NY: Manpower Demonstration Research Corporation Methodology Working Paper.

Bos, J. M. (1995). The labor market value of remedial education: Evidence from time series data on an experimental program for school dropouts (Doctoral dissertation, Robert F. Wagner Graduate School of Public Service, New York University, 1995). *Dissertation Abstracts International, 56*, 4139.

Boudett, K. P., & Friedlander, D. (1997). Does mandatory basic education improve achievement test scores of AFDC recipients? A reanalysis of data from California's GAIN Program. *Evaluation Review, 21*, 568–588.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association, 94*, 1053–1062.

Eberwein, C., Ham, J. C., & LaLonde, R. J. (1997). The impact of being offered and receiving classroom training on the employment histories of disadvantaged women: Evidence from experimental data. *Review of Economic Studies, 64*, 655–682.

Fein, D. J., Beecroft, E., Hamilton, W., Lee, W. S., Holcomb, P. A., Thompson, T. S., & Ratcliffe, C. E. (1998). *The Indiana welfare reform evaluation: Program implementation and economic impacts after two years*. Cambridge, MA: Abt Associates Inc.

Frangakis, C. E., & Rubin, D. B. (2000). The defining role of "principal causal effects" in comparing treatments using general post-treatment variables. In *Proceedings of the Epidemiology Section, American Statistical Association* (pp. 23–32).

Friedlander, D. (1993). Subgroup impacts of large-scale welfare employment programs. *Review of Economics and Statistics, 75*, 138–143.

Friedlander, D., & Robins, P. K. (1997). The distributional impacts of social programs. *Evaluation Review, 21*, 531–553.

Hamilton, W. L., Burstein, N. R., Baker, A. J., Earle, A., Gluckman, S., Peck, L., & White, A. (1996). *The New York State Child Assistance Program: Five year impacts, costs, and benefits*. Cambridge, MA: Abt Associates Inc.

Hargreaves, M. (1992). *The New York Child Assistance Program: Interim report on implementation*. Cambridge, MA: Abt Associates Inc.

Heckman, J. J. (1999). *Accounting for heterogeneity, diversity and general equilibrium in evaluating social programs*. Cambridge, MA: National Bureau of Economic Research, Working Paper #7230.

Heckman, J. J., & Robb, R., Jr. (1985a). Alternative methods for evaluating the impact of interventions. In J. Heckman & B. Singer (Eds.), *Longitudinal analysis of labor market data* (pp. 145–245). New York, NY: Cambridge University Press.

Heckman, J. J., & Robb, R., Jr. (1985b). Alternative methods for evaluating the impact of interventions: An overview. *Journal of Econometrics, 30*, 239–367.

Heckman, J., Smith, J., & Clements, N. (1997). Making the most out of program evaluations and social experiments: Accounting for heterogeneity in program impacts. *Review of Economic Studies, 64*, 487–535.

Heckman, J., Smith, J., & Taber, C. (1998). Accounting for dropouts in evaluations of social programs. *Review of Economics and Statistics, 130*, 1–14.

Hirano, K., Imbens, G. W., Rubin, D., & Zhou, X. (1999). Assessing the effect on an influenza vaccine in an encouragement design. Unpublished manuscript.

Hollister, R. G., & Metcalf, C. E. (1977). Family labor-supply response in the New Jersey experiment. In H. Watts & A. Rees (Eds.), *The New Jersey income-maintenance experiment: Volume II: Labor supply responses* (pp. 185–220). New York, NY: Academic Press.

Imbens, G. W., & Rubin, D. B. (1997). Estimating outcome distributions for compliers in instrumental variable models. *Review of Economic Studies, 64*, 555–574.

Kemple, J. J., & Snipes, J. C. (2000). *Career academies: Impacts on students' engagement and performance in high school*. New York, NY: Manpower Demonstration Research Corporation.

Manski, C. F. (1995). *Identification problems in the social sciences*. Cambridge, MA: Harvard University Press.

Manski, C. F. (1996). Learning about treatment effects from experiments with random assignment of treatments. *Journal of Human Resources, 31*, 709–733.

Manski, C. F. (1997). The mixing problem in program evaluation. *Review of Economic Studies, 64*, 537–553.

Michalopolous, C., & Schwartz, C. (2000). *What works best for whom: Impacts of 20 welfare-to-work programs by subgroup*. Washington, DC: U.S. Department of Health and Human Services, Office of Assistant Secretary for Planning and Evaluation and Administration for Children and Families; and U.S. Department of Education.

Nagin, D. S., & Land, K. C. (1993). Age, criminal careers, and population heterogeneity: Specification and estimation of a nonparametric mixed poisson model. *Criminology, 31*, 327–362.

Peck, L. (1999). *Do social programs affect nonparticipants? Evidence from the child assistance program*. Paper presented at the New York University Robert F. Wagner Graduate School of Public Service Doctoral Colloquium, New York, NY.

Peck, L. R. (2002). Subgroup analysis in social experiments. (Doctoral dissertation, Robert F. Wagner Graduate School of Public Service, New York University, 2002). *Dissertation Abstracts International, 63*, 359.

Yoshikawa, H., Rosman, E. A., & Hsueh, J. (2001). Variation in teenage mothers' experiences of child care and other components of welfare reform: Selection processes and developmental consequences. *Child Development, 72*, 299–317.