

The Economics and Econometrics of Active Labor Market Programs

James J. Heckman, University of Chicago

Robert J. LaLonde, Michigan State University

and

Jeffrey A. Smith, University of Western Ontario

Prepared for the Handbook of Labor Economics, Volume III, Orley Ashenfelter and David Card, editors. We thank Susanne Ackum Agell for her helpful comments on Scandinavian active labor market programs and Costas Meghir for very helpful comments on Sections 1-7.

The Economics and Econometrics of Active Labor Programs
Contents

1. Introduction
2. Public Job Training and Active Labor Market Policies
3. The Evaluation Problem and the Parameters of Interest in Evaluating Social Programs
 - 3.1 The Evaluation Problem
 - 3.2 The Counterfactuals of Interest
 - 3.3 The Counterfactuals Most Commonly Estimated in the Literature
 - 3.4 Is Treatment on the Treated an Interesting Economic Parameter?
4. The Prototypical Solutions to the Evaluation Problem
 - 4.1 The Before-After Estimator
 - 4.2 The Difference-in-Differences Estimator
 - 4.3 The Cross Section Estimator
5. Social Experiments
 - 5.1 How Social Experiments Solve the Evaluation Problem
 - 5.2 Intention to Treat and Substitution Bias
 - 5.3 Social Experiments in Practice
 - 5.3.1 Two Important Social Experiments
 - 5.3.2 The Practical Importance of Dropping Out and Substitution
 - 5.3.3 Additional Problems Common to All Evaluations
6. Econometric Models of Outcomes and Program Participation
 - 6.1 Uses of Economic Models
 - 6.2 Prototypical Models of Earnings and Program Participation
 - 6.3 Expected Present Value of Earnings Maximization
 - 6.3.1 Common Treatment Effect
 - 6.3.2 A Separable Representation
 - 6.3.3 Variable Treatment Effect
 - 6.3.4 Imperfect Credit Markets
 - 6.3.5 Training as a Form of Job Search
 - 6.4 The Role of Program Eligibility Rules in Determining Participation
 - 6.5 Administrative Discretion and the Efficiency and Equity of Training Provision
 - 6.6 The Conflict between the Economic Approach to Program Evaluation and the Modern Approach to Social Experiments
7. Non-experimental Evaluations
 - 7.1 The Problem of Causal Inference in Non-experimental Evaluations

- 7.2 Constructing a Comparison Group
- 7.3 Econometric Evaluation Estimators
- 7.4 Identification Assumptions for Cross-Section Estimators
 - 7.4.1 The Method of Matching
 - 7.4.2 Index Sufficient Methods and the Classical Econometric Selection Model
 - 7.4.3 The Method of Instrumental Variables
 - 7.4.4 The Instrumental Variable Estimator as a Matching Estimator
 - 7.4.5 IV Estimators and the Local Average Treatment Effect
 - 7.4.6 Regression Discontinuity Estimators
- 7.5 Using Aggregate Time Series Data on Cohorts of Participants to Evaluate Programs
- 7.6 Panel Data Estimators
 - 7.6.1 Analysis of the Common Coefficient Model
 - 7.6.2 The Fixed Effects Method
 - 7.6.3 U_t Follows a First-Order Autoregressive Process
 - 7.6.4 U_t is Covariance Stationary
 - 7.6.5 Repeated Cross-Section Analogs of Longitudinal Procedures
 - 7.6.6 The Fixed Effect Model
 - 7.6.7 The Error Process Follows a First-Order Autoregression
 - 7.6.8 Covariance Stationary Errors
 - 7.6.9 The Anomalous Properties of First Difference or Fixed Effect Models
 - 7.6.10 Robustness of Panel Data Methods in the Presence of Heterogeneous Responses to Treatment
 - 7.6.11 Panel Data Estimators as Matching Estimators
- 7.7 Robustness to Biased Sampling Plans
 - 7.7.1 The IV Estimator and Choice-Based Sampling
 - 7.7.2 The IV Estimator and Contamination Bias
 - 7.7.3 Repeated Cross-Section Methods with Unknown Training Status and Choice-Based Sampling
- 7.8 Bounding and Sensitivity Analysis
- 8. Econometric Practice
 - 8.1 Data Sources
 - 8.1.1 Using Existing General Survey Data Sets
 - 8.1.2 Using Administrative Data
 - 8.1.3 Collecting New Survey Data
 - 8.1.4 Combining Data Sources
 - 8.2 Characterizing Selection Bias
 - 8.3 A Simulation Study of the Sensitivity of Nonexperimental Methods
 - 8.3.1 A Model of Earnings and Program Participation

- 8.3.2 The Data Generating Process
- 8.3.3 The Estimators We Examine
- 8.3.4 Results from the Simulations
- 8.4 Specification Testing and the Fallacy of Alignment
- 9. Indirect Effects, Displacement, and General Equilibrium Treatment Effects
 - 9.1 Review of Traditional Approaches to Displacement and Substitution
 - 9.2 General Equilibrium Approaches
 - 9.2.1 Davidson and Woodbury
 - 9.2.2 Heckman, Lochner, and Taber
 - 9.3 Summary on General Equilibrium Approaches
- 10. A Survey of Empirical Findings
 - 10.1 The Objectives of Program Evaluations
 - 10.2 The Impact of Government Programs on Labor Market Outcomes
 - 10.3 The Findings from U.S. Social Experiments
 - 10.4 The Findings from Non-experimental Evaluations of U.S. Programs
 - 10.5 The Findings from European Evaluations
- 11. Conclusions

1 Introduction

Public provision of job training, of wage subsidies and of job search assistance is a feature of the modern welfare state. These activities are cornerstones of European “active labor market policies,” and have been a feature of U.S. social welfare policy for more than three decades. Such policies also have been advocated as a way to soften the shocks administered to the labor markets of former East Bloc and Latin economies currently in transition to market-based systems.

A central characteristic of the modern welfare state is a demand for “objective” knowledge about the effects of various government tax and transfer programs. Different parties benefit and lose from such programs. Assessments of these benefits and losses often play critical roles in policy decision-making. Recently, interest in evaluation has been elevated as many economies with modern welfare states have floundered, and as the costs of running welfare states have escalated.

This chapter examines the evidence on the effectiveness of such welfare state active labor market policies such as training, job search and job subsidy policies, and the methods used to obtain the evidence on their effectiveness. Our methodological discussion of alternative approaches to evaluating programs has more general interest. Few U.S. government programs have received such intensive scrutiny, and have been subject to so many different

types of evaluation methodologies as has governmentally-supplied job training. In part, this is due to the fact that short run measures of government training programs are more easily obtained and are more readily accepted. Outcomes such as earnings, employment, and educational and occupational attainment are all more easily measured than the outcomes of health and public school education programs. In addition, short run measures of the outcomes of training programs are more closely linked to the “treatment” of training. In public school and health programs, a variety of inputs over the life cycle often give rise to measured outcomes. For these programs, attribution of specific effects to specific causes is more problematic.

A major focus of this chapter is on the general lessons learned from over thirty years of experience in evaluating government training programs. Most of our lessons come from American studies because the U.S. government has been much more active in promoting evaluations than have other governments, and the results from the evaluations are often used to expand – or contract – government programs. We demonstrate that recent studies in Europe indicate that the basic patterns and lessons from the American case apply more generally.

The two relevant empirical questions in this literature are (i) adjusting for their lower skills and abilities, do participants in government employment and training programs benefit from these programs? and (ii) are these programs worthwhile social investments? As currently constituted, these programs are often ineffective on both counts. For most groups of participants, the benefits are modest, and at worst participation in government programs is harmful. Moreover, many programs and initiatives can not pass a cost-benefit test. Even when programs are cost effective, they are rarely associated with a large scale improvement in skills. But, at the same time, there is substantial heterogeneity in the impacts of these programs. For some groups these programs appear to generate significant benefits both to the participants and to society.

We believe that there are two reasons why the private and social gains from these programs are generally small. First, the per-capita expenditures on participants are usually small relative to the deficits that these programs are being asked to address. In order for such interventions to generate large gains they would have to be associated with very large internal rates of return. Moreover, these returns would have to be larger than what is estimated for private sector training (Mincer, 1993). Another reason that the gains from these programs are generally low is that these services are targeted toward relatively unskilled and less able individuals. Evidence on the complementarity between the returns to training and skill in the private sector suggests that the returns to training in the public sector should be relatively low.

We also survey the main methodological lessons learned from thirty years of evaluation activity conducted mainly in the United States. We have identified eight lessons from the

evaluation literature that we believe should guide practice in the future. First, there are many parameters of interest in evaluating any program. This multiplicity of parameters results in part because of the heterogeneous impacts of these programs. As a result of this heterogeneity, some popular estimators that are well-suited for estimating one set of parameters are poorly suited for estimating others. Understanding that responses to the same measured treatment are heterogeneous across people, that measured treatments themselves are heterogeneous, that in many cases people participate in programs based in part on this heterogeneity and that econometric estimators should allow for this possibility, is an important insight of the modern literature that challenges traditional approaches to program evaluation. Because of this heterogeneity, many different parameters are required to answer the interesting evaluation questions.

Second, there is inherently no method of choice for conducting program evaluations. The choice of an appropriate estimator should be guided by the economics underlying the problem, the data that are available or that can be acquired, and the evaluation question being addressed.

A third lesson from the evaluation literature is that better data helps a lot. The data available to most analysts have been exceedingly crude as we document below. Too much has been asked of econometric methods to remedy the defects of the underlying data. When certain features of the data are improved, the evaluation problem becomes much easier. The best solution to the evaluation problem lies in improving the quality of the data on which evaluations are conducted and not in the development of formal econometric methods to circumvent inadequate data.

Fourth, it is important to compare comparable people. Many non-experimental evaluations identify the parameter of interest by comparing observationally different persons using extrapolations based on inappropriate functional forms imposed to make incomparable people comparable. A major advantage of nonparametric methods for solving the problem of selection bias is that, rigorously applied, they force analysts to compare only comparable people.

Fifth, evidence that different non-experimental estimators produce different estimates of the same parameter does not indicate that non-experimental methods cannot address the underlying self-selection problem in the data. Instead, different estimates obtained from different estimators simply indicate that different estimators address the selection problem in different ways and that non-random participation in social programs is an important problem that deserves more attention in its own right. Different methods produce the same estimates only if there is no problem of selection bias.

Sixth, a corollary lesson, derived from lessons three, four and five, is that the message from LaLonde's (1986) influential study of nonexperimental estimators has been misunderstood. Once analysts define bias clearly, compare comparable people, know a little about

the unemployment histories of trainees and comparison group members, administer them the same questionnaire and place them in the same local labor market, much of the bias in using nonexperimental methods is attenuated. Variability in estimates across estimators arises from the fact that different nonexperimental estimators solve the selection problem under different assumptions, and these assumptions are often incompatible with each other. Only if there is no selection bias would all evaluation estimators identify the same parameter.

Seventh, three decades of experience with social experimentation have enhanced our understanding of the benefits and limitations of this approach to program evaluation. Like all evaluation methods, this method is based on implicit identifying assumptions. Experimental methods estimate the effect of the program compared to no programs at all when they are used to evaluate the effect of a program for which there are few good substitutes. They are less effective when evaluating ongoing programs in part because they appear to disrupt established bureaucratic procedures. The threat of disruption leads local bureaucrats to oppose their adoption. To the extent that programs are disrupted, the program evaluated by the method is not the ongoing program that one seeks to evaluate. The parameter estimated in experimental evaluations is often not likely to be of primary interest to policy makers and researchers, and under any event has to be more carefully interpreted than is commonly done in most public policy discussions. However, if there is no disruption, and the other problems that plague experiments are absent, the evidence from social experiments provides a benchmark for learning about the performance of alternative non-experimental methods.

Eighth, and finally, programs implemented at a national or regional level affect both participants and nonparticipants. The current practice in the entire “treatment effect” literature is to ignore the indirect effects of programs on nonparticipants by assuming they are negligible. This practice can produce substantially misleading estimates of program impacts if indirect effects are substantial. To account for the impacts of programs on both participants and nonparticipants, general equilibrium frameworks are required when programs substantially impact the economy.

The remainder of the chapter is organized as follows. In Section 2, we distinguish among several types of active labor market policies and describe the types of employment and training services offered both in U.S. and in Europe, their approximate costs, and their intended effects. We introduce the evaluation problem in Section 3. We discuss the importance of heterogeneity in the response to treatment for defining counterfactuals of interest. We consider what economic questions the most widely used counterfactuals answer. In section 4, we present three prototypical solutions to the problem cast in terms of mean impacts. These prototypes are generalized throughout the rest of this chapter, but three basic principles introduced in this section underlie all approaches to program evaluation when the

parameters of interest are means or conditional means. In Section 5, we present conditions under which social experiments solve the evaluation problem and assess the effectiveness of social experiments as a tool for evaluating employment and training programs. In Section 6, we outline two prototypical models of program participation and outcomes that represent the earliest and the latest thinking in the literature. We demonstrate the implications of these decision rules for the choice of an econometric evaluation estimator. We discuss the empirical evidence on the determinants of participation in government training programs.

The econometric models used to evaluate the impact of training programs in non-experimental settings are described in Section 7. The interplay between the economics of program participation and the choice of an appropriate evaluation estimator is stressed. In Section 8, we discuss some of the lessons learned from implementing various approaches to evaluation. Included in this section are the results of a simulation analysis based on the empirical model of Ashenfelter and Card (1985), where we demonstrate the sensitivity of the performance of alternative estimators to assumptions about heterogeneity in impact among persons and other data generating processes of the underlying econometric model. We also reexamine LaLonde's (1986) evidence on the performance of nonexperimental estimators and reinterpret the main lessons from his study.

Section 9 discusses the problems that arise in using microeconomic methods to evaluate programs with macroeconomic consequences. A striking example of the problems that can arise from this practice is provided. Two empirically operational general equilibrium frameworks are presented, and the lessons from applying them in practice are summarized. Section 10 surveys the findings from the non-experimental literature, and contrasts them with those from experimental evaluations. We conclude in Section 11 by surveying the main methodological lessons learned from the program evaluation literature on job training.

2 Public Job Training and Active Labor Market Policies

Many government policies affect employment and wages. The “active labor market” policies we analyze have two important features that distinguish them from general policies, such as income taxes, that also affect the labor market. First, they are targeted toward the unemployed or toward those with low skills or little work experience who have completed (usually at a low level) their formal schooling. Second, the policies are aimed at promoting employment and/or wage growth among this population, rather than just providing income support.

Table 2.1 describes the set of policies we consider. This set includes: (a) classroom training (CT) consisting of basic education to remedy deficiencies in general skills or vocational training to provide the skills necessary for particular jobs; (b) subsidized employment with public or private employers (WE), which includes public service employment (wholly subsidized temporary government jobs) and work experience (subsidized entry-level jobs at public or non-profit employers designed to introduce young people to the world of work) as well as wage supplements and fixed payments to private firms for hiring new workers; (c) subsidies to private firms for the provision of on-the-job training (OJT); (d) training in how to obtain a job; and (e) in-kind subsidies to job search such as referrals to employers and free access to job listings. Policies (d) and (e) fall under the general heading of job search assistance (JSA), which also includes the job matching services provided by the U.S. Employment Service and similar agencies in other countries.

As we argue in more detail below, distinguishing the types of training provided is important for two reasons. First, different types of training often imply different economic models of training participation and impact and therefore different econometric estimation strategies. Second, because most existing training programs provide a mix of these services, heterogeneity in the impact of training becomes an important practical concern. As we show in Section 7, this heterogeneity has important implications for the choice of econometric methods for evaluating active labor market policies.

We do not analyze privately supplied job training despite its greater quantitative importance to modern economies (see Heckman, Lochner and Taber, 1998a, or Mincer, 1962, 1993). For example, in the United States, Jacob Mincer has estimated that such training amounts to approximately 4 to 5 percent of GDP, annually. Despite the magnitude of this investment there are surprisingly few publicly-available studies of the returns to private job training, and many of those that are available do not control convincingly for the non-random allocation of training among private sector workers. Governments demand publicly-justified evaluations of training programs while private firms, to the extent that

they formally evaluate their training programs, keep their findings to themselves. An emphasis on objective publicly accessible evaluations is a distinctive feature of the modern welfare state, especially in an era of limited funds and public demands for accountability.

Table 2.2 presents the amount spent on active labor market policies by a number of OECD countries. Most OECD countries provide some mix of the employment and training services described in Table 2.1. Differences among countries include the relative emphasis on each type of service, the particular populations targeted for service, the total resources spent on the programs, how resources are allocated among programs and the extent to which employment and training services are integrated with other programs such as unemployment insurance or social assistance. In addition, although the programs we study are funded by governments, they are not always conducted by governments, especially in the U.S. and the U.K. In decentralized training systems, private firms and local organizations play an important role in providing employment and training services.

Table 2.2 reveals that many OECD countries spend substantial sums on active labor market policies. In nearly all countries, total expenditures are more than one-third of total expenditures on unemployment benefits, and some countries' expenditures on active labor market policies exceed those on unemployment benefits. Usually only a fraction of these expenditures are for CT. Further, even in countries that emphasize classroom training, governments spend substantial sums on other active labor market policies. Denmark spends 1 percent of its GDP on CT for adults, the most of any OECD country. However, this expenditure amounts to only 40 percent of its total spending on active labor market programs. Only in Canada is the fraction spent on CT larger. At the opposite extreme, Japan and the U.S. spend only 0.03 percent and 0.04 percent, respectively, of their GDP on CT. However, as the table shows, these two countries also spend the smallest share of GDP on active labor market policies.

The low percentage of GDP spent on active labor market programs in the U.S. has led some researchers to comment on the irony that despite these low expenditures, U.S. programs have been evaluated more extensively and over a longer period of time than programs elsewhere (Haveman and Saks, 1985; Björklund, 1993). Indeed, much of what is known about the impacts of these programs and many of the methodological developments associated with evaluating them come from U.S. evaluations.¹

¹However, the level of total expenditure in the U.S. is still quite large. Relative total expenditures on active labor market policies can be inferred from Table 2.2 using the relative sizes of each economy compared with the U.S. For example, the German economy is somewhat less than one-fourth the size of the U.S. economy, and the French, Italian and British economies are approximately one-sixth the size of the U.S. economy. Accordingly, training expenditures are somewhat greater in Germany and France, about the same in Italy, and less in the United Kingdom than in the U.S. See OECD, *Employment Outlook* (1996), Table 1.1, p.2.

We now consider in detail each type of employment and training service in Table 2.1. This discussion motivates the consideration of alternative economic models of program participation and impact in Sections 6 and 7, and our focus on heterogeneity in program impacts. It also provides a context for the empirical literature on the impact these programs that we review in Section 10.

The first category listed in Table 2.1 is classroom training. In many countries, CT represents the largest fraction of government expenditures on active labor market policy, and most of that expenditure is devoted to vocational training. Even in the U.S., where remedial programs aimed at high school dropouts and other low-skill individuals play a larger role than elsewhere, most CT programs provide vocational training. By design, most CT programs in the OECD are of limited duration. For example in Denmark, CT typically lasts 2 to 4 weeks (Jensen, et al., 1993) while in Sweden a duration of four months and in the United Kingdom, and the United States three months is the more typical duration. Per capita expenditures on such training varies substantially, with a training slot costing approximately \$7,500 in Sweden and between \$2,000 and \$3,000 in the United States.² The Swedish figures include stipends for participants while the U.S. figures do not.

An important difference among OECD countries that provide CT is the extent to which the training is relatively standardized and therefore less tailored to the requirements of firms or the market in general. In the 1980s and early 1990s, the Nordic countries usually provide CT in government training centers that use standardized materials and teaching methods. However, the emphasis has shifted recently, especially in Sweden, toward decentralized and firm based training. In the United Kingdom and the U.S., the provision of CT is highly decentralized and its content depends on the choices made by local councils of business, political, and labor leaders. The local councils receive funding from the federal government and then subcontract for CT with private vocational and proprietary schools and local community colleges. Due to this highly decentralized structure, both participant characteristics and training content can vary substantially among locales, which suggests that the impact of training is likely to vary substantially across individuals in evaluations of such programs.

The second category of services listed in Table 2.1 are wage and employment subsidies. This category encompasses several different specific services which we group together due to their analytic similarity. The simplest example of this type of policy provides subsidies to private firms for hiring workers in particular groups. These subsidies may take the form of a fixed amount for each new employee hired or some fraction of the employee's wage for a period of time. In the U.S., the Targeted Jobs Tax Credit is an example of this type of program. Heckman, Lochner, Smith and Taber (1997) discuss the empirical evidence on

²Unless otherwise indicated all monetary units are expressed in 1997 U.S. dollars.

the effectiveness of wage and employment subsidies in greater detail.

Temporary work experience (WE) usually targets low skilled youth or adults with poor employment histories and provides them with a job lasting 3 to 12 months in the public or nonprofit sector. The idea of these programs is to ease the transition of these groups into regular jobs, by helping them learn about the world of work and develop good work habits. Such programs constitute a very small proportion of U.S. training initiatives, but substantial fractions of services provided to youth in countries such as France (TUC) and the United Kingdom (Community Programmes). In public sector employment (PSE) programs, governments create temporary public sector jobs. These jobs usually require some amount of skill and are aimed at unemployed adults with recent work experience rather than youth or the disadvantaged. Except for a brief period during the late 1970s, they have not been used in the United States since the Depression era. However, they have been and remain an important component of active labor market policy in several European countries.

The third category in Table 2.1 is subsidized on-the-job training at private firms. The goal of subsidized OJT programs is to induce employers provide job-relevant skills, including firm-specific skills, to disadvantaged workers. In the U.S., employers receive a 50 percent wage subsidy for up to six months; in the U.K. employers receive a lump sum per week (O'Higgins, 1994). Although evidence is limited and firm training is difficult to measure, there is a widespread view that these programs in fact provide little training, even informal on-the-job training, and are better characterized as a work experience or wage subsidy program (e.g., Breen, 1988; Hutchinson and Church, 1989).³ Survey responses by employers who have hired or sponsored OJT trainees suggest that they value the program for its help in reducing the costs associated with hiring and retaining suitable employees more than for the opportunity to increase the skills of new workers (Begg, et al., 1991).

For purposes of evaluation, it is almost always impossible to distinguish those OJT experiences from which new skills were acquired from those that amounted to work experience or wage subsidy without a training component. In addition, because OJT is provided by individual employers, this indeterminacy is not simply a program-specific feature, but holds among individuals within the same program. Consequently, OJT programs will likely have heterogeneous effects, and the impact, if any, of these programs will result from some combination of learning by doing, the usual training provided by the firm to new workers

³The provision of subsidized OJT is particularly hard to monitor both because on-the-job training has proven difficult to measure with survey methods (Barron, Berger and Black, 1997) and because trainees often do not perceive that they have been treated any differently than their co-workers who are not subsidized. In fact, both groups may have received substantial amounts of informal on-the-job training. For evidence of the importance of informal on-the-job training in the U.S., see Barron, Black and Lowenstein (1989).

and incremental training beyond that provided to unsubsidized workers.

The fourth category of services in Table 2.1 is job search assistance. The purpose of these services is to facilitate the matching process between workers and firms both by reducing time unemployed and by increasing match quality. The programs are usually operated by the national or local employment service, but sometimes may be subcontracted out to third parties. Included under this category are direct placement in vacant jobs, employer referrals, in-kind subsidies to search such as free access to job listings and telephones for contacting employers, career counseling, and instruction in job search skills. The last of these, which often includes instruction in general social skills, was developed in the U.S., but is now used in U.K., Sweden, and recently France (Björklund and Regner, 1996, p. 24). In recent years, JSA has become more popular due to its low cost, usually just a few hundred dollars per participant, and relatively solid record of performance (which we discuss in detail in Section 10).

To conclude this section, we discuss five features of employment and training programs that should be kept in mind when evaluating them. First, as the operation of these programs has become more decentralized in OECD countries, there have emerged differences between how these programs were designed and how they are implemented (Hollister and Freedman, 1988). Actual practice can deviate substantially from explicit written policy.⁴ Therefore, the evaluator must be careful to characterize the program as implemented when assessing its impacts.

Second, participants often receive services from more than one category in Table 2.1. For example, classroom training in vocational skills might be followed by job search assistance. In the U.K., the Youth Training Scheme (now Youth Training) was explicitly designed to combine OJT with 13 weeks of CT. Some expensive programs combine several of the services listed in Table 2.1 into a single package. For example, in the U.S. the Job Corps program for youth combines classroom training with work experience and job search assistance in a residential setting at a current cost of around \$19,000 per participant. Many available survey data sets do not identify all the services received by a participant. In this case, the practice of combining together various types of training, particularly when combinations are tailored to the needs of individual trainees as in the U.S. JTPA program, constitutes another source of heterogeneity in the impact of training. Even when administrative data are available that identify the services received, isolating the impact of particular individual services often proves difficult or impossible in practice due to the small samples receiving particular combinations of services or due to difficulties in determining the process by which

⁴For example, see Breen (1988) and Hollister and Freedman (1990) describing the implementation of WEP in Ireland and Hollister and Freedman (1990) and Leigh (1995) describing the implementation of JTPA in the United States.

individuals come to receive particular service combinations.

Third, certain features of active labor market programs affect individuals' decisions to participate in training. In some countries, such as Sweden and the United Kingdom, participation in training is a condition for receiving unemployment benefits rather than less generous social assistance payments. In the U.S., participation is sometimes required by a court order in lieu of alternative punishment.

Fourth, program administrators often have considerable discretion over whom they admit into government training programs. This discretion results from the fact that the number of applicants often exceeds the number of available training positions. It has long been a feature of U.S. programs, but also has characterized programs in Austria, Denmark, Germany, Norway, and the United Kingdom (Björklund and Regner, 1996; Westergaard-Neilsen, 1993; Kraus, et al., 1997). Consequently, when modeling participation in training, it may be important to account for not only individual incentives, but also those of the program operators. In Section 6, we discuss the incentives facing program operators and how they affect the characteristics of participants in government training programs.

Finally, the different types of services require different economic models of program participation and impact. For example, the standard human capital model captures the essence of individual decisions to invest in vocational skills (CT). It provides little guidance to behavior regarding job search assistance or wage subsidies. In Section 6 we describe economic models that describe participation in alternative programs and discuss their implications for evaluation research.

3 The Evaluation Problem and the Parameters of Interest in Evaluating Social Programs

3.1 The Evaluation Problem

Constructing counterfactuals is the central problem in the literature on evaluating social programs. In the simplest form of the evaluation problem, persons are imagined as being able to occupy one of two mutually exclusive states: “0” for the untreated state and “1” for the treated state. Treatment is associated with participation in the program being evaluated.⁵ Associated with each state is an outcome, or set of outcomes. It is easiest to think of each state as consisting of only a single outcome measure, such as earnings, but just as easily, we can use the framework to model vectors of outcomes such as earnings, employment and participation in welfare programs. In the models presented in section 6, we study an entire vector of earnings or employment at each age that result from program participation.

We can express these outcomes as a function of conditioning variables, X . Denote the potential outcomes by Y_0 and Y_1 , corresponding to the untreated and treated states. Each person has a (Y_0, Y_1) pair. Assuming that means exist, we may write the (vector) of outcomes in each state as

$$(3.1a) \quad Y_0 = \mu_0(X) + U_0$$

$$(3.1b) \quad Y_1 = \mu_1(X) + U_1$$

where $E(Y_0|X) = \mu_0(X)$ and $E(Y_1|X) = \mu_1(X)$. To simplify the notation, we keep the conditioning on X implicit unless it serves to clarify the exposition by making it explicit. The potential outcome actually realized depends on decisions made by individuals, firms, families or government bureaucrats. This model of potential outcomes is variously attributed to Fisher (1935), Neyman (1935), Roy (1951), Quandt (1972, 1988) or Rubin (1974).

To focus on main ideas, throughout most of this chapter we assume $E(U_1|X) = E(U_0|X) = 0$, although as we note at several places in this paper, this is not strictly required. For many of the estimators that we consider in this chapter we allow for the more general case

$$Y_0 = g_0(X) + U_0$$

$$Y_1 = g_1(X) + U_1$$

where $E(U_0 | X) \neq 0$ and $E(U_1 | X) \neq 0$. Then $\mu_0(X) = g_0(X) + E(U_0|X)$ and $\mu_1(X) = g_1(X) + E(U_1|X)$.⁶ Thus X is not necessarily exogenous in the ordinary econometric usage

⁵In this paper, we only consider a two potential state model in order to focus on the main ideas. Heckman (1998a) develops a multiple state model of potential outcomes for a large number of mutually exclusive states. The basic ideas in his work are captured in the two outcome models we present here.

⁶For example, an exogeneity assumption is not required when using social experiments to identify $E(Y_1 - Y_0|X, D = 1)$.

of that term. These conditions do *not* imply that $E(U_1 - U_0|X, D = 1) = 0$. D may depend on U_1 , U_0 or $U_1 - U_0$ and X .

Note also that Y may be a vector of outcomes or a time series of potential outcomes: (Y_{0t}, Y_{1t}) , for $t = 1, \dots, T$, on the same type of variable. We will encounter the latter case when we analyze panel data on outcomes. In this case, there is usually a companion set of X variables which we will sometimes assume to be strictly exogenous in the conventional econometric meaning of that term: $E(U_{0t}|X) = 0, E(U_{1t}|X) = 0$ where $X = (X_1, \dots, X_T)$. In defining a sequence of “treatment on the treated” parameters, $E(Y_{1t} - Y_{0t}|X, D = 1)$ $t = 1, \dots, T$, this assumption allows us to abstract from any dependence between U_{1t} , U_{0t} and X . It excludes differences in U_{1t} and U_{0t} arising from X dependence and allows us to focus on differences in outcomes solely attributable to D . While convenient, this assumption is overly strong.

However, we stress that the exogeneity assumption in either cross section or panel contexts is only a matter of convenience and is not strictly required. What is required for an interpretable definition of the “treatment on the treated” parameter is avoiding conditioning on X variables *caused* by D even holding $Y^P = ((Y_{01}, Y_{11}), \dots, (Y_{0T}, Y_{1T}))$ fixed where Y^P is the vector of potential outcomes. More precisely, we require that for the conditional density of the data

$$f(X|D, Y^P) = f(X|Y^P)$$

i.e. we require that the realization of D does not determine X given the vector of potential outcomes. Otherwise, the parameter $E(Y_1 - Y_0|X, D = 1)$ does not capture the full effect of treatment on the treated as it operates through all channels and certain other technical problems discussed in Heckman (1998a) arise. In order to obtain $E(Y_{1t} - Y_{0t}|X, D = 1)$ defined on subsets of X , say X_c , simply integrate out $E(Y_{1t} - Y_{0t}|X, D)$ against the density $f(\tilde{X}_c|D = 1)$ where \tilde{X}_c is the portion of X not in X_c : $X = (X_c, \tilde{X}_c)$.

Note, finally, that the choice of a base state “0” is arbitrary. Clearly the roles of “0” and “1” can be reversed. In the case of human capital investments, there is a natural base state. But for many other evaluation problems the choice of a base is arbitrary. Assumptions appropriate for one choice of “0” and “1” need not carry over to the opposite choice. With this cautionary note in mind, we proceed as if a well-defined base state exists.

In many problems it is convenient to think of “0” as a benchmark “no treatment” state. The gain to the individual of moving from “0” to “1” is given by

$$(3.2) \quad \Delta = Y_1 - Y_0.$$

If one could observe both Y_0 and Y_1 for the same person at the same time, the gain Δ would be known for each person. The fundamental evaluation problem arises because we do not know both coordinates of (Y_1, Y_0) and hence Δ for anybody. All approaches to

solving this problem attempt to estimate the missing data. These attempts to solve the evaluation problem differ in the assumptions they make about how the missing data are related to the available data, and what data are available. Most approaches to evaluation in the social sciences accept the impossibility of constructing Δ for anyone. Instead, the evaluation problem is redefined from the individual level to the population level to estimate the mean of Δ , or some other aspect of the distribution of Δ , for various populations of interest. The question becomes what features of the distribution of Δ should be of interest and for what populations should it be defined?

3.2 The Counterfactuals of Interest

There are many possible counterfactuals of interest for evaluating a social program. One might like to compare the state of the world in the presence of the program to the state of the world if the program were operated in a different way, or to the state of the world if the program did not exist at all, or to the state of the world if alternative programs were used to replace the present program. A full evaluation entails an enumeration of all outcomes of interest for all persons both in the current state of the world and in all the alternative states of interest, and a mechanism for valuing the outcomes in the different states.

Outcomes of interest in program evaluations include the direct benefits received, the level of behavioral variables for participants and nonparticipants and the payments for the program, for both participants and nonparticipants, including taxes levied to finance a publicly provided program. These measures would be displayed for each individual in the economy to characterize each state of the world.

In a Robinson Crusoe economy, participation in a program is a well-defined event. In a modern economy, almost everyone participates in each social program either directly or indirectly. A training program affects more than the trainees. It also affects the persons with whom the trainees compete in the labor market, the firms that hire them and the taxpayers who finance the program. The impact of the program depends on the number and composition of the trainees. Participation in a program does not mean the same thing for all people.

The traditional evaluation literature usually defines the effect of participation to be the effect of the program on participants explicitly enrolled in the program. These are the “*Direct Effects*.” They exclude the effects of a program that do not flow from direct participation, known as the “*Indirect Effects*”. This distinction appears in the pioneering work of H. G. Lewis on measuring union relative wage effects (Lewis, 1963). His insights apply more generally to all evaluation problems in social settings.

There may be indirect effects for both direct participants and direct nonparticipants. Thus a direct participant may pay taxes to support the program just as persons who do not

directly participate may also pay taxes. A firm may be an indirect beneficiary of the lower wages resulting from an expansion of the trained workforce. The conventional econometric and statistical literature ignores the indirect effects of programs and equates “treatment” outcomes with the direct outcome Y_1 in the program state and “no treatment” with the direct outcome Y_0 in the no program state.

Determining all outcomes in all states is not enough to evaluate a program. Another aspect of the evaluation problem is the valuation of the outcomes. In a democratic society, aggregation of the evaluations and the outcomes in a form useful for social deliberations also is required. Different persons may value the same state of the world differently even if they experience the same “objective” outcomes and pay the same taxes. Preferences may be interdependent. Redistributive programs exist, in part, because of altruistic or paternalistic preferences. Persons may value the outcomes of other persons either positively or negatively. Only if one person’s preferences are dominant (the idealized case of a social planner with a social welfare function) is there a unique evaluation of the outcomes associated for each possible state from each possible program.

The traditional program evaluation literature assumes that the valuation of the direct effects of the program boil down to the effect of the program on GDP. This assumption ignores the important point that different persons value the same outcomes differently and that the democratic political process often entails coalitions of persons who value outcomes in different ways. Both efficiency and equity considerations may receive different weights from different groups. Different mechanisms for aggregating evaluations and resolving social conflicts exist in different societies. Different types of information are required to evaluate a program under different modes of social decision making.

Both for pragmatic and political reasons, government social planners, statisticians or policy makers may value objective output measures differently than the persons or institutions being evaluated. The classic example is the value of nonmarket time (Greenberg, 1997). Traditional program evaluations exclude such valuations largely because of the difficulty of inputting the value and quantity of nonmarket time. By doing this, however, these evaluations value labor supply in the market sector at the market wage, but value labor supply in the nonmarket sector at a zero wage. By contrast, individuals value labor supply in the nonmarket sector at their reservation wage. In this example, two different sets of preferences value the same outcomes differently. In evaluating a social program in a society that places weight on individual preferences, it is appropriate to recognize personal evaluations and that the same outcome may be valued in different ways by different social actors.

Programs that embody redistributive objectives inherently involve different groups. Even if the taxpayers and the recipients of the benefits of a program have the same preferences, their valuations of a program will, in general, differ. Altruistic considerations often

motivate such programs. These often entail private valuations of *distributions* of program *impacts* - how much recipients gain over what they would experience in the absence of the program. (See Heckman and Smith, 1993, 1995, 1998a and Heckman, Smith and Clements, 1997.)

Answers to many important evaluation questions require knowledge of the distribution of program gains especially for programs that have a redistributive objective or programs for which altruistic motivations play a role in motivating the existence of the program. Let $D = 1$ denote direct participation in the program and $D = 0$ denote direct nonparticipation. To simplify the argument in this section, ignore any indirect effects. From the standpoint of a detached observer of a social program who takes the base state values (denoted “0”) as those that would prevail in the absence of the program, it is of interest to know, among other things,

- (A) the proportion of people taking the program who benefit from it:
 $\Pr(Y_1 > Y_0 \mid D = 1) = \Pr(\Delta > 0 \mid D = 1)$;
- (B) the proportion of the total population benefiting from the program:
 $\Pr(Y_1 > Y_0 \mid D = 1) \cdot \Pr(D = 1) = \Pr(\Delta > 0 \mid D = 1) \cdot \Pr(D = 1)$;
- (C) selected quantiles of the impact distribution
 $\inf_{\Delta} \{ \Delta : F(\Delta \mid D = 1) > q \}$, where q is a quantile of the distribution
and where “inf” is the smallest attainable value of Δ that satisfies the
condition stated in the braces;
- (D) the distribution of gains at selected base state values:
 $F(\Delta \mid D = 1, Y_0 = y_0)$;
- (E) the increase in the level of outcomes above a certain threshold \bar{y} due to a policy:
 $\Pr(Y_1 > \bar{y} \mid D = 1) - \Pr(Y_0 > \bar{y} \mid D = 1)$.

Measure (A) is of interest in determining how widely program *gains* are distributed among participants. Participants in the political process with preferences over distributions of program outcomes would be unlikely to assign the same weight to two programs with the same mean outcome, one of which produced favorable outcomes for only a few persons while the other distributed gains more broadly. When considering a program, it is of interest to determine the proportion of participants who are harmed as a result of program participation, indicated by $\Pr(Y_1 < Y_0 \mid D = 1)$. Negative mean impact results might be acceptable if most participants gain from the program. These features of the outcome distribution are likely to be of interest to evaluators even if the persons studied do not know their Y_0 and Y_1 values in advance of participating in the program.

Measure (B) is the proportion of the entire population that benefits from the program, assuming that the costs of financing the program are broadly distributed and are not perceived to be related to the specific program being evaluated. If voters have correct

expectations about the joint distribution of outcomes, it is of interest to politicians to determine how widely program benefits are distributed. At the same time, large program gains received by a few persons may make it easier to organize interest groups in support of a program than if the same gains are distributed more widely.

Evaluators interested in the distribution of program benefits would be interested in measure (C). Evaluators who take a special interest in the impact of a program on recipients in the lower tail of the base state distribution would find measure (D) of interest. It reveals how the distribution of gains depends on the base state for participants. Measure (E) provides the answers to the question “do the distributions of gains for the participants dominate the distribution of outcomes if they did not participate?” (See Heckman, Smith and Clements, 1997; and Heckman and Smith, 1998a.) Expanding the scope of the discussion to evaluate the indirect effects of the program makes it more likely that estimating distributional impacts is an important part in conducting program evaluations.

3.3 The Counterfactuals Most Commonly Estimated In The Literature

The evaluation problem in its most general form for distributions of outcomes is formidable and is not considered in depth either in this chapter or in the literature. (Heckman and Smith, 1998a, and Heckman, Smith and Clements, 1997, consider identification and estimation of counterfactual distributions.) Instead, in this chapter we focus on counterfactual means, and consider a form of the problem in which analysts have access to information on persons who are in one state or the other at any time, and for certain time periods there are some persons in both states, but there is no information on any single person who is in both states at the same time. As discussed in Heckman (1998a) and Heckman and Smith (1998a), a crucial assumption in the traditional evaluation literature is that the no treatment state approximates the no program state. This would be true if indirect effects are negligible.

Most of the empirical work in the literature on evaluating government training programs focuses on means and in particular on one mean counterfactual: the mean direct effect of treatment on those who take treatment. The transition from the individual to the group level counterfactual recognizes the inherent impossibility of observing the same person in both states at the same time. By dealing with aggregates, rather than individuals, it is sometimes possible to estimate group impact measures even though it may be impossible to measure the impacts of a program on any particular individual. To see this point more formally, consider the switching regression model with two regimes denoted by “1” and “0” (Quandt, 1972). The observed outcome Y is given by

$$(3.3) \quad Y = DY_1 + (1 - D)Y_0.$$

When $D = 1$ we observe Y_1 ; when $D = 0$ we observe Y_0 .

To cast the foregoing model in a more familiar-looking form, and to distinguish it from conventional regression models, express the means in (3.1a) and (3.1b) in more familiar linear regression form:

$$E(Y_j|X) = \mu_j(X) = X\beta_j, j = 0, 1.$$

With these expressions, substitute from (3.1a) and (3.1b) into (3.3) to obtain

$$Y = D(\mu_1(X) + U_1) + (1 - D)(\mu_0(X) + U_0).$$

Rewriting,

$$Y = \mu_0(X) + D(\mu_1(X) - \mu_0(X) + U_1 - U_0) + U_0.$$

Using the linear regression representation, we obtain

$$(3.4) \quad Y = X\beta_0 + D(X(\beta_1 - \beta_0) + U_1 - U_0) + U_0.$$

Observe that from the definition of a conditional mean, $E(U_0 | X) = 0$ and $E(U_1 | X) = 0$.

The parameter most commonly invoked in the program evaluation literature, although not the one actually estimated in social experiments, or in most nonexperimental evaluations, is the effect of randomly picking a person with characteristics X and moving that person from “0” to “1”:

$$E(Y_1 - Y_0|X) = E(\Delta|X).$$

In terms of the switching regression model this parameter is the coefficient on D in the “regression” non-error component of following equation:

$$(3.5) \quad \begin{aligned} Y &= \mu_0(X) + D(\mu_1(X) - \mu_0(X)) + \{U_0 + D(U_1 - U_0)\} \\ &= \mu_0(X) + D(E(\Delta|X)) + \{U_0 + D(U_1 - U_0)\} \\ &= X\beta_0 + DX(\beta_1 - \beta_0) + \{U_0 + D(U_1 - U_0)\} \end{aligned}$$

where the term in braces is the “error.”

If the model is specialized so that there are K regressors plus an intercept and $\beta_1 = (\beta_{10}, \dots, \beta_{1K})$ and $\beta_0 = (\beta_{00}, \dots, \beta_{0K})$, where the intercepts occupy the first position, and the slope coefficients are the same in both regimes:

$$\beta_{1j} = \beta_{0j} = \beta_j, \quad j = 1, \dots, K$$

and $\beta_{00} = \beta_0$ and $\beta_{10} - \beta_{00} = \alpha$, the parameter under consideration reduces to α :

$$(3.6) \quad E(Y_1 - Y_0|X) = \beta_{10} - \beta_{00} = \alpha.$$

The regression model for this special case maybe written as

$$(3.7) \quad Y = X\beta + D\alpha + \{U_0 + D(U_1 - U_0)\}.$$

It is nonstandard from the standpoint of elementary econometrics because the error term has a component that switches on or off with D . In general, its mean is not zero because $E[U_0 + D(U_1 - U_0)] = E(U_1 - U_0|D = 1) \Pr(D = 1)$. If $U_1 - U_0$, or variables statistically dependent on it, help determine D , $E(U_1 - U_0 | D = 1) \neq 0$. Intuitively, if persons who have high gains ($U_1 - U_0$) are more likely to appear in the program, than this term is positive.

In practice most non-experimental and experimental studies do not estimate $E(\Delta | X)$. Instead, most nonexperimental studies estimate the effect of treatment on the treated, $E(\Delta | X, D = 1)$. This parameter conditions on participation in the program as follows:

$$(3.8) \quad E(\Delta|X, D = 1) = E(Y_1 - Y_0|X, D = 1) = X(\beta_1 - \beta_0) + E(U_1 - U_0|X, D = 1).$$

It is the coefficient on D in the non-error component of the following regression equation:

$$(3.9) \quad Y = \mu_0(X) + D(E(\Delta|X, D = 1)) \\ + \{U_0 + D[(U_1 - U_0) - E(U_1 - U_0|X, D = 1)]\} \\ = X\beta_0 + D(X(\beta_1 - \beta_0) + E(U_1 - U_0|X, D = 1)) \\ + \{U_0 + D[(U_1 - U_0) - E(U_1 - U_0|X, D = 1)]\}.$$

$E(\Delta | X, D = 1)$ is a nonstandard parameter in conventional econometrics. It combines “structural” parameters $X(\beta_1 - \beta_0)$ with the means of the unobservables ($E(U_1 - U_0|X, D = 1)$). It measures the average gain in the outcome for persons who choose to participate in a program compared to what they would have experienced in the base state. It computes the average gain in terms of both observables and unobservables. It is the latter that makes the parameter look nonstandard. Most econometric activity is devoted to separating β_0 and β_1 from the effects of the regressors on U_1 and U_0 . Parameter (3.8) combines these effects.

This parameter is implicitly defined conditional on the current levels of participation in the program in society at large. Thus it recognizes social interaction. But at any point in time the aggregate participation level is just a single number, and the composition of trainees is fixed. From a single cross section of data, it is not possible to estimate how variation in the levels and composition of participants in a program affect the parameter.

The two evaluation parameters we have just presented are the same if we assume that $U_1 - U_0 = 0$, so the unobservables are common across the two states. From (3.9) we now have $Y_1 - Y_0 = \mu_1(X) - \mu_0(X) = X(\beta_1 - \beta_0)$. The difference between potential outcomes in the two states is a function of X but not of unobservables. Further specializing the model to one of intercept differences (*i.e.* $Y_1 - Y_0 = \alpha$), requires that the difference between potential

outcomes is a constant. The associated regression can be written as the familiar-looking dummy variable regression model:

$$(3.10) \quad Y = X\beta + D\alpha + U, \text{ where } E(U) = 0.$$

The parameter α is easy to interpret as a standard structural parameter and the specification (3.10) looks conventional. In fact, model (3.10) dominates the conventional evaluation literature. The validity of many conventional instrumental variables methods and longitudinal estimation strategies is contingent on this specification as we document below. The conventional econometric evaluation literature focuses on α , or more rarely, $X(\beta_1 - \beta_0)$, and the selection problem arises from the correlation between D and U .

While familiar, the framework of (3.10) is very special. Potential outcomes (Y_1, Y_0) differ only by a constant ($Y_1 - Y_0 = \alpha$). The best Y_1 is the best Y_0 . All people gain or lose the same amount in going from “0” to “1”. There is no heterogeneity in gains. Even in the more general case, with $\mu_1(X)$ and $\mu_0(X)$ distinct, or $\beta_1 \neq \beta_0$ in the linear regression representation, so long as $U_1 = U_0$ among people with the same X , there is no heterogeneity in the outcomes moving from “0” to “1”. This assumed absence of heterogeneity in response to treatments is strong. When tested, it is almost always rejected (see Heckman, Smith and Clements, 1997 and the evidence presented below).

There is one case when $U_1 \neq U_0$, where the two parameters of interests are still equal even though there is dispersion in gain Δ . This case occurs when

$$(3.11) \quad E(U_1 - U_0 | X, D = 1) = 0.$$

Condition (3.11) arises when conditional on X , D does not explain or predict $U_1 - U_0$. This condition could arise if agents who select into state “1” from “0” either do not know or do not act on $U_1 - U_0$, or information dependent on $U_1 - U_0$, in making their decision to participate in the program. Ex post, there is heterogeneity, but ex ante it is not acted on in determining participation in the program.

When the gain does not affect individuals’ decisions to participate in the program, the error terms (the terms in braces in (3.7) and (3.9)) have conventional properties. The only bias in estimating the coefficients on D in the regression models arise from the dependence between U_0 and D just as the only source of bias in the common coefficient model is the covariance between U and D when $E(U|X) = 0$. To see this point take the expectation of the terms in braces in (3.7) and (3.9), respectively, to obtain the following:

$$E(U_0 + D(U_1 - U_0) | X, D) = E(U_0 | X, D)$$

and

$$E(U_0 + D [(U_1 - U_0) - E(U_1 - U_0 | X, D = 1)] | X, D) = E(U_0 | X, D).$$

A problem that remains when condition (3.11) holds is that, the D component in the error terms contributes a component of variance to the model and so makes the model heteroscedastic:

$$\begin{aligned} \text{Var}(U_0 + D(U_1 - U_0)|X, D) &= \text{Var}(U_0|X, D) \\ &+ 2\text{COV}(U_0, U_1 - U_0|X, D)D + \text{Var}(U_1 - U_0|X, D)D. \end{aligned}$$

The distinction between a model with $U_1 = U_0$, and one with $U_1 \neq U_0$, is fundamental to understanding modern developments in the program evaluation literature. When $U_1 = U_0$ and we condition on X , *everyone* with the same X has the same treatment effect. The evaluation problem greatly simplifies and one parameter answers all of the conceptually distinct evaluation questions we have posed. “Treatment on the treated” is the same as the effect of taking a person at random and putting him/her into the program. The distributional questions (A)–(E) all have simple answers because everyone with the same X has the same Δ . Equation (3.10) is amenable to analysis by conventional econometric methods. Eliminating the covariance between D and U is the central problem in this model.

When $U_1 \neq U_0$, but (3.11) characterizes the program being evaluated, most of the familiar econometric intuition remains valid. This is the “random coefficient” model with the coefficient on D “random” (from the standpoint of the observing economist), but uncorrelated with D . The central problem in this model is covariance between U_0 and D and the only additional econometric problem arises in accounting for heteroscedasticity in getting the right standard errors for the coefficients. In this case, the response to treatment varies among persons with the same X values. The mean effect of treatment on the treated and the effect of treatment on a randomly chosen person are the same.

In the general case when $U_1 \neq U_0$ and (3.11) no longer holds, we enter a new world not covered in the traditional econometric evaluation literature. A variety of different treatment effects can be defined. Conventional econometric procedures often break down or require substantial modification. The error term for the model (3.5) has a non-zero mean.⁷ Both error terms are heteroscedastic. The distinctions among these three models — (a) the coefficient on D is fixed (given X) for everyone; (b) the coefficient on D is variable (given X), but does not help determine program participation; and (c) the coefficient on D is variable (given X) and does help determine program participation — are fundamental to this chapter and the entire literature on program evaluation.

⁷ $E[U_0 + D(U_1 - U_0)|X] = E(U_1 - U_0 | X, D = 1) \Pr(D = 1 | X) \neq 0.$

3.4 Is Treatment on the Treated an Interesting Economic Parameter?

What economic question does parameter (3.2) answer? How does it relate to the conventional parameter of interest in cost-benefit analysis - the effect of a program on GDP? In order to relate the parameter (3.2) with the parameters needed to perform traditional cost-benefit analysis, it is fruitful to consider a more general framework. Following our previous discussion, we consider two discrete states or sectors corresponding to direct participation and nonparticipation and a vector of policy variables φ that affect the outcomes in both states and the allocation of persons to states or sectors. The policy variables may be discrete or continuous. Our framework departs from the conventional treatment effect literature and allows for general equilibrium effects.

Assuming that costless lump-sum transfers are possible, that a single social welfare function governs the distribution of resources and that prices reflect true opportunity costs, traditional cost-benefit analysis (see, e.g., Harberger, 1971) seeks to determine the impact of programs on the total output of society. Efficiency becomes the paramount criterion in this framework, with the distributional aspects of policies assumed to be taken care of by lump sum transfers and taxes engineered by an enlightened social planner. In this framework, impacts on total output are the only objects of interest in evaluating programs. The distribution of program impacts is assumed to be irrelevant. This framework is favorable to the use of mean outcomes to evaluate social programs.

Within the context of the simple framework discussed in Section 3.1, let Y_1 and Y_0 be individual output which trades at a constant relative price of “1” set externally and not affected by the decisions of the agents we analyze. Alternatively, assume that the policies we consider do not alter relative prices. Let φ be a vector of policy variables which operate on all persons. These generate indirect effects. $c(\varphi)$ is the social cost of φ denominated in “0” units. We assume that $c(0) = 0$ and that c is convex and increasing in φ . Let $N_1(\varphi)$ be the number of persons in state “1” and $N_0(\varphi)$ be the number of persons in state “0”. The total output of society is

$$N_1(\varphi)E(Y_1 | D = 1, \varphi) + N_0(\varphi)E(Y_0 | D = 0, \varphi) - c(\varphi),$$

where $N_1(\varphi) + N_0(\varphi) = \bar{N}$ is the total number of persons in society. For simplicity, we assume that all persons have the same person-specific characteristics X . Vector φ is general enough to include financial incentive variables for participation in the program as well as mandates that assign persons to a particular state. A policy may benefit some and harm others.

Assume for convenience that the treatment choice and mean outcome functions are differentiable and for the sake of argument further assume that φ is a scalar. Then the change in output in response to a marginal increase in φ from any given position is:

$$(3.12) \quad \Delta(\varphi) = \frac{\partial N_1(\varphi)}{\partial \varphi} [E(Y_1 | D = 1, \varphi) - E(Y_0 | D = 0, \varphi)] + N_1(\varphi) \left[\frac{\partial E(Y_1 | D = 1, \varphi)}{\partial \varphi} \right] + N_0(\varphi) \left[\frac{\partial E(Y_1 | D = 0, \varphi)}{\partial \varphi} \right] - \frac{\partial c(\varphi)}{\partial \varphi}.$$

The first term arises from the transfer of persons across sectors that is induced by the policy change. The second term arises from changes in output within each sector induced by the policy change. The third term is the marginal social cost of the change.

In principle, this measure could be estimated from time-series data on the change in aggregate GDP occurring after the program parameter φ is varied. Assuming a well-defined social welfare function and making the additional assumption that prices are constant at initial values, an increase in GDP evaluated at base period prices raises social welfare provided that feasible bundles can be constructed from the output after the social program parameter is varied so that all losers can be compensated. (See, e.g., Laffont, 1989, p. 155, or the comprehensive discussion in Chipman and Moore, 1976).

If marginal policy changes have no effect on intra-sector mean output, the bracketed elements in the second set of terms inside the braces are zero. In this case, the parameters of interest for evaluating the impact of the policy change on GDP are:

- (i) $\frac{\partial N_1(\varphi)}{\partial \varphi}$; the number of people entering or leaving state 1.
- (ii) $E(Y_1 | D = 1, \varphi) - E(Y_0 | D = 0, \varphi)$; the mean output difference between sectors.
- (iii) $\frac{\partial c(\varphi)}{\partial \varphi}$; the social marginal cost of the policy.

It is revealing that nowhere on this list are the parameters that receive the most attention in the econometric policy evaluation literature. (See, e.g., Heckman and Robb, 1985a). These are “the effect of treatment on the treated”:

- (a) $E(Y_1 - Y_0 | D = 1, \varphi)$
- or
- (b) $E(Y_1 | \varphi = \bar{\varphi}) - E(Y_0 | \varphi = 0)$ where $\varphi = \bar{\varphi}$ sets $N_1(\bar{\varphi}) = \bar{N}$. This is the effect of universal coverage for the program.

Parameter (ii) can be estimated by taking simple mean differences between the outputs in the two sectors; no adjustment for selection bias is required. Parameter (i) can be obtained from knowledge of the net movement of persons across sectors in response to the policy change, something usually neglected in micro policy evaluation (for exceptions, see Moffitt, 1992, or Heckman, 1992). Parameter (iii) can be obtained from cost data. Full social marginal costs should be included in the computation of this term. The typical micro evaluation neglects all three terms. Costs are rarely collected and gross outcomes are typically reported; entry effects are neglected and term (ii) is usually “adjusted” to avoid selection bias when in fact, no adjustment is needed to estimate the impact of the program on GDP.

It is informative to place additional structure on this model. This leads to a representation of a criterion that is widely used in the literature on microeconomic program evaluation and also establishes a link with the models of program participation used in the later sections of this chapter. Assume a binary choice random utility framework. Suppose that agents make choices based on net utility and that policies affect participant utility through an additively-separable term $k(\varphi)$ that is assumed scalar and differentiable. Net utility is

$$U = X + k(\varphi)$$

where k is monotonic in φ and where the joint distributions of (Y_1, X) and (Y_0, X) are $F(y_1, x)$ and $F(y_0, x)$, respectively. The underlying variables are assumed to be continuously distributed. In the special case of the Roy model of self-selection (see, Heckman and Honoré, 1990, for one discussion) $X = Y_1 - Y_0$,

$$D = 1(U \geq 0) = 1(X \geq -k(\varphi)),$$

where “1” is the indicator function ($1(Z > 0) = 1$ if $Z > 0$; = 0 otherwise)

$$N_1(\varphi) = \bar{N} \Pr(U \geq 0) = \bar{N} \int_{-k(\varphi)}^{\infty} f(x)dx,$$

and

$$N_0(\varphi) = \bar{N} \Pr(U < 0) = \bar{N} \int_{-\infty}^{-k(\varphi)} f(x)dx.$$

Total output is

$$\bar{N} \int_{-\infty}^{\infty} y_1 \int_{-k(\varphi)}^{\infty} f(y_1, x | \varphi) dx dy_1 + \bar{N} \int_{-\infty}^{\infty} y_0 \int_{-\infty}^{-k(\varphi)} f(y_0, x | \varphi) dx dy_0 - c(\varphi).$$

Under standard conditions (see, e.g., Royden, 1968), we may differentiate this expression to obtain the following expression for the marginal change in output with respect to a change in φ :

$$(3.13) \quad \Delta(\varphi) = \bar{N}k'(\varphi)f_x(-k(\varphi))[E(Y_1 | D = 1, x = -k(\varphi), \varphi) - E(Y_0 | D = 0, x = -k(\varphi), \varphi)] \\ + \bar{N}[\int_{-\infty}^{\infty} y_1 \int_{-k(\varphi)}^{\infty} \frac{\partial f(y_1, x | \varphi)}{\partial \varphi} dx dy_1 + \int_{-\infty}^{\infty} y_0 \int_{-\infty}^{-k(\varphi)} \frac{\partial f(y_0, x | \varphi)}{\partial \varphi} dx dy_0] \\ - \frac{\partial c(\varphi)}{\partial \varphi}.$$

This model has a well-defined margin: $X = -k(\varphi)$, which is the utility of the marginal entrant into the program. The utility of the participant might be distinguished from the objective of the social planner who seeks to maximize total output. The first set of terms corresponds to the gain arising from the movement of persons at the margin (the term in brackets) weighted by the proportion of the population at the margin, $k'(\varphi)f_x(-k(\varphi))$, times the number of people in the population. This term is the net gain from switching sectors. The expression in brackets in the first term is a limit form of the “local average treatment effect” of Imbens and Angrist (1994) which we discuss further in our discussion of instrumental variables in Section 7.4.3. The second set of terms is the intrasector change in output resulting from a policy change. This includes both direct and indirect effects. The second set of terms is ignored in most evaluation studies. It describes how people who do not switch sectors are affected by the policy. The third term is the direct marginal social cost of the policy change. It includes the cost of administering the program plus the opportunity cost of consumption foregone to raise the taxes used to finance the program. Below we demonstrate the empirical importance of accounting for the full social costs of programs.

At an optimum, $\Delta(\varphi) = 0$, provided standard second order conditions are satisfied. Marginal benefit should equal the marginal cost. We can use either a cost-based measure of marginal benefit or a benefit-based measure of cost to evaluate the marginal gains or marginal costs of the program, respectively.

Observe that the local average treatment effect is simply the effect of treatment on the treated for persons at the margin ($X = -k(\varphi)$):

$$(3.14) \quad E(Y_1 | D = 1, X = -k(\varphi), \varphi) - E(Y_0 | D = 0, X = -k(\varphi), \varphi) \\ = E(Y_1 - Y_0 | D = 1, X = -k(\varphi), \varphi).$$

This expression is obvious once it is recognized that the set $X = -k(\varphi)$ is the indifference set. Persons in that set are indifferent between participating in the program and not participating. The Imbens and Angrist (1994) parameter is a marginal version of the “treatment on the treated” evaluation parameter for gross outcomes. This parameter is one of the ingredients required to produce an evaluation of the impact of a marginal change

in the social program on total output but it ignores costs and the effect of a change in the program on the outcomes of persons who do not switch sectors.⁸

The conventional evaluation parameter,

$$E(Y_1 - Y_0 \mid D = 1, x, \varphi)$$

does not incorporate costs, does not correspond to a marginal change and includes rents accruing to persons. This parameter is in general inappropriate for evaluating the effect of a policy change on GDP. However, under certain conditions which we now specify, this parameter is informative about the gross gain accruing to the economy from the existence of a program at level $\tilde{\varphi}$ compared to the alternative of shutting it down. This is the information required for an “all or nothing” evaluation of a program.

The appropriate criterion for an all or nothing evaluation of a policy at level $\varphi = \tilde{\varphi}$ is

$$A(\tilde{\varphi}) = \{N_1(\tilde{\varphi})E(Y_1 \mid D = 1, \varphi = \tilde{\varphi}) + N_0(\tilde{\varphi})E(Y_0 \mid D = 0, \varphi = \tilde{\varphi}) - c(\tilde{\varphi})\} \\ - \{N_1(0)E(Y_1 \mid D = 1, \varphi = 0) + N_0(0)E(Y_0 \mid D = 0, \varphi = 0)\}$$

where $\varphi = 0$ corresponds to the case where there is no program, so that $N_1(0) = 0$ and $N_0(0) = \bar{N}$. If $A(\tilde{\varphi}) > 0$, total output is increased by establishing the program at level $\tilde{\varphi}$.

In the special case where the outcome in the benchmark state “0” is the same whether or not the program exists,

$$(3.15) \quad E(Y_0 \mid D = 0, \varphi = \tilde{\varphi}) = E(Y_0 \mid D = 0, \varphi = 0).$$

This condition defines the absence of general equilibrium effects in the base state so the no program state for nonparticipants is the same as the nonparticipation state. Assumption (3.15) is what enables analysts to generalize from partial equilibrium to general equilibrium settings. Recalling that $\bar{N} = N_1(\varphi) + N_0(\varphi)$, when (3.15) holds we have

$$(3.16) \quad A(\tilde{\varphi}) = N_1(\tilde{\varphi})E(Y_1 - Y_0 \mid D = 1, \varphi = \tilde{\varphi}) - c(\tilde{\varphi}).^9$$

Given costless redistribution of the benefits, the output-maximizing solution for φ also maximizes social welfare. For this important case, which is applicable to small-scale social programs with partial participation, the measure “treatment on the treated” which we focus on in this chapter is justified. For evaluating the effect of marginal variation or “fine-tuning” of existing policies, measure $\Delta(\varphi)$ is more appropriate.¹⁰

⁸Heckman and Smith (1998a) and Heckman (1997) present comprehensive discussions of the Imbens and Angrist (1994) parameter. We discuss this parameter further in Section 7. One important difference between their parameter and the traditional treatment on the treated parameter is that the latter excludes variables like φ from the conditioning set, but the Imbens-Angrist parameter includes it.

⁹Condition (3.15) is stronger than what is required to justify (3.16). The condition only has to hold for the subset of the population ($N_0(\varphi)$ in number) who would not participate in the presence of the program.

¹⁰Björklund and Moffitt (1987) estimate both the marginal gross gain and the average gross gain from participating in a program. However, they do not present estimates of marginal or average costs.

4 Prototypical Solutions to the Evaluation Problem

An evaluation entails making some comparison between “treated” and “untreated” persons. This section considers three widely-used comparisons for estimating the impact of treatment on the treated: $E(Y_1 - Y_0 | X, D = 1)$. All use some form of comparison to construct the required counterfactual $E(Y_0 | X, D = 1)$. Data on $E(Y_1 | X, D = 1)$ are available from program participants. A person who has participated in a program is paired with an “otherwise comparable” person or set of persons who have not participated in it. The set may contain just one person. In most applications of the method, the paired partner is not literally assumed to be a replica of the treated person in the untreated state although some panel data evaluation estimators make such an assumption. Thus, in general, $\Delta = Y_1 - Y_0$ is not estimated exactly. Instead, the outcome of the paired partners is treated as a proxy for Y_0 for the treated individual and the population mean difference between treated and untreated persons is estimated by averaging over all pairs. The method can be applied symmetrically to nonparticipants to estimate what they would have earned if they had participated. For that problem the challenge is to find $E(Y_1 | X, D = 0)$ since the data on nonparticipants enables one to identify $E(Y_0 | X, D = 0)$.

A major difficulty with the application of this method is providing some objective way of demonstrating that a candidate partner or set of partners is “otherwise comparable.” Many econometric and statistical methods are available for adjusting differences between persons receiving treatment and potential matching partners which we discuss in Section 7.

4.1 The Before-After Estimator

In the empirical literature on program evaluation, the most commonly-used evaluation strategy compares a person with himself/herself. This is a comparison strategy based on longitudinal data. It exploits the intuitively-appealing idea that persons can be in both states at different times, and that outcomes measured in one state at one time are good proxies for outcomes in the same state at other times at least for the no-treatment state. This gives rise to the motivation for the simple “before-after” estimator which is still widely used. Its econometric descendent is the fixed effect estimator without a comparison group.

The method assumes that there is access either (i) to longitudinal data on outcomes measured before and after a program for a person who participates in it, or (ii) to repeated cross section data from the same population where at least one cross section is from a period prior to the program. To incorporate time into our analysis, we introduce “ t ” subscripts. Let Y_{1t} be the post-program earnings of a person who participates in the program. When longitudinal data are available, $Y_{0t'}$ is the pre-program outcome of the

person. For simplicity, assume that program participation occurs only at time period k , where $t > k > t'$. The “before-after” estimator uses preprogram earnings $Y_{0t'}$ to proxy the treatment state in the post program period. In other words, the underlying identifying assumption is

$$(4.A.1) \quad E(Y_{0t} - Y_{0t'} \mid D = 1) = 0.$$

If this assumption is valid, the “Before-After” estimator is given by

$$(4.1) \quad (\bar{Y}_{1t} - \bar{Y}_{0t'})_1,$$

where the subscript “1” denotes conditioning on $D = 1$, and the “-” denotes sample means.

To see how this estimator works, observe that for each individual the gain from the program may be written as

$$Y_{1t} - Y_{0t} = (Y_{1t} - Y_{0t'}) + (Y_{0t'} - Y_{0t}).$$

The second term ($Y_{0t'} - Y_{0t}$) is the approximation error. If this term averages out to zero, we may estimate the impact of participation on those who participate in a program by subtracting participants’ mean pre-program earnings from the mean of their post-program earnings. These means also may be defined for different values of participants’ characteristics, X .

The before-after estimator does not literally require longitudinal data to identify the means (Heckman and Robb, 1985a,b). As long as the approximation error averages out, repeated cross-sectional data that sample the same population over time, but not necessarily the same persons, are sufficient to construct a before-after estimate. An advantage of this approach is that it only requires information on the participants and their pre-participation histories to evaluate the program.

The major drawback to this estimator is its reliance on the assumption that the approximation errors average out. This assumption requires that among participants, the mean outcome in the no-treatment state is the same in t and t' . Changes in the overall state of the economy between t and t' , or changes in the life cycle position of a cohort of participants, can violate this assumption.

A good example of a case in which assumption (4.A.1) is likely violated is provided in the work of Ashenfelter (1978). Ashenfelter observed that prior to enrollment in a training program, participants experience a decline in their earnings. Later research demonstrates that Ashenfelter’s “dip” is a common feature of the pre-program earnings of participants in government training programs. See Figures 4.1 to 4.6 which show the dip for a variety of programs in different countries. If this decline in earnings is transitory, and earnings is a mean-reverting process so that the dip is eventually restored, even in the absence of participation in the program, and if period t' falls in the period of transitorily low

earnings, then the approximation error will not average out. In this example, the before-after estimator overstates the average effect of training on the trained and attributes mean reversion that would occur under any event to the effect of the program. On the other hand, if the decline is permanent, the before-after estimator is unbiased for the parameter of interest. In this case, any improvement in earnings is properly attributable to the program. Another potential defect of this estimator is that it attributes to the program any trend in earnings due to macro or lifecycle factors.

Two different approaches have been used to solve these problems with the before-after estimators. One controversial method generalizes the before-after estimator by making use of many periods of pre-program data and extrapolating from the period before t' to generate the counterfactual state in period t . It assumes that Y_{0t} and $Y_{0t'}$ can be adjusted to equality using data on the same person, or the same populations of persons, followed over time. As an example, suppose that Y_{0t} is a function of t , or is a function of t -dated variables. If we have access to enough data on pre-program outcomes prior to date t' to extrapolate post-program outcomes Y_{0t} , and if there are no errors of extrapolation, or if it is safe to assume that such errors average out to zero across persons in period t , one can replace the missing data or at least averages of the missing data, using extrapolated values. This method is appropriate if population mean outcomes evolve as deterministic functions of time or macroeconomic variables like unemployment. This procedure is discussed further in Section 7.5.¹¹ The second approach is based on the before-after estimator which we discuss next.

4.2 The Difference-in-Differences Estimator

A more widely used approach to the evaluation problem assumes access either (i) to longitudinal data or (ii) to repeated cross-section data on nonparticipants in periods t and t' . If the mean change in the no-program outcome measures are the same for participants and nonparticipants *i.e.* if the following assumption is valid:

$$(4.A.2) \quad E(Y_{0t} - Y_{0t'} \mid D = 1) = E(Y_{0t} - Y_{0t'} \mid D = 0)$$

then the *difference-in-differences* estimator given by

$$(4.2) \quad (\bar{Y}_{1t} - \bar{Y}_{0t'})_1 - (\bar{Y}_{0t} - \bar{Y}_{0t'})_0 \quad t > k > t' :$$

is valid for $E(\Delta_t \mid D = 1) = E(Y_{1t} - Y_{0t} \mid D = 1)$ where $\Delta_t = Y_{1t} - Y_{0t}$ because

$$E[(\bar{Y}_{1t} - \bar{Y}_{0t'})_1 - (\bar{Y}_{0t} - \bar{Y}_{0t'})_0] = E(\Delta_t \mid D = 1).^{12}$$

¹¹See also Heckman and Robb (1985a), p. 210-215.

¹²The proof is immediate. Make the following decomposition

$$(\bar{Y}_{1t} - \bar{Y}_{0t'})_1 = (\bar{Y}_{1t} - \bar{Y}_{0t})_1 + (\bar{Y}_{0t} - \bar{Y}_{0t'})_1.$$

If assumption (4.A.2) is valid, the change in the outcome measure in the comparison group serves to benchmark common year or age effects among participants.

Because we cannot form the change in outcomes between the treated and untreated states, the expression

$$(Y_{1t} - Y_{0t'})_1 - (Y_{0t} - Y_{0t'})_0,$$

cannot be formed for anyone, although we can form one or the other of these terms for everyone. Thus, we cannot use the difference-in-differences estimator to identify the *distribution* of gains without making further assumptions.¹³ Like the before-after estimator, we can implement the difference-in-differences estimator for means (4.2) on repeated cross sections. It is not necessary to sample the same persons in periods t and t' —just persons from the same populations.

Ashenfelter’s dip provides an example of a case where assumption (4.A.2) is likely to be violated. If Y is earnings, and t' is measured at the time of a transitory earnings dip, and if non-participants do not experience the dip, then (4.A.2) will be violated, because the time path of no-program earnings between t' and t will be different between participants and nonparticipants. In this example, the difference-in-differences estimator overstates the average impact of training on the trainee.

4.3 The Cross-Section Estimator

A third estimator compares mean outcomes of participants and nonparticipants at time t . This estimator is sometimes called the cross-section estimator. It does not compare the same persons because by hypothesis a person cannot be in both states at the same time. Because of this fact, cross-section estimators cannot estimate the distribution of gains unless additional assumptions are invoked beyond those required to estimate mean impacts.

The key identifying assumption for the cross-section estimator of the mean is that

$$(4.A.3) \quad E(Y_{0t} \mid D = 1) = E(Y_{0t} \mid D = 0),$$

i.e., that on average persons who do not participate in the program have the same no-treatment outcome as those who do participate. If this assumption is valid, then the *cross-section* estimator is given by

The claim follows upon taking expectations.

¹³One assumption that identifies the distribution of gains is to assume that $(Y_{1t} - Y_{0t})_1$ is independent of $(Y_{0t} - Y_{0t'})_1$ and that the distribution of $(Y_{1t} - Y_{0t})_1$ is the same as the distribution of $(Y_{0t} - Y_{0t'})_0$. Then the results on deconvolution in Heckman, Smith and Clements (1997) can be applied. See their paper for details.

$$(4.3) \quad (\bar{Y}_{1t})_1 - (\bar{Y}_{0t'})_0.$$

This estimator is valid under assumption (4.A.3) because

$$E((\bar{Y}_{1t})_1 - (\bar{Y}_{0t'})_0) = E(\Delta_t \mid D = 1).^{14}$$

If persons go into the program based on outcome measures in the *post-program* state, then assumption (4.A.3) will be violated. The assumption would be satisfied if participation in the program is unrelated to outcomes in the no program state *in the post-program period*. Thus, it is possible for Ashenfelter’s dip to characterize the data on earnings in the pre-program period, and yet for (4.A.3) to be satisfied. Moreover, as long as the macro economy and aging process operate identically on participants and nonparticipants, the cross section estimator is not vulnerable to the problems that plague the before-after estimator.

The cross section estimator (4.3), the difference-in-differences estimator (4.2), and the before-after estimator (4.1) comprise the trilogy of conventional non-experimental evaluation estimators. All of these estimators can be defined conditional on observable characteristics X . Conditioning on X or additional “instrumental” variables make it more likely that modified versions of assumptions (4.A.3), (4.A.2), or (4.A.1) will be satisfied but this is not guaranteed. If, for example, the distribution of X characteristics is different between participants ($D = 1$) and nonparticipants ($D = 0$), conditioning on X may eliminate systematic differences in outcomes between the two groups. Using modern nonparametric procedures, it is possible to exploit each of the identifying conditions to estimate nonparametric versions of all three estimators. On the other hand, if the difference between participants and nonparticipants is due to unobservables, conditioning may accentuate, and not eliminate, differences between participants and nonparticipants in the no-program state.¹⁵

The three estimators exploit three different principles but all are based on making some comparison. The assumptions that justify one method will not, in general, justify any of the other methods. All of the estimators considered in this chapter exploit one of these three principles. They extend the simple mean differences just discussed by making a variety of adjustments to the means. Throughout the rest of the chapter, we organize our discussion of alternative estimators by discussing how they modify the simple mean differences used in the three intuitive estimators to account for nonstationary environments and different regressors in the different comparison groups. We first consider social experimentation and how it constructs the counterfactuals used in policy evaluations.

¹⁴Proof:

$$(\bar{Y}_{1t})_1 - (\bar{Y}_{0t'})_0 = (\bar{Y}_{1t})_1 - (\bar{Y}_{0t})_1 + (\bar{Y}_{0t})_1 - (\bar{Y}_{0t'})_0$$

and take expectations invoking assumption (4-A-3).

¹⁵Thus if $|E(Y_0 \mid D = 1) - E(Y_0 \mid D = 0)| = M$, there is no guarantee that $|E(Y_0 \mid D = 1, X) - E(Y_0 \mid D = 0, X)| < M$. For some values of X , the gap could widen.

5 Social Experiments

Randomization is one solution to the evaluation problem. Recent years have witnessed increasing use of experimental designs to evaluate North American employment and training programs. This approach has been less common in Europe, though a small number of experiments have been conducted in Britain, Norway and Sweden. When the appropriate qualifications are omitted, the impact estimates from these social experiments are easy for analysts to calculate and for policymakers to understand (see, e.g., Burtless, 1995). As a result of its apparent simplicity, evidence from social experiments has had an important impact on the design of U.S. welfare and training programs.¹⁶ Because of the importance of experimental designs in this literature, in this section we show how they solve the evaluation problem, describe how they have been implemented in practice, and discuss their advantages and limitations.

5.1 How Social Experiments Solve the Evaluation Problem.

An important lesson of this section is that social experiments, like other evaluation methods, provide estimates of the parameters of interest only under certain behavioral and statistical assumptions. To see this, let “*” denote outcomes in the presence of random assignment. Thus, conditional on X for each person we have (Y_1^*, Y_0^*, D^*) in the presence of random assignment and (Y_1, Y_0, D) when the program operates normally without randomization. Let $R = 1$ if a person for whom $D^* = 1$ is randomized into the program and $R = 0$ if the person is randomized out. Thus, $R = 1$ corresponds to the experimental treatment group and $R = 0$ to the experimental control group.

The essential assumption required to use randomization to solve the evaluation problem for estimating the mean effect of treatment on the treated is that

$$(5.A.1) \quad E(Y_1^* - Y_0^* | X, D^* = 1) = E(Y_1 - Y_0 | X, D = 1).$$

A stronger set of conditions, not strictly required, are

$$(5.A.2a) \quad E(Y_1^* | X, D^* = 1) = E(Y_1 | X, D = 1)$$

and

$$(5.A.2b) \quad E(Y_0^* | X, D^* = 1) = E(Y_0 | X, D = 1).$$

Assumption (5.A.2a) states that the means from the treatment and control groups generated by random assignment produce the desired population parameter. With certain exceptions discussed below, this assumption rules out changes in the impact of participation due to the presence of random assignment as well as changes in the process of program participation. The first part of this assumption can in principle be tested by comparing the

¹⁶We discuss this evidence in Section 10.

outcomes of participants under a regime of randomization with the outcome of participants under the usual regime.

If (5.A.2a) is true, among the population for whom $D = 1$ and $R = 1$ we can identify

$$E(Y_1 | X, D = 1, R = 1) = E(Y_1 | X, D = 1).$$

Under (5.A.2a) information sufficient to estimate this mean without bias is routinely produced from data collected on participants in social programs. The new information produced by an experiment comes from those randomized out of the program. Using the experimental control group it is possible to estimate:

$$E(Y_0 | X, D = 1, R = 0) = E(Y_0 | X, D = 1).$$

Thus, experiments produce data that satisfy assumption (4.A.3). Simple application of the cross-section estimator identifies

$$E(\Delta | X, D = 1) = E(Y_1 - Y_0 | X, D = 1).$$

Within the context of the model of equation (3.10), an experiment that satisfies (5.A.1) or (5.A.2a) and (5.A.2b) *does not* make D orthogonal to U . It simply equates the bias in the two groups $R = 1$ and $R = 0$. Thus in the model of equation (3.1), under (5.A.2a), $E(Y|X, D = 1, R = 1) = g_1(X) + E(U_1|X, D = 1)$ and $E(Y|X, D = 1, R = 0) = g_0(X) + E(U_0|X, D = 1)$.¹⁷

Rewriting the first conditional mean, we obtain

$$E(Y|X, D = 1, R = 1) = g_1(X) + E(U_1 - U_0|X, D = 1) + E(U_0|X, D = 1).$$

Subtracting the second mean from the first eliminates the common selection bias component $E(U_0|X, D = 1)$ so

$$E(Y|X, D = 1, R = 1) - E(Y|X, D = 1, R = 0) = g_1(X) - g_0(X) + E(U_1 - U_0|X, D = 1).$$

When the model (3.1) is specialized to one of intercept differences, as in (3.10), this parameter simplifies to α . Notice, that the method of social experiments does *not* set either $E(U_1|X, D = 1)$ or $E(U_0|X, D = 1)$ equal to zero. Rather, it balances the selection bias in the treatment and control groups.

¹⁷Notice that in this section we allow for the more general model $Y_0 = g_0(X) + U_0$, $Y_1 = g_1(X) + U_1$ where $E(U_0 | X) \neq 0$ and $E(U_1 | X) \neq 0$.

Stronger assumptions must be made to identify the distribution of impacts $F(\Delta | D = 1)$.¹⁸ Without invoking further assumptions, data from experiments, like data from non-experimental sources, are unable to identify the distribution of impacts because the same person is not observed in both states at the same time (Heckman, 1992; Heckman, Smith and Clements, 1997; Heckman and Smith, 1993, 1995, 1998a).

If assumption (5.A.1) or assumptions (5.A.2a) and (5.A.2b) fail to hold because the program participation probabilities are affected, so D^* and D are different, then the composition of the participant population differs in the presence of random assignment. In two important special cases, experimental data still provide unbiased estimates of the effect of treatment on the treated. First, if the effect of training is the same for everyone, changing the composition of the participants has no effect because the parameter of interest is the same for all possible participant populations (Heckman, 1992). This assumption is sometimes called the common treatment effect assumption and, letting i denote a variable value for individual i , may be formally expressed as

$$(5.A.3) \quad Y_{1i} - Y_{0i} = \Delta_i \equiv \Delta \text{ for all } i.$$

This assumption is equivalent to setting $U_1 = U_0$ in (3.9). Assumption (5.A.3) can be defined conditionally on observed characteristics, so we may write $\Delta = \Delta(X)$. Notice, however, that in this case, if randomization induces persons with certain X values not to participate in the program, then estimates of $\Delta(X)$ can only be obtained for values of X possessed by persons who participate in the program. In this case (5.A.1) is satisfied but (5.A.2a) and (5.A.2b) are not.

The second special case where experimental data still provide unbiased estimates of the effect of treatment on the treated arises when decisions about training are not affected by the realized gain from participating in the program. This case could arise if potential trainees know $E(\Delta | X)$ but not Δ at the time participation decisions are made. Formally, the second condition is

$$(5.A.4) \quad E(\Delta | X, D = 1) = E(\Delta | X),$$

which is equivalent to condition (3.11) in the model (3.9). If either (5.A.3) or (5.A.4) holds, the simple experimental mean difference estimator is unbiased for $E(\Delta | X, D = 1)$.

Randomization improves on the non-experimental cross-section estimator even if there is no selection bias. In an experiment, for all values of X for which $D = 1$, one can identify

$$E(\Delta | X, D = 1) = E(Y_1 - Y_0 | X, D = 1).$$

Using assumption (4.A.3) in an ordinary nonexperimental evaluation, there may be values of X such that $\Pr(D = 1 | X) = 1$; that is, there may be values of X with no comparison group members. Randomization avoids this difficulty by balancing the distribution of X

¹⁸Replace “ E ” with “ F ” in (5.A.2a) and (5.A.2b) to obtain one necessary condition.

values in the treatment and control groups (Heckman, 1996). At the same time, however, random assignment conditional on $D = 1$ cannot provide estimates of $\Delta(X)$ for values of X such that $\Pr(D = 1 | X) = 0$.

The stage of potential program participation at which randomization is applied - eligibility, application, or acceptance into a program - determines what can be learned from a social experiment. For randomization conditional on acceptance into a program ($D = 1$), we can estimate the effect of treatment on the treated:

$$E(\Delta | X, D = 1) = E(Y_1 - Y_0 | X, D = 1)$$

using simple experimental means. We cannot estimate the effect of randomly selecting a person to go into the program:

$$E(\Delta | X) = E(Y_1 - Y_0 | X),$$

by using simple experimental means unless one of two conditions prevails. The first condition is just the common effect assumption (5.A.3). This assumption is explicit in the widely-used dummy endogenous variable model (Heckman, 1978). The second condition is that embodied in assumption (5.A.4), that participation decisions are independent of the person-specific component of the impact. In both cases, the mean impact of treatment on a randomly selected person is the same as the mean impact of treatment on the treated.

In the general case, it is difficult to estimate the effect of randomly assigning a person with characteristics X to go into a program. This is because persons randomized into a program cannot be compelled to participate in it. In order to secure compliance, it may be necessary to compensate or persuade persons to participate. For example, in many U.S. social experiments, program operators threaten to reduce participants' social assistance benefits, if they refuse to participate in training. Such actions, even if successful, alter the environment in which persons operate and may make it impossible to estimate $E(\Delta | X)$ using experimental means. One assumption that guarantees compliance is the existence of a "compensation" or "punishment" level c such that

$$(5.A.5a) \quad \Pr(D = 1 | X, c) = 1$$

and

$$(5.A.5b) \quad E(\Delta | X, c) = E(\Delta | X).$$

The first part of the assumption guarantees that a person with characteristics X can be "bribed" or "persuaded" to participate in the program. The second part of the assumption guarantees that compensation c does not affect the outcome being evaluated.¹⁹ If c is a

¹⁹Observe that the value of c is not necessarily unique.

monetary payment, it would be optimal from the standpoint of an experimental analyst to find the minimal value of c that satisfies these conditions.

Randomization of eligibility is sometimes proposed as a less disruptive alternative to randomization conditional on $D = 1$. Randomizing eligibility avoids the application and screening costs that are incurred when accepted individuals are randomized out of a program. Because the randomization is performed outside of training centers, it also avoids some of the political costs that have accompanied the use of the experimental method.

Consider a population of persons who are usually eligible for the program. Randomize eligibility within this population. Let $e = 1$ if a person retains eligibility and $e = 0$ if a person becomes ineligible. Assume that eligibility does not disturb the underlying structure of the random variables (Y_0, Y_1, D, X) and that $\Pr(D = 1 | X) \neq 0$. Then Heckman (1996) shows that

$$\frac{E(Y | X, e = 1) - E(Y | X, e = 0)}{\Pr(D = 1 | X, e = 1)} = E(\Delta | X, D = 1).$$

Randomization of eligibility produces samples that can be used to identify $E(\Delta | X, D = 1)$ and also to recover $\Pr(D = 1 | X)$. The latter is not recovered from samples which condition on $D = 1$ (Heckman, 1992; Moffitt, 1992). Without additional assumptions of the sort previously discussed, randomization on eligibility will not, in general, identify $E(\Delta | X)$.

5.2 Intention to Treat and Substitution Bias

The objective of most experimental designs is to estimate the conditional mean impact of training, or $E(\Delta | X, D = 1)$. However, in many experiments a significant fraction of the treatment group drops out of the program and does not receive the services being evaluated.²⁰ In general, in the presence of dropping out $E(\Delta | X, D = 1)$ cannot be identified using comparisons of means. Instead, the experimental mean difference estimates the mean effect of the offer of treatment, or what is sometimes called the “intent to treat.” For many purposes, this is the policy-relevant parameter. It is informative on how the availability of a program affects participant outcomes. Attrition is a normal feature of an ongoing program.

To obtain an estimate of the impact of training on those who actually receive it, additional assumptions are required beyond (5.A.1) or (5.A.2a) and (5.A.2b). Let T be an indicator for actual receipt of treatment, with $T = 1$ for persons actually receiving training, and $T = 0$ otherwise. Let T^* be a similarly defined latent variable for control group

²⁰Using the analysis in the preceding subsection, dropping out by experimental treatment group members could be reduced by compensating them for completing training.

members indicating whether or not they would have actually received training, had they been in the treatment group. Define

$$E(\Delta | X, D = 1, R = 1, T = 1) = E(\Delta | X, D = 1, T = 1)$$

as the mean impact of training on those members of the treatment group who actually receive it. This parameter will equal the original parameter of interest $E(\Delta | X, D = 1)$ only in the special cases where (5.A.3), the common effect assumption, holds, or where an analog to (5.A.4) holds so that the decision of treatment group members to drop out is independent of $(\Delta - E(\Delta))$, the person-specific component of their impact.

A consistent estimate of the impact of training on those who actually received it can be obtained under the assumption that the mean outcome of the treatment group dropouts is the same as that of their analogs in the control group, so that

$$(5.A.6) \quad E(Y | X, D = 1, R = 1, T = 0) = E(Y | X, D = 1, R = 0, T^* = 0).$$

Note that this assumption rules out situations where the treatment group dropouts receive potentially valuable partial treatment. Under (5.A.6),

$$(5.1) \quad \frac{E(Y | X, D = 1, R = 1) - E(Y | X, D = 1, R = 0)}{P(T = 1 | X, D = 1, R = 1)}$$

identifies the mean impact of training on those who receive it.²¹ This estimator scales up the experimental mean difference estimate by the fraction of the treatment group receiving training. When all treatment group members receive training, the denominator equals one and the estimator reduces to the simple experimental mean difference. Estimator (5.1) also shows that the simple mean difference estimator provides a downward biased estimate of the mean impact of training on the trained when there are dropouts from the treatment group, because the denominator always lies between zero and one. Heckman, Smith and Taber (1998) present methods for estimating distributions of outcomes and for testing the identifying assumptions in the presence of dropping out. They present evidence on the validity of the assumptions that justify (5.1) in the National JTPA Study data.

In an experimental evaluation, the converse problem can also arise for the control group members. In an ideal experiment, no control group members would receive either the experimental treatment or close substitutes to it from other sources. In practice, a significant fraction of controls often receives similar services from other sources. In this situation, the mean earnings of control group members no longer correspond to $E(Y_0 | X, D = 1)$ and neither the experimental mean difference estimator nor the adjusted estimator (5.1) identifies the impact of training relative to no training for those who receive it. However, under certain conditions discussed in Section 3, the experimental estimate can be interpreted as the mean incremental effect of the program relative to a world in which it does not exist.

²¹See, e.g., Mallar (1978), Bloom (1984) and Heckman, Smith and Taber (1998).

As in the case of treatment group dropouts, identifying the impact of training on the trained in the presence of control group substitution requires additional assumptions beyond (5.A.1) or (5.A.2a) and (5.A.2b). Let $S = 1$ denote control group members receiving substitute training from alternative sources and let $S = 0$ denote control group members receiving no training and let Y_2 be the outcome conditional on receipt of alternative training. Consider the general case with both treatment group dropping out and control group substitution. In this context, one approach would be to invoke the assumptions required to apply non-experimental techniques as described in Section 7 to the treatment group data to obtain an estimate of the impact of the training being evaluated on those who receive it. Heckman, Hohmann, Khoo and Smith (1998) employ this and other strategies using data from the National JTPA Study.

Alternatively, two other assumptions allow use of the control group data to estimate the impact of training on the trained. The first assumption is a generalized common effect assumption, where to distinguish individuals we restore subscript i

$$(5.A.3') \quad Y_{1i} - Y_{0i} = Y_{2i} - Y_{0i} = \Delta_i \equiv \Delta \text{ for all } i.$$

This assumption states that (a) the impact of the program being evaluated is the same as the impact of substitute programs for each person and (b) that all persons respond exactly the same way to the program (a common effect assumption). The second assumption is a generalized version of (5.A.4), where

$$(5.A.4') \quad E(Y_1 - Y_0 \mid X, D = 1, T = 1, R = 1) = E(Y_2 - Y_0 \mid X, D = 1, S = 1, R = 0).$$

This assumption states that the mean impact of the training being evaluated received by treatment group members who do not drop out equals the mean impact of substitute training on those control group members who receive it. Both (5.A.3') and (5.A.4') are strong assumptions. To be plausible, either would require evidence that the training received by treatment group members was similar in content and duration to that received by control group members. Note that (5.A.3') implies (5.A.4'). Under either assumption, the ratio

$$(5.2) \quad \frac{E(Y \mid X, D = 1, R = 1) - E(Y \mid X, D = 1, R = 0)}{\Pr(T = 1 \mid X, D = 1, R = 1) - \Pr(S = 1 \mid X, D = 1, R = 0)}$$

identifies the mean impact of training on those who receive it in both the experimental treatment and control groups, provided that the denominator is not zero. The similarity of estimator (5.2) to the instrumental variable estimator defined in Section 7 is not accidental; under assumptions (5.A.3') or (5.A.4'), random assignment is a valid instrument for training because it is correlated with training receipt but not with any other determinants of the outcome Y . Without one of these assumptions, random assignment is not, in general, a valid instrument (Heckman, 1997; Heckman, Hohmann, Khoo and Smith, 1998). To see this point, consider a model in which individuals know their gain from training, but because the treatment group has access to the program being evaluated, it faces a lower cost of training. In this case, controls are less likely to be trained, but the mean gross impact

would be larger among control trainees than among the treatment trainees. Drawing on the analysis of Section 7, this correlation violates the condition required for the IV estimator to identify the parameter of interest.

5.3 Social Experiments in Practice

In this subsection we discuss how social experiments operate in practice. We present empirical evidence on some of the theoretical issues surrounding social experiments discussed in the preceding subsections and provide a context for the discussion of the experimental evidence on the impact of training in Section 10. To make the discussion concrete, we focus in particular on two of the best known U.S. social experiments: the National Supported Work (NSW) demonstration (Hollister, et al., 1984) and the recent National JTPA Study (NJS).²² We begin with a brief discussion of the implementation of these two experiments.

5.3.1 Two Important Social Experiments

The NSW Demonstration was one of the first employment and training experiments. It tested the effect of 9 to 18 months of guaranteed work experience in unskilled occupations on groups of long-term AFDC (welfare) recipients, ex-drug addicts, ex-criminal offenders, and economically disadvantaged youths in 10 sites across the U.S. These jobs were in a sheltered environment in which productivity standards were gradually raised over time and participants met frequently with program counselors to discuss grievances and performance.

The NSW enrollment process began with a referral, usually by a welfare agency, drug rehabilitation agency, or prisoners' assistance society. Program operators then interviewed potential participants and eliminated any persons that they believed "would be disruptive to their programs" (Hollister, et al., 1984, p. 35). Following this screening, a third party randomly assigned one-half of the qualified applicants to the treatment group. The remainder were assigned to the control group and prevented from receiving NSW services. Although the controls could not receive NSW services, program administrators could not prevent them from receiving other training services in their community, such as those offered under another widely available training program with the acronym CETA. Follow-up data on the experimental treatment and control groups were collected via both surveys and administrative earnings records.

In contrast to the NSW, the NJS sought to evaluate the effectiveness of an ongoing training program. From the start, the goal of evaluating an ongoing program without significantly disrupting its operations – and thereby violating assumption (5.A.1) or assumptions (5.A.2a) and (5.A.2b) – posed significant problems. The first of these arose

²²See, among others, Doolittle and Traeger (1990), Bloom, et al. (1993) and Orr, et al. (1994).

in selecting the training centers at which random assignment would take place. Initially, evaluators planned to use a random sample of the nearly 600 U.S. JTPA training sites. Randomly choosing the evaluation sites would enhance the “external validity” of the experiment – the extent to which its findings can be generalized to the population of JTPA training centers. Yet, it was difficult to persuade local administrators to participate in an evaluation that required them to randomly deny services to eligible applicants. When only four of the randomly selected sites or their alternates agreed to participate, the study was redesigned to include a “diverse” group of 16 centers willing to participate in a random assignment study (see Doolittle and Traeger, 1990; or the summary of their analysis presented in Hotz, 1992). Evaluators had to contact 228 JTPA training centers in order to obtain these sixteen volunteers.²³ The option of forcing centers to participate was rejected because of the importance of securing the cooperation of local administrators in preserving the integrity of random assignment. Such concerns are not without foundation, as the integrity of an experimental training evaluation in Norway was undermined by the behavior of local operators (Torp, et al., 1993).

Concerns about disrupting normal program operations and violating (5.A.1) or (5.A.2a)-(5.A.2b) also led to an unusual approach to the evaluation of the specific service types provided by JTPA. This program offers a personalized mix of employment and training services including all those listed in Table 2.1 with the exception of public service employment. During their enrollment in the program, participants may receive two or more of these services in sequence, where the sequence may depend on the participant’s success or failure in those services provided first. As a result of this heterogeneous, fluid structure, it was impossible without changing the character of the program to conduct random assignment conditional on (planned) receipt of particular services or sets of services. Instead, JTPA staff recommended particular services for each potential participant prior to random assignment, and impact estimates were calculated conditional on these recommendations. In particular, the recommendations were grouped into three “treatment streams”: the “CT-OS stream” which included persons recommended for classroom training, CT, (and possibly other services), OS, but not on the job training or OJT; the “OJT stream” which included persons recommended for OJT (and possibly other services) but not CT; and the “other stream” which included the rest of the admitted applicants, most of whom ended up receiving only job search assistance. Note that this issue did not arise in the NSW, which provided a single service to all of its participants. In the NJS, follow-up data on earnings, employment and other outcomes were obtained from both surveys and multiple

²³Very large training centers (e.g., Los Angeles) and small, rural centers were excluded from the study design from the outset of the center enrollment process, for administrative and cost reasons, respectively. The final set of 16 training centers received a total of \$1 million in payments to cover the cost of participating in the experiment.

administrative data sources.

5.3.2 The Practical Importance of Dropping Out and Substitution

The most important problems affecting social experiments are treatment group dropout and control group substitution. These problems are not unique to experiments. Persons drop out of programs whether or not they are experimentally evaluated. There is no evidence that the rate of dropping out increases during an experimental evaluation. Most programs have good substitutes so that the estimated effect of a program as typically estimated is in relation to the full range of activities in which nonparticipants engage. Experiments exacerbate this problem by creating a pool of persons who attempted to take training who then flock to substitute programs when they are placed in an experimental control group.

Table 5.1 demonstrates the practical importance of these problems in experimental evaluations by reporting the rates of treatment group dropout and control group substitution from a variety of social experiments. It reveals that the fraction of treatment group members receiving program services is often less than 0.7, and sometimes less than 0.5. Furthermore, the observed characteristics of the treatment group members who drop out often differ from those who remain and receive the program services.²⁴ In regard to substitution, Table 5.1 shows that as many as 40 percent of the controls in some experiments received substitute services elsewhere. In an ideal experiment, all treatments receive the treatment and there is no control group substitution, so that the difference between the fraction of treatments and controls that receive the treatment equals 1.0. In practice, this difference is often well below 1.0.

The extent of both substitution and dropout depends on the characteristics of the treatment being evaluated and the local program environment. In the NSW, where the treatment was relatively unique and of high enough quality to be clearly perceived as valuable by participants, dropout and substitution rates were low enough to approximate the ideal case. In contrast, in the NJS and other evaluations of programs that provide low cost services widely available from other sources, substitution and dropout rates are high.²⁵ In the NJS, the substitution problem is accentuated by the fact that JTPA relies on

²⁴For the NSW, see LaLonde (1984); for the NJS see Smith (1992).

²⁵For the NJS, Table 5.1 reveals the additional complication that estimates of the rate of training receipt in the treatment and control groups depend on the data source used to make the calculation. In particular, because many treatment group members do not report training that administrative records show they received, dropout rates measured using only the survey data are substantially higher than those that combine the survey and administrative data. At the same time, because administrative data are not available on control group training receipt (other than the very small number of persons who defeated the experimental protocol), using only self-report data on controls but the combined data for the treatment group will likely overstate the difference in service receipt levels between the two groups.

outside vendors to provide most of its training. Many of these vendors, such as community colleges, provide the same training to the general public, often with subsidies from other government programs such as Pell Grants. In addition, in order to help in recruiting sites to participate in the NJS, evaluators allowed them to provide control group members with a list of alternative training providers in the community. Of the 16 sites in the NJS, 14 took advantage of this opportunity to alert control group members to substitute training opportunities.

To see the effect of high of dropping out and substitution on the interpretation of the experimental evidence, consider Project Independence. The unadjusted experimental impact estimate is \$264 over the 2-year follow-up period, while application of the IV estimator that uses sample moments in place of (5.2) yields an adjusted impact estimate of \$1,100 (\$264/0.24). The first estimate indicates the mean impact of the offer of treatment relative to the other employment and training opportunities available in the community. Under assumptions (5.A.3') or (5.A.4'), the latter estimate indicates the impact of training relative to no training in both the treatment and control groups. Under these assumptions, the high rates of dropping out and substitution suggest that, the experimental mean difference estimate is strongly downward biased as an estimate of the impact of treatment on the treated, the primary parameter of policy interest.

A problem unique to experimental evaluations is violation of (5.A.1), or (5.A.2a) and (5.A.2b) which produces what Heckman (1992) and Heckman and Smith (1993, 1995) call “randomization bias.” In the NJS, this problem took the form of concerns that expanding the pool of accepted applicants, which was required to keep the number of participants at normal levels while creating a control group, would change the process of selection of persons into the program. Specifically, training centers were concerned that the additional recruits brought in during the experiment would be less motivated and harder to train and therefore benefit less from the program. Concerns about this problem were frequently cited by training centers that declined to participate in the NJS (Doolittle and Traeger, 1990). To partially allay these concerns, random assignment was changed from the 1:1 ratio that minimizes the sampling variance of the experimental impact estimator to a 2:1 ratio of treatments to controls.

Although we have no direct evidence on the empirical importance of changes in participation patterns on measured outcomes during the NJS, there is some indirect evidence about the validity of (5.A.1) or (5.A.2a) and (5.A.2b) in this instance. First of all, a number of training centers in the NJS streamlined their intake processes during the experiment – sometimes with the help of an intake consulting firm whose services were subsidized as part of the evaluation. In so doing, they generally reduced the number of visits and other costs paid by potential trainees, thereby including among those randomly assigned less motivated persons than were normally served. Second, some training centers asked for, and received,

additional temporary reductions in the random assignment ratio during the course of the experiment when they experienced difficulties recruiting sufficient qualified applicants to keep the program operating at normal levels.

A second problem unique to experiments involves obtaining experimental estimates of the effects of individual components of services provided in sequence as part of a single program. Experimental designs can readily determine how access to a bundle of services affects participants' earnings. More difficult is the question of how participation at each stage influences earnings, when participants can drop out during the sequence. Providing an experimental answer to this question requires randomization at each stage in the sequence.²⁶ In a program with several stages, this would lead to a proliferation of treatments and either large (and costly) samples or insufficient sample sizes. In practice, such sequential randomization has not been attempted in evaluating job training programs.

A final problem unique to experimental designs is that even under ideal conditions, they are unable to answer many questions of interest besides the narrow impact of "treatment on the treated" parameter. For example, it is not possible in practice to obtain simple experimental estimates of the duration of post-random assignment employment due to post-random assignment selection problems (Ham and LaLonde, 1990). An elaborate analysis of self-selection of the sort sought to be avoided by social experiments is required. As another example, consider estimating the impact of training on wage rates. The problem that arises in this case is that we observe wages only for those employed following random assignment. If the experimental treatment affects employment, then the sample of employed treatments will have different observed and unobserved characteristics than the employed controls. In general, we would expect that the persons without wages will be less skilled. The experimental impact estimate cannot separate out differences between the distribution of observed wages in the treatment and control groups that result from the effect of the program on wage rates from the effect of the program on selection into employment. Under these circumstances, only non-experimental methods such as those discussed in Section 7 can provide an answer to the question of interest.

5.3.3 Additional Problems Common to All Evaluations

There are a number of other problems that arise in both social experiments and non-experimental evaluations. Solving these problems in an experimental setting requires an-

²⁶Alternatively, in a program with three stages, program administrators might randomly assign eligible participants to one of several treatment groups, with the first group receiving only stage 1 services, the second receiving stage 1 and stage 2 services and the third receiving services from all three stages. However, a problem may arise with this scheme if participants assigned to the second and third stages of the program at some point decline to participate. In that case, the design described in the text would be more effective.

alysts to make the same types of choices (and assumptions) that are required in a non-experimental analysis. An important point of this subsection is that experimental impact estimates are sensitive to these choices in the same way as non-experimental estimates. A related concern is that experimental evaluations should, but often do not, include sensitivity analyses indicating the effect of the choices made on the impact estimates obtained.

The first common evaluation problem arises from imperfect data. Different survey instruments can yield different measures for the same variable for the same person in a given time period (see Smith, 1997a,b, and the citations therein). For example, self-reported measures of earnings or welfare receipt from surveys typically differ from administrative measures covering the same period (LaLonde and Maynard, 1987; Bloom, et al., 1993). As we discuss in Section 8, in the case of earnings, data sources commonly used for evaluation research differ in the types of earnings covered, the presence or absence of top-coding and the extent of missing or incorrect values. The evaluator must trade off these factors when choosing which data source to rely on. Whatever the data source used, the analyst must make decisions about how to handle outliers and missing values.

To underscore the point that experimental impacts for the same program can differ due to different choices about data sources and data handling, we compare the impact estimates for NJS presented in the two official experimental impact reports, Bloom, et al. (1993) and Orr, et al. (1994).²⁷ As shown in Table 5.2, these two reports give substantially different estimates of the impact of JTPA training for the same demographic groups over the same time period. The differences result from different decisions about whom to include in the evaluation sample, how to combine earnings information from surveys and administrative data, how to treat seemingly anomalous reports of overtime earnings in the survey data and so on. Several of the point estimates differ substantially, as do the implications about the relative effectiveness of the three treatment streams for adult women. The estimated 18-month impact for adult women in the “other services” stream triples from the 18-month impact report to the 30-month impact report, making it the service with the largest estimated impact despite the low average cost of the services provided to persons in this stream.

The second problem common to experimental and non-experimental evaluations is sample attrition. Note that sample attrition is not the same as dropping out of the program. Both control and treatment group members can attrit from the sample and treatment group members who drop out of the program will often remain in the data. In the NSW, attrition from the evaluation sample by the 18 month follow-up interview was 10 percent for the adult women, but more than 30 percent for the male participants. In the NJS study, sample attrition by the 18 month follow-up was 12 percent for the adult women and ap-

²⁷A complete discussion of the impact estimates from the NJS appears in Section 10.

proximately 20 percent of the adult males. Such high rates of attrition are common among the disadvantaged due to relatively frequent changes in residence and other difficulties with making follow-up contacts.

Sample attrition poses a problem for experimental evaluations when it is correlated with individual characteristics or with the impact of treatment conditional on characteristics. In practice, persons with poorer labor market characteristics tend to have higher attrition rates (see, e.g., Brown, 1979). Even if attrition affects both experimental and control groups in the same way, the experiment estimates the mean impact of the program only for those who remain in the sample. Usually, attrition rates are both non-random and larger for controls than for treatments. In this case, the experimental estimate of training is biased because individuals' experimental status, R , is correlated with their likelihood of being in the sample. In this setting, experimental evaluations become non-experimental evaluations because evaluators must make some assumption to deal with selection bias.

6 Econometric Models of Outcomes and Program Participation

The economic approach to program evaluation is based on estimating behavioral relationships that can be applied to evaluate policies not yet implemented. A focus on invariant behavioral relationships is the cornerstone of the econometric approach. Economic relationships provide frameworks within which empirical knowledge can be accumulated across different studies. They offer guidance on the specification of empirical relationships for any given study and the type of data required to estimate a behaviorally-motivated evaluation model. Alternative empirical evaluation strategies can be judged, in part, by the economic justification for them. Estimators that make economically implausible or empirically unjustified assumptions about behavior should receive little support.

The approach to evaluation guided by economic models is in contrast with the case-by-case approach of statistics that at best offers intuitive frameworks for motivating estimators. The emphasis in statistics is on particular estimators and not on the models motivating the estimators. The output of such case by case studies often does not cumulate. Since no articulated behavioral theory is used in this approach, it is not helpful in organizing evidence across studies or in suggesting explanatory variables or behaviorally motivated empirical relationships for a given study. It produces estimated parameters that are very difficult to use in answering well posed evaluation questions.

All economic evaluation models have two ingredients: (a) a model of outcomes and (b) a model of program participation. This section presents several prototypical econometric models. The first was developed by Heckman (1978) to rationalize the evidence in Ashenfelter (1978). The second rationalizes the evidence presented in Heckman and Smith (1998b) and Heckman, Ichimura, Smith and Todd (1998).

6.1 Uses of Economic Models

There are several distinct uses of economic models. (1) They suggest lists of explanatory variables that might belong in both outcome and participation equations. (2) They sometimes suggest plausible “exclusion restrictions” - variables that influence participation but do not directly influence outcomes, that can be used to help identify models in the presence of self-selection by participants. (3) They sometimes suggest specific functional forms of estimating equations motivated by *a priori* theory or by cumulated empirical wisdom.

6.2 Prototypical Models of Earnings and Program Participation

To simplify the discussion, and start where the published literature currently stops, assume that persons have only one period in their lives - period k - where they have the chance to take job training. From the beginning of economic life, $t = 1$ up through $t = k$, persons have one outcome associated with the no-training state “0”:

$$Y_{0j} \quad j = 1, \dots, k.$$

After period k , there are two potential outcomes corresponding to the training outcome (denoted “1”) and the no-training outcome (“0”):

$$(Y_{0j}, Y_{1j}) \quad j = k + 1, \dots, T$$

where T is the end of economic life.

Persons participate in training only if they apply to a program and are accepted into it. Several decision makers may be involved: individuals, family members and bureaucrats. Let $D = 1$ if a person participates in a program; $D = 0$ otherwise. Then the full description of participation and potential outcomes is

$$(6.1) \quad (D; Y_{0t}, t = 1, \dots, k; (Y_{0t}, Y_{1t}), t = k + 1, \dots, T).$$

As before, observed outcomes after period k can be written as a switching regression model:

$$Y_{0t} = DY_{1t} + (1 - D)Y_{0t}.$$

The most familiar model and the one that is most widely used in the training program evaluation literature assumes that program participation decisions are based on individual choices based on the maximization of the expected present value of earnings. It ignores family and bureaucratic influences on participation decisions.

6.3 Expected Present Value of Earnings Maximization

In period k , a prospective trainee seeks to measure the expected present value of earnings. Earnings is the outcome of interest. The information available to the agent in period k is I_k . The cost of program participation consists of two components: c (direct costs) and foregone earnings during the period. Training takes one period to complete. Assume that credit markets are perfect so that agents can lend and borrow freely at interest rate r . The expected present value of earnings maximizing decision rule is to participate in the program ($D = 1$) if

$$(6.2) \quad E \left[\sum_{j=1}^{T-k} \frac{Y_{1,k+j}}{(1+r)^j} - c - \sum_{j=0}^{T-k} \frac{Y_{0,k+j}}{(1+r)^j} \mid I_k \right] \geq 0,$$

and not to participate in the program ($D = 0$) if this inequality does not hold. In (6.2), the expectations are computed with respect to the information available to the person in period $k(I_k)$. It is important to notice that the expectations in (6.2) are the private expectations of the decision maker. They may or may not conform to the expectations computed against the true ex ante distribution. Note further that I_k may differ among persons in the same environment or may differ among environments. Many variables external to the model may belong in the information sets of persons. Thus friends, relatives and other channels of information may affect personal expectations.²⁸

The following are consequences of this decision rule. (a) Older persons, and persons with higher discount rates, are less likely to take training. (b) Earnings prior to time period k are irrelevant for determining participation in the program except for their value in forecasting future earnings. (*i.e.* except as they enter the person's information set I_k). (c) Only current costs and the discounted gain to earnings determine participation in the program. Persons with lower foregone earnings and lower direct costs of program participation are more likely to go into the program. (d) Any dependence between the realized (measured) income at date t and D is induced by the decision rule. It is the relationship between the expected outcomes at the time decisions are made and the realized outcomes that generate the structure of the bias for any econometric estimator of a model. This framework underlies much of the empirical work in the literature on evaluating job training programs (see, e.g., Ashenfelter, 1978, Bassi, 1983, 1984, and Ashenfelter and Card, 1985). We now consider various specializations of it.

6.3.1 Common Treatment Effect

As discussed in Section 3, the common treatment effect model is implicitly assumed in much of the literature evaluating job training programs. It assumes that $Y_{1t} - Y_{0t} = \alpha_t, t > k$, where α_t is a common constant for everyone. Another version writes α_t as a function of X , $\alpha_t(X)$. We take it as a point of departure for our analysis. The model we first presented was in Heckman (1978). Ashenfelter and Card (1985) and Heckman and Robb (1985a, 1986a) develop it. In this model, the effect of treatment on the treated and the effect of randomly assigning a person to treatment come to the same thing, *i.e.* $E(Y_{1t} - Y_{0t} | X, D = 1) = E(Y_{1t} - Y_{0t} | X)$ since the difference between the two income streams is the same for all persons with the same X characteristics. Under this model, decision rule (6.2) specializes to the discrete choice model

²⁸A sharp contrast between a model of perfect certainty and model of uncertainty is that the latter introduces the possibility of incorporating many more "explanatory variables" in the model in addition to the direct objects of the theory.

$$(6.3) \quad \begin{aligned} D &= 1, & \text{if } E \left(\sum_{j=1}^{T-k} \frac{\alpha_{k+j}}{(1+r)^j} - c - Y_{0k} \mid I_k \right) &\geq 0, \\ D &= 0 & \text{otherwise.} \end{aligned}$$

If the α_{k+j} are constant in all periods and T is large ($T \rightarrow \infty$) the criterion simplifies to

$$(6.4) \quad \begin{aligned} D &= 1 & \text{if } E \left(\frac{\alpha}{r} - c - Y_{0k} \mid I_k \right) &\geq 0, \\ D &= 0 & \text{otherwise.} \end{aligned}$$

Even though agents are assumed to be farsighted, and possess the ability to make accurate forecasts, the decision rule is simple. Persons compare current costs (both direct costs c and foregone earnings, Y_{0k}) with expected future rewards $E \left[\left(\sum_{j=1}^{T-k} \frac{\alpha_{k+j}}{(1+r)^j} \right) \mid I_k \right]$. Future rewards are the same for everyone of the same age and with the same discount rate. Future values of Y_{0t} , $t > k$, comes through the dependence with Y_{0k} and any dependence on cost c . If one knew, or could proxy, Y_{0k} and c , one could condition on these variables and eliminate selective differences between participants and nonparticipants. Since returns are identical across persons, only variation across persons in the direct cost and foregone earnings components determine the variation in the probability of program participation across persons. Assuming that c and Y_{0k} are unobserved by the econometrician, but known to the agent making the decision to go into training,

$$\Pr(D = 1) = \Pr \left(\sum_{j=1}^{T-k} \frac{\alpha_{k+j}}{(1+r)^j} > c + Y_{0k} \right).$$

In the case of an infinite-horizon, temporally-constant treatment effect, α , the expression simplifies to

$$\Pr(D = 1) = \Pr \left(\frac{\alpha}{r} \geq c + Y_{0k} \right).$$

This simple model is rich enough to be consistent with Ashenfelter's dip. As discussed in Section 4, the "dip" refers to the pattern that the earnings of program participants decline just prior to their participation in the program. If earnings are temporarily low in enrollment period k , and c does not offset Y_{0k} , persons with low earnings in the enrollment period enter the program. Since the return is the same for everyone, it is low opportunity costs or tuition that drive program participation in this model. If the α , c or Y_{0k} depend on observed characteristics, one can condition on those characteristics in constructing the probability of program participation.

This model is an instance of a more general approach to modelling behavior that is used in the economic evaluation literature. Write the net utility of program participation of the

decision maker as IN . An individual participates in the program ($D = 1$) if and only if $IN > 0$. Adopting a separable specification, we may write

$$IN = H(X) - V.$$

In terms of the previous example, $H(X) = \sum_{j=1}^{T-k} \frac{\alpha_{k+j}}{(1+r)^j}$ is a constant, and $V = c + Y_{0k}$.

The probability that $D = 1$ given X is

$$(6.5) \quad \Pr(D = 1 | X) = \Pr(V < H(X) | X).$$

If V is stochastically independent of X , we obtain the important special case

$$\Pr(D = 1 | X) = \Pr(V < H(X))$$

which is widely assumed in econometric studies of discrete choice.²⁹

If V is normal with mean μ_1 and variance σ_V^2 , then

$$(6.6) \quad \Pr(D = 1 | X) = \Pr(V < H(X)) = \Phi\left(\frac{H(X) - \mu_1}{\sigma_V}\right)$$

where Φ is the cumulative distribution function of a standard normal random variable. If V is a standardized logit,

$$\Pr(D = 1 | X) = \frac{\exp(H(X))}{1 + \exp(H(X))}.$$

Although these functional forms are traditional, they are restrictive and are not required by the econometric approach. Conditions for nonparametric identifiability of $\Pr(D = 1 | X)$ given different assumptions about the dependence of X and V are presented in Cosslett (1983), and Matzkin (1992). Cosslett (1983), Matzkin (1993) and Ichimura (1993) consider nonparametric estimation of H and the distribution of V . Lewbel (1998) demonstrates how discrete choice models can be identified under much weaker assumptions than independence between X and V . Under certain conditions, information about agent decisions to participate in a training program can be informative about their preferences and the outcomes of a program.

Heckman and Smith (1998a) demonstrate conditions under which knowledge of the self-selection decisions of agents embodied in $\Pr(D = 1 | X)$ is informative about the value of Y_1 relative to Y_0 . In the Roy model (see, e.g., Heckman and Honoré, 1990), $IN = Y_1 - Y_0 = (\mu_1(X) - \mu_0(X)) + (U_1 - U_0)$. Assuming X is independent of $U_1 - U_0$, from

²⁹Conditions for the existence of a discrete choice random utility representation of a choice process are given in McLennan (1990).

self selection decisions of persons into a program, it is possible to estimate $\mu_1(X) - \mu_0(X)$ up to scale, where the scale is $[Var(U_1 - U_0)]^{1/2}$. This is a standard result in discrete choice theory. Thus in the Roy model it is possible to recover $E(Y_1 - Y_0 | X)$ up to scale just from knowledge of the choice probability. Under additional assumptions on the support of X , Heckman and Smith (1998a) demonstrate that it is possible to recover the full joint distribution $F(y_0, y_1 | X)$ and to answer *all* of the evaluation questions about means and distributions posed in Section 3. Under more general self-selection rules, it is still possible to infer the personal valuations of a program from observing selection into the program and attrition from it. The Roy model is the one case where personal evaluations of a program, as revealed by the choice behavior of the agents studied, coincide with the “objective” evaluations based on $Y_1 - Y_0$.

Within the context of a choice-theoretic model, it is of interest to consider the assumptions that justify the three intuitive evaluation estimators introduced in section 4, starting with the cross-section estimator (3.3) - which is valid if assumption (4.A.3) is correct. Given decision rule (6.3), under what conditions is it plausible to assume that

$$(4.A.3) \quad E(Y_{0t} | D = 1) = E(Y_{0t} | D = 0), \quad t > k$$

so that cross section comparisons identify the true program effect? (Recall that in a model with homogeneous treatment impacts, the various mean treatment effects all come to the same thing.) We assume that evaluators do not observe costs nor do they observe Y_{0k} for trainees.

Assumption (4.A.3) would be satisfied in period t if

$$E(Y_{0t} | \sum_{j=1}^{T-k} \frac{\alpha_{k+j}}{(1+r)^j} - c - Y_{0k} \geq 0) = E(Y_{0t} | \sum_{j=1}^{T-k} \frac{\alpha_{k+j}}{(1+r)^j} - c - Y_{0k} < 0), \quad t > k.$$

One way this condition can be satisfied is if earnings are distributed independently over time (Y_{0k} independent of Y_{0t}), $t > k$, and direct costs c are independent of Y_{0t} , $t > k$. More generally, only independence in the means with respect to $c + Y_{0k}$ is required.³⁰ If the dependence in earnings vanishes for earnings measured more than ℓ periods apart (*e.g.* if earnings are a moving average of order ℓ), then for $t > k + \ell$, assumption (4.A.3) would be satisfied in such periods.

Considerable evidence indicates that earnings have an autoregressive component (see, *e.g.*, Ashenfelter 1978; Ashenfelter and Card, 1985; MaCurdy, 1982; Farber and Gibbons, 1994). Then (4.A.3) seems implausible except for special cases.³¹ Moreover if stipends (a component of c) are determined in part by current and past income because they are targeted toward low-income workers, then (4.A.3) is unlikely to be satisfied.

Access to better information sometimes makes it more likely that a version of assumption

³⁰Formally, it is required that $E(Y_{0t}|c + Y_{0k})$ does not depend on c and Y_{0k} for all $t > k$.

³¹Note, however, much of this evidence is for log earnings and not earnings levels.

(4.A.3) will be satisfied if it is revised to condition on observables X :

$$(4.A.3') \quad E(Y_{0t} \mid D = 1, X) = E(Y_{0t} \mid D = 0, X).$$

In this example, let $X = (c, Y_{0k})$. Then if we observe Y_{0k} for everyone, and can condition on it, and if c is independent of Y_{0t} given Y_{0k} , then

$$\begin{aligned} E(Y_{0t} \mid D = 1, Y_{0k}) &= E\left(Y_{0t} \mid \sum_{j=1}^{T-k} \frac{\alpha_{k+j}}{(1+r)^j} - Y_{0k} \geq c, Y_{0k}\right) \\ &= E(Y_{0t} \mid Y_{0k}) \\ &= E(Y_{0t} \mid D = 0, Y_{0k}). \end{aligned}$$

Then for common values of Y_{0k} , assumption (4.A.3') is satisfied for $X = Y_{0k}$.

Ironically, using too much information may make it difficult to satisfy (4.A.3'). To see this, suppose that we observe c and Y_{0k} and $X = (c, Y_{0k})$. Now

$$E(Y_{0t} \mid D = 1, (c, Y_{0k})) = E(Y_{0t} \mid c, Y_{0k})$$

and

$$E(Y_{0t} \mid D = 0, (c, Y_{0k})) = E(Y_{0t} \mid c, Y_{0k})$$

because c and Y_{0k} perfectly predict D . But (4.A.3') is *not* satisfied because decision rule (6.3) perfectly partitions the (c, Y_{0k}) space into disjoint sets. There are no common values of $X = (c, Y_{0k})$ such that (4.A.3') can be satisfied. In this case, the “regression discontinuity design” estimator of Campbell and Stanley (1966) is appropriate. We discuss this estimator in Section 7.4.6 below.

If we assume that

$$0 < Pr(D = 1 \mid X) < 1$$

we rule out the phenomenon of perfect predictability of D given X . This condition guarantees that persons with the same X values have a positive probability of being both participants and nonparticipants.³² Ironically, having too much information may be a bad thing. We need some “random” variation that places observationally equivalent people in both states. The existence of this fortuitous randomization lies at the heart of the method of matching.

Next consider assumption (4.A.1). It is satisfied in this example if in a time homogeneous environment, a “fixed effect” or “components of variance structure” characterizes Y_{0t} so that there is an invariant random variable φ such that Y_{0t} can be written as

$$(6.7) \quad Y_{0t} = \beta_t + \varphi + U_{0t} \quad \text{for all } t$$

$$\text{and} \quad E(U_{0t} \mid \varphi) = 0 \quad \text{for all } t$$

where the U_{0t} are mutually independent, and c is independent of U_{0t} . If Y_{0t} is earnings, then φ is “permanent income” and the U_{0t} are “transitory deviations” around it. Then using (6.3) for $t > k > t'$, we have

³²This is one of two conditions that Rosenbaum and Rubin (1983) call “strong ignorability” and is central to the validity of matching. We discuss these conditions further in section 7.3.

$$E(Y_{0t} - Y_{0t'} | D = 1) = \alpha_t + \beta_t - \beta_{t'},$$

since

$$E(U_{0t} | D = 1) - E(U_{0t'} | D = 1) = 0.$$

From the assumption of time homogeneity, $\beta_t = \beta_{t'}$. Thus assumption (4.A.1) is satisfied and the before-after estimator identifies α_t . It is clearly not necessary to assume that the U_{0t} are mutually independent, just that

$$(6.8) \quad E(U_{0t} - U_{0t'} | D = 1) = 0$$

i.e. that the innovation $U_{0t} - U_{0t'}$ is *mean* independent of $U_{0k} + c$. In terms of the economics of the model, it is required that participation does not depend on transitory innovations in earnings in periods t and t' . For decision model (6.3), this condition is satisfied as long as U_{0k} is independent of U_{0t} and $U_{0t'}$, or as long as $U_{0k} + c$ is mean independent of both terms.

If, however, the U_{0t} are serially correlated, then (4.A.1) will generally not be satisfied. Thus if a transitory decline in earnings persists over several time periods (as seems to be true as a consequence of Ashenfelter's dip), so that there is stochastic dependence of $(U_{0t}, U_{0t'})$ with U_{0k} , then it is unlikely that the key identifying assumption is satisfied. One special case where it is satisfied, developed by Heckman (1978) and Heckman and Robb (1985a) and applied by Ashenfelter and Card (1985) and Finifter (1987) among others, is a "symmetric differences" assumption. If t and t' are symmetrically aligned (so that $t = k + \ell$ and $t' = k - \ell$) and conditional expectations forward and backward are symmetric, so that

$$(6.9) \quad E(U_{0t} | c + \beta_t + U_{0k}) = E(U_{0t'} | c + \beta_k + U_{0k}),$$

then assumption (4.A.1) is satisfied. This identifying condition motivates the symmetric differences estimator discussed in Section 7.6.

Some evidence of non-stationary wage growth presented by Farber and Gibbons (1994), MaCurdy (1982), Topel and Ward (1992) and others suggests that earnings can be approximated by a "random walk" specification. If

$$(6.10) \quad Y_{0t} = \beta_t + \eta + \sum_{j=0}^t \nu_j,$$

where the ν_j are mean zero, mutually independent and identically-distributed random variables independent of η , then (6.8) and (6.9) will not generally be satisfied. Thus even if conditional expectations are linear, both forward and backward, it does not follow that (4.A.1) will hold. Let the variance of η and the variance of ν_j be finite. Assume that $E(\eta) = 0$. Suppose c is independent of all the ν_j and η , and

$$E(U_{0t} | c + \beta_t + U_{0k}) = \frac{\sigma_\eta^2 + k\sigma_\nu^2}{\sigma_c^2 + \sigma_\eta^2 + k\sigma_\nu^2} (c + U_{0k} - E(c))$$

and

$$E(U_{0t'} | c + \beta_t + U_{0k}) = \frac{\sigma_\eta^2 + t'\sigma_\nu^2}{\sigma_c^2 + \sigma_\eta^2 + t'\sigma_\nu^2} (c + U_{0k} - E(c)).$$

These two expressions are not equal unless $\sigma_\nu^2 = 0$.

A more general model that is consistent with the evidence reported in the literature writes

$$Y_{0t} = \mu_{0t}(X) + \eta + U_{0t},$$

where

$$U_{0t} = \sum_{j=1}^k \rho_{0j} U_{0,t-j} + \sum_{j=1}^m m_{0j} \nu_{t-j},$$

where the ν_{t-j} satisfy $E(\nu_{t-j}) = 0$ at all leads and lags, and are uncorrelated with η , where U_{0t} is an autoregression of order k and moving average of length m . Some authors like MaCurdy (1982) or Gibbons and Farber (1994) allow the coefficients (ρ_{0j}, m_{0j}) to depend on t and do not require that the innovations be identically distributed over time. For the logarithm of white male earnings in the United States, MaCurdy (1982) finds that a model with a permanent component (η), plus one autoregressive coefficient ($k = 1$) and two moving average terms ($m = 2$) describes his data.³³ Gibbons and Farber report similar evidence.

These times series models suggest generalizations of the before-after estimator that exploit the longitudinal structure of earnings processes but work with more general types of differences that align future and past earnings. These are developed at length in Heckman and Robb (1985, 1986), Heckman (1998a) and in Section 7.6.

If there are “time effects,” so that $\beta_t \neq \beta_{t'}$, (4.A.1) will not be satisfied. Before-after estimators will confound time effects with program gains. The “difference in differences” estimator circumvents this problem for models in which (4.A.1) is satisfied for the unobservables of the model but $\beta_t \neq \beta_{t'}$. Note, however, that in order to apply this assumption it is necessary that time effects be additive in some transformation of the dependent variable and identical across participants and nonparticipants. If they are not, then (4.A.2) will not be satisfied.

For example, if the decision rule for program participation is such that persons with lower life cycle wage growth paths are admitted into the program, or persons who are more vulnerable to the national economy are trained, then the assumption of common time (or age) effects across participants and nonparticipants will be inappropriate and the difference-in-difference estimator will not identify true program impacts.

³³The estimated value of ρ_{01} is close to 1 so that the model is close to a random walk in levels of log earnings.

6.3.2 A Separable Representation

In implementing econometric evaluation strategies, it is common to control for observed characteristics X . Invoking the separability assumption, we write the outcome equation for Y_{0t} as

$$Y_{0t} = g_{0t}(X) + U_{0t}$$

where g_{0t} is a behavioral relationship and U_{0t} has a finite mean conditioning on X . A parallel expression can be written for Y_{1t} :

$$Y_{1t} = g_{1t}(X) + U_{1t}.$$

The expression for $g_{0t}(X)$ is a structural relationship that may or may not be different from $\mu_{0t}(X)$, the conditional mean. It is a ceteris paribus relationship that informs us of the effect of changes of X on Y_{0t} holding U_{0t} constant. Throughout this chapter we distinguish μ_{1t} from g_{1t} and μ_{0t} from g_{0t} . For the latter, we allow for the possibility that $E(U_{1t} | X) \neq 0$ and $E(U_{0t} | X) \neq 0$. The separability enables us to isolate the effect of self selection, as it operates through the “error term”, from the structural outcome equation:

$$(6.11a) \quad E(Y_{0t} | D = 0, X) = g_{0t}(X) + E(U_{0t} | D = 0, X).$$

$$(6.11b) \quad E(Y_{1t} | D = 1, X) = g_{1t}(X) + E(U_{1t} | D = 1, X).$$

The $g_{0t}(X)$ and $g_{1t}(X)$ functions are invariant across different conditioning schemes and decision rules provided that X is available to the analyst. One can borrow knowledge of these functions from other studies collected under different conditioning rules including the conditioning rules that define the samples used in social experiments. Although the conditional mean of the errors differs across studies, the $g_{0t}(X)$ and analogous $g_{1t}(X)$ functions are invariant across studies. If they can be identified, they can be meaningfully compared across studies, unlike the parameter treatment on the treated which, in the case of heterogeneous response to treatment that is acted on by agents, differs across programs with different decision rules and different participant compositions.

A special case of this representation is the basis for an entire literature. Suppose that

(P.1) The random utility representation (6.5) is valid.

Further, suppose that

(P.2) $(U_{0t}, U_{1t}, V) \perp\!\!\!\perp X$,

(“ $\perp\!\!\!\perp$ ” denotes stochastic independence)

and finally assume that

(P.3) the distribution of V , $F(V)$ is strictly increasing in V .

Then

$$(6.12a) \quad E(U_{0t} | D = 1, X) = K_{0t}(\Pr(D = 1 | X)).$$

and

$$(6.12b) \quad E(U_{1t} | D = 1, X) = K_{1t}(\Pr(D = 1 | X)).^{34}$$

The mean error term is a function of P , the probability of participation in the program. This special case receives empirical support in Heckman, Ichimura, Smith and Todd (1998) and Heckman, Ichimura and Todd (1997). It enables analysts to characterize the dependence between U_{0t} and X by the dependence of U_{0t} on $\Pr(D = 1 | X)$ which is a scalar function of X . As a practical matter, this greatly reduces the empirical task of estimating selection models. Instead of having to explore all possible dependence relationships between U and X , the analyst can confine attention to the more manageable task of exploring the dependence between U and $\Pr(D = 1 | X)$. An investigation of the effect of conditioning on program eligibility rules or self selection on Y_{0t} comes down to an investigation of the effect of the conditioning on Y_{0t} as it operates through the probability P . It motivates a focus on the determinants of participation in the program in order to understand selection bias and is the basis for the “control function” estimators developed in Section 7.

³⁴The proof is immediate. The proof of (6.12b) follows by similar reasoning. We follow Heckman (1980) and Heckman and Robb (1985a, 1986b). Assume that U_{0t}, V are jointly continuous random variables, with density $f(U_{0t}, V | X)$. From (P.2)

$$f(U_{0t}, V | X) = f(U_{0t}, V).$$

Thus

$$E(U_{0t} | X, D = 1) = \frac{\int_{-\infty}^{\infty} U_{0t} \int_{-\infty}^{H(X)} f(U_{0t}, V) dU_{0t} dV}{\int_{-\infty}^{H(X)} f(V) dV}.$$

Now

$$\Pr(D = 1 | X) = \int_{-\infty}^{H(X)} f(V) dV.$$

Inverting, we obtain

$$H(X) = F_V^{-1}(\Pr(D = 1 | X)).$$

Thus

$$E(U_{0t} | X, D = 1) = \frac{\int_{-\infty}^{\infty} U_{0t} \int_{-\infty}^{F_V^{-1}(\Pr(D=1|X))} f(U_{0t}, V) dV dU_{0t}}{\Pr(D = 1 | X)}$$

$$\stackrel{def}{=} K_{0t}(\Pr(D = 1 | X)).$$

If, however, (P.2) is not satisfied, then the separable representation is not valid. Then it is necessary to know more than the probability of participation to characterize $E(U_{0t} | X, D = 1)$. In this case it is necessary to characterize both the dependence between U_{0t} and X given $D = 1$ and the probability of participation.

6.3.3 Variable Treatment Effect

A more general version of the decision rule, given by (6.2), allows (Y_{0t}, Y_{1t}) to be a pair of random variables with no necessary restriction connecting them. In the more general case,

$$\alpha_t = Y_{1t} - Y_{0t}, \quad t > k$$

is now a random variable. In this case as previously discussed in Section 3, there is a distinction between the parameter “the mean effect of treatment on the treated” and the “mean effect of randomly assigning a person with characteristics X into the program”.

In one important case discussed in Heckman and Robb (1985a), the two parameters have the same ex post mean value even if treatment effect α_t is heterogeneous after conditioning on X . Suppose that α_t is unknown to the agent at the time enrollment decisions are made. The agent forecasts α_t using the information available in his/her information set I_k . $E(\alpha_t | I_k)$ is the private expectation of gain by the agent. If ex post gains of participants with characteristics X are the same as what the ex post gains of nonparticipants would have been had they participated, then the two parameters are the same. This would arise if both participants and nonparticipants have the same ex ante expected gains

$$E(\alpha_t | D = 1, I_k) = E(\alpha_t | D = 0, I_k) = E(\alpha_t | I_k),$$

and if

$$E[E(\alpha_t | I_k) | X, D = 1] = E[E(\alpha_t | I_k) | X, D = 0],$$

where the expectations are computed with respect to the observed ex-post distribution of the X . This condition requires that the information in the participant’s decision set has the same relationship to X as it has for nonparticipants. The interior expectations in the preceding expression are subjective. The exterior expectations in the expression are computed with respect to distributions of objectively-observed characteristics. The condition for the two parameters to be the same is

$$E[E(\alpha_t | I_k, D = 1) | X, D = 1] = E[E(\alpha_t | I_k, D = 0) | X, D = 0].$$

As long as the ex-post objective expectation of the subjective expectations is the same, the two parameters ($E(\alpha_t | X, D = 1)$ and $E(\alpha_t(X))$) are the same. This condition would be satisfied if, for example, all agents, irrespective of their X values, place themselves at the mean of the objective distribution, i.e.,

$$E(\alpha_t | I_k, D = 1) = E(\alpha_t | I_k, D = 0) = \bar{\alpha}_t$$

(see, e.g., Heckman and Robb, 1985a). Differences across persons in program participation are generated by factors other than potential outcomes. In this case, the ex-post surprise,

$$(\alpha_t - \bar{\alpha}_t)$$

does not depend on X or D in the sense that

$$E(\alpha_t - \bar{\alpha}_t | X, D = 1) = 0.$$

So

$$E(Y_{1t} - Y_{0t} | X, D = 1) = \bar{\alpha}_t.$$

This discussion demonstrates the importance of understanding the decision rule and its relationship to measured outcomes in formulating an evaluation model. If agents do not make their decisions based on the unobserved components of gains from the program or on variables statistically related to those components, the analysis for the common coefficient model presented in section (a) remains valid even if there is variability in $U_{1t} - U_{0t}$. If agents anticipate the gains, and base decisions on them, at least in part, then a different analysis is required.

The conditions for the absence of bias for one parameter are different from the conditions for the absence of bias for another parameter. The difference between the “random assignment” parameter $E(Y_{1t} - Y_{0t} | X)$ and the “treatment on the treated” parameter is gain in the unobservables going from one state to the next:

$$E(U_{1t} - U_{0t} | X, D = 1) = E(\Delta_t | X, D = 1) - E(\Delta_t | X).$$

The only way to avoid bias for *both* mean parameters is if $E(U_{1t} - U_{0t} | X, D = 1) = 0$.

Unlike the other estimators, the before-after estimators are non-robust to time effects that are common across participants and nonparticipants. The difference-in-differences estimators and the cross-section estimators are unbiased under different conditions. The cross-section estimator for the period t common effect and the “treatment on the treated”

variable-effect version of the model require that mean unobservables in the no program state be the same for participants and nonparticipants. The difference-in-differences estimator requires a *balance of the bias* in the *change* in the unobservables from period t' to period t . If the cross-section conditions for the absence of bias are satisfied for all t , then the assumption justifying the difference-in-differences estimator is satisfied.

However, the converse is not true. Even if the conditions for the absence of bias in the difference-in-differences estimator are satisfied, the conditions for absence of bias for the cross section estimator are not necessarily satisfied. Moreover, failure of the difference-in-differences condition for the absence of bias does not imply failure of the condition for absence of bias for the cross-section estimator. Ashenfelter's dip provides empirically relevant example of this point. If t' is measured during the period of the dip, but the dip is mean-reverting in post-program periods, then the condition for the absence of cross-section bias could be satisfied because post-program, there could be no selective differences among participants.

6.3.4 Imperfect Credit Markets

How robust is the analysis of Sections 6.2 and 6.3, and in particular the conditions for bias, to alternative specifications of decision rules and the economic environments in which individuals operate? To answer this question, we first reexamine the decision rule after dropping our assumption of perfect credit markets. There are many ways to model imperfect credit markets. The most extreme approach assumes that persons consume their earnings each period. This changes the decision rule (6.2) and produces a new interpretation for the conditions for absence of bias. Let G denote a time-separable strictly concave utility function and let β be a subjective discount rate. Suppose that persons have exogenous income flow η_t per period. Expected utility maximization given information I_k produces the following program participation rule:

$$(6.13) \quad D = \begin{cases} 1 & \text{if } E \left[\sum_{j=1}^{T-k} \beta^j \{G(Y_{1,k+j} + \eta_{k+j}) - G(Y_{0,k+j} + \eta_{k+j})\} \right. \\ & \left. + G(\eta_k - c_k) - G(Y_{0k} + \eta_k) \mid I_k \right] \geq 0; \\ 0 & \text{otherwise.} \end{cases}$$

As in the previous cases, earnings prior to time period k are only relevant for forecasting future earnings (i.e., as elements of I_k). However, the decision rule (6.2) is fundamentally altered in this case. Future earnings in both states determine participation in a different way. Common components of earnings in the two states do not difference out unless G is a linear function.³⁵

³⁵Due to the nonlinearity of G , there are wealth effects in the decision to take training.

Consider the permanent-transitory model of equation (6.7). That model is favorable to the application of longitudinal before-after estimators. Suppose that the U_{0t} are independent and identically distributed, and there is a common-effect model. Condition (6.8) is not satisfied in a perfect foresight environment when there are credit constraints, or in an environment in which the U_{0t} can be partially forecast³⁶ because for $t > k > t'$

$$E(U_{0t} | X, D = 1) \neq 0$$

even though

$$E(U_{0t'} | X, D = 1) = 0,$$

so

$$E(U_{0t} - U_{0t'} | X, D = 1) \neq 0.$$

The before-after estimator is now biased. So is the difference in differences estimator. If, however, the U_{0t} are not known, and cannot be partially forecast, then condition (6.8) is valid, so both the before-after and difference in difference estimators are unbiased.

Even in a common effect model, with Y_{0t} (or U_{0t}) independently and identically distributed, the cross section estimator is biased for period $t > k$ in an environment of perfect certainty with credit constraints because D depends on Y_{0t} through decision rule (6.13). On the other hand, if Y_{0t} is not forecastable with respect to the information in I_k , the cross-section estimator is unbiased.

The analysis in this subsection and the previous subsections has major implications for a certain style of evaluation research. Understanding the stochastic model of the outcome process is not enough. It is also necessary to know how the decision makers process the information, and make decisions about program participation.

6.3.5 Training As A Form of Job Search

Heckman and Smith (1998b) find that among persons eligible for the JTPA program, the unemployed are much more likely to enter the program than are other eligible persons. Persons are defined to be unemployed if they are not working but report themselves as actively seeking work. The relationship uncovered by Heckman and Smith is not due to eligibility requirements. In the United States, unemployment is not a precondition for participation in the program.

Several previous studies suggest that Ashenfelter's dip results from changes in labor force status, instead of from declines in wages or hours among those who work. Using

³⁶ "Partially forecastable" means that some component of U_{0t} resides in the information set I_k . That is, letting $f(y | x)$ be the density of y given x , $f(U_{0t} | I_k) \neq f(U_{0t})$ so that I_k predicts U_{0t} in this sense. One could define "moment forecastability" using conditional expectations of certain moments of function " φ ". If $E(\varphi(U_{0t}) | I_k) \neq E(\varphi(U_{0t}))$, then $\varphi(U_{0t})$ is partially moment forecastable using the information in I_k . More formally, a random variable is fully-forecastable if the σ -algebra generating U_{0t} is contained in the σ -algebra of I_k . It is partially forecastable if the complement of the projection of the σ -algebra of U_{0t} onto the σ -algebra of I_k is not the empty set. It is fully unforecastable if the projection of the σ -algebra of U_{0t} onto the σ -algebra of I_k is the empty set.

even a crude measure of employment rates, namely whether a person was employed at all during a calendar year, Card and Sullivan (1988) observed that U.S. CETA training participants' employment rates declined prior to entering training.³⁷ Their evidence suggests that changes in labor force dynamics instead of changes in earnings may be a more precise way to characterize participation in training.

Heckman and Smith (1998b) show that whether or not a person is employed, unemployed (not employed and looking for work), or out of the labor force is a powerful predictor of participation in training programs. Moreover, they find that recent changes in labor force status are important determinants of participation for all demographic groups. In particular, eligible persons who have just become unemployed, either through job loss or through re-entry into the labor force, have the highest probabilities of participation. For women, divorce, another form of job termination, is a predictor of who goes into training. Among those who either are employed or out of the labor force, persons who have recently entered these states have much higher participation program probabilities than persons in those states for some time. Their evidence is formalized by the model presented in this section.

The previous models that we have considered are formulated in terms of *levels* of costs and earnings. When opportunity costs are low, or tuition costs are low, the persons are more likely to enter training. The model presented here recognizes that *changes* in labor force states account for participation in training. Low earnings levels are a subsidiary predictor of program participation that are overshadowed in empirical importance by unemployment dynamics in the analyses of Heckman and Smith (1998b).

Persons with zero earnings differ substantially in their participation probabilities depending on their recent labor force status histories. Yet, in models based on pre-training earnings dynamics, such as the one presented in Section 6.3, such persons are assumed to have the same behavior irrespective of their labor market histories.

The importance of labor force status histories also is not surprising given that many employment and training services, such as job search assistance, on-the-job training at private firms, and direct placement are all designed to lead to immediate employment. By providing these services, these programs function as a form of job search for many participants. Recognizing this role of active labor market policies is an important development in recent research. It indicates that in many cases, participation in active labor market programs should not be modeled as if it were like a schooling decision, such as we have modeled it in the preceding sections.

In this section, we summarize the evidence on the determinants of participation in the program and construct a simple economic model in which job search makes two contribu-

³⁷Ham and LaLonde (1990) report the same result using semi-monthly employment rates for adult women participating in NSW.

tions to labor market prospects: (a) it facilitates the rate of arrival of job offers and (b) it improves the distribution of wages in the sense of giving agents a stochastically dominant wage distribution compared to the one they face without search. Training is one form of unemployment that facilitates job search. Different training options will produce different job prospects characterized by different wage and layoff distributions. Searchers might participate in programs that subsidize the rate of arrival of job offers (JSA as described in Section 2), or that improve the distribution from which wage offers are drawn (i.e., basic educational and training investments).

Instead of motivating participation in training with a standard human capital model, we motivate participation as a form of search among options. Because JSA constitutes a large component of active labor market policy, it is of interest to see how the decision rule is altered if enhanced job search rather than human capital accumulation is the main factor motivating individuals' participation in these programs.

Our model is based on the idea that in program j , wage offers arrive from a distribution F_j at rate λ_j . Persons pay c_j to sample from F_j . (The costs can be negative). Assume that the arrival times are statistically independent of the wage offers and that arrival times and wage offers from one search option are independent of the wages and arrival times of other search options. At any point in time, persons pick the search option with the highest expected return. To simplify the analysis, suppose that all distributions are time invariant and denote by N the value of nonmarket time. Persons can select among any of J options, denoted by j . Associated with each option is a rate at which jobs appear, λ_j . Let the discount rate be r . These parameters may vary among persons but for simplicity we assume that they are constant for the same person over time. This heterogeneity among persons produces differences among choices in training options, and differences in the decision to undertake training.

In the unemployed state, a person receives a nonmarket benefit, N . The choice between search from any of the training and job search options can be written in "Gittens Index" form. (See, e.g., Berry and Fristedt, 1986). Under our assumptions, being in the nonmarket state has constant per-period value N irrespective of the search option selected. Letting V_{je} be the value of employment arising from search option j , the value of being unemployed under training option j is:

$$(6.14a) \quad V_{ju} = N - c_j + \frac{\lambda_j}{1+r} E_j \max[V_{je}; V_{ju}] + \frac{(1-\lambda_j)}{1+r} V_{ju}.$$

The first term, $(N - c_j)$, is the value of nonmarket time minus the j -specific cost of search. The second term is the discounted product of the probability that an offer arrives next period if the j^{th} option is used, and the expected value of the maximum of the two options: work (valued at V_{je}) or unemployment V_{ju} . The third term is the probability that the person will continue to search times the value of doing so. In a stationary environment, if

it is optimal to search from j today, it is optimal to do so tomorrow.

Let σ_{je} be the exogenous rate at which jobs disappear. For a job holder, the value of employment is V_{je} :

$$(6.14b) \quad V_{je} = Y_j + \frac{(1 - \sigma_{je})}{1 + r} V_{je} + \frac{\sigma_{je}}{1 + r} E_j[\max(V_N, V_{ju})].$$

V_{ju} is the value of optimal job search under j . The expression consists of the current flow of earnings (Y_j) plus the discounted ($\frac{1}{1+r}$) expected value of employment (V_{je}) times the probability that the job is retained ($1 - \sigma_{je}$). The third term arises from the possibility that a person loses his/her job (this happens with probability (σ_{je})) times the expected value of the maximum of the search and nonmarket value options (V_N).

To simplify this expression, assume that $V_{ju} > V_N$. If this is not so, the person would never search under any training option under any event. In this case, V_{je} simplifies to

$$V_{je} = Y_j + \frac{(1 - \sigma_{je})}{1 + r} V_{je} + \frac{\sigma_{je}}{1 + r} V_{ju}$$

so

$$(6.14c) \quad V_{je} = \frac{\sigma_{je}}{r + \sigma_{je}} V_{ju} + \frac{(1 + r)Y_j}{r + \sigma_{je}}.$$

Substituting (6.14c) into (6.14a), we obtain, after some rearrangement,

$$V_{ju} = \frac{(1 + r)(N - c_j) + \lambda_j E_j (V_{je} | V_{je} > V_{ju}) \Pr(Y_j > V_{ju}(r/1 + r))}{r + \lambda_j \Pr(Y_j > V_{ju}(r/1 + r))}.$$

In deriving this expression, we assume that the environment is stationary so that the optimal policy at time t is also the optimal policy at t' provided that the state variables are the same in each period.

The optimal search strategy is

$$\hat{j} = \arg \max_j \{V_{ju}\}$$

provided that $V_{ju} > V_N$ for at least one j . The lower c_j and the higher λ_j , the more attractive is option j . The larger the F_j —in the sense that j stochastically dominates j' ($F_j(x) < F_{j'}(x)$), so more of the mass of F_j is the upper portion of the distribution—the more attractive is option j . Given the search options available to individuals, enrollment in a job training program may be the most effective option.

The probability that a training from option j lasts $T_j = t_j$ periods or more is

$$\Pr(T_j \geq t_j) = [1 - \lambda_j(1 - F_j(V_{ju}(r/(1 + r))))]^{t_j}$$

where $1 - \lambda_j(1 - F_j(V_{ju}(r/1 + r)))$ is the sum of the probability of receiving no offer ($1 - \lambda_j$) plus the probability of receiving an offer that is not acceptable ($\lambda_j F_j(V_{ju}(r/1 + r))$). This model is nonlinear in the basic parameters. Because of this nonlinearity, many estimators relying on additive separability of the unobservables, such as difference-in-differences or the fixed effect schemes for eliminating unobservables, are ineffective evaluation estimators.

This simple model summarizes the available empirical evidence on job training programs. (a) It rationalizes variability in the length of time persons with identical characteristics spend in training. Persons receive different wage offers at different times and leave the program to accept the wage offers at different dates. (b) It captures the notion that training programs might facilitate the rate of job arrivals - the λ_j (this is an essential function of “job search assistance” programs) or they might produce skills - by improving the F'_j - or both. (c) It accounts for why there might be recidivism back into training programs. As jobs are terminated (at rate σ_{je}), persons re-enter the program to search for a replacement job. Recidivism is an important feature of major job training programs. Trott and Baj (1993) estimate that as many as 20 percent of all JTPA program participants in Northern Illinois have been in the program at least twice with the modal number being three. This has important implications for the contamination bias problem that we discuss in Section 7.7.

A less attractive feature of the model is that persons do not switch search strategies. This is a consequence of the assumed stationarity of the environment and the assumption that agents know both arrival rates and wage offer distributions. Relaxing the stationarity assumption produces switching among strategies which seems to be consistent with the evidence. A more general - but less analytically tractable model - allows for learning about wage offer distributions as in Weitzman (1979). In such a model, persons may switch strategies as they learn about the arrival rates or the wage offers obtained under a given strategy. The learning can take place within each type of program and may also entail word of mouth learning from fellow trainees taking the option.

Weitzman’s model captures this idea in a very simple way and falls within the Gitten’s index framework. The basic idea is as follows. Persons have J search options. They pick the option with the highest value and take a draw from it. They accept the draw if the value of the realized draw is better than the expected value of the best remaining option. Otherwise they try out the latter option. If the draws from the J options are independently distributed, a Gittens-index strategy describes this policy. In this framework, unemployed persons may try a variety of options - including job training - before they take a job, or drop out of the labor force.

One could also extend this model to allow the value of non-market time, N , to become stochastic. If N fluctuates, persons would enter or exit the labor force depending on the value of N . Adding this feature captures the employment dynamics of trainees described

by Card and Sullivan (1988).

In this more general model, shocks to the value of leisure or termination of previous jobs make persons contemplate taking training. Whether or not they do so depends on the value of training compared to the value of other strategies for finding jobs. Allowing for these considerations produces a model broadly consistent with the evidence presented in Heckman and Smith (1998b) that persons enter training as a consequence of displacement from both the market and nonmarket sector.

The full details of this model remain to be developed (see Heckman and Smith, 1999, for a start). We suggest that future analyses of program participation be based on this empirically more concordant model. For the rest of this chapter, however, we take decision rule (6.3) as canonical in order to motivate and justify the choice of alternative econometric estimators. We urge our readers to modify our analysis to incorporate the lessons from this framework of labor force dynamics sketched here.

6.4 The Role of Program Eligibility Rules In Determining Participation

Several institutional features of most training programs suggest that the participation rule is more complex than that characterized by the simple model presented above in Section 6.2. For example, eligibility for training is often based on a set of objective criteria, such as current or past earnings being below some threshold. In this instance, individuals can take training at time k only if they have had low earnings, regardless of its potential benefit to them. For example, enrollees satisfy

$$(6.15) \quad \alpha/r - Y_{ik} - c_i > 0 \text{ and the eligibility rules } Y_{i,k-1} < K$$

where K is a cutoff level. More general eligibility rules can be analyzed in the same framework.

The universality of Ashenfelter's dip in pre-program earnings among program participants occurs despite the substantial variation in eligibility rules among training programs. This suggests that earnings or employment dynamics drive the participation process and that Ashenfelter's dip is not an artifact of eligibility rules. Few major training programs in the United States have required earnings declines to qualify for program eligibility. Certain CETA programs in the late 1970s required participants to be unemployed during the period just prior to enrollment, while NSW required participants to be unemployed at the date of enrollment. MDTA contained no eligibility requirements, but restricted training stipends to persons who were unemployed or "underemployed."³⁸ For the JTPA program, eligibility

³⁸Eligibility for CETA varied by subprogram. CETA's controversial Public Sector Employment (PSE) program required participants to have experienced a minimum number of days of unemployment or "un-

has been confined to the economically disadvantaged (defined by low family income over the past six months, participation in a cash welfare program or Food Stamps or being a foster child or disabled). There is also a 10 percent “audit window” of eligibility for persons facing other unspecified “barriers to employment.”

It is possible that Ashenfelter’s dip results simply from a mechanical operation of program eligibility rules that condition on recent earnings. Such rules select individuals with particular types of earnings patterns into the eligible population. To illustrate this point, consider the monthly earnings of adult males who were eligible for JTPA in a given month from the 1986 panel of the U.S. Survey of Income and Program Participation (SIPP). For most people, eligibility is determined by family earnings over the past six months. The mean monthly earnings of adult males appear in Figure 4.1 aligned relative to month ‘ k ,’ the month when eligibility is measured. The figure reveals a dip in the mean earnings of adult male *eligibles* centered in the middle of the six month window over which family income is measured when determining JTPA eligibility.

Figure 4.1 also displays the mean earnings of adult males in the experimental control group from the NJS.³⁹ The earnings dip for the controls, who applied and were admitted in the program, is larger than for the sample of JTPA eligibles from the SIPP. Moreover, this dip reaches its minimum during month ‘ k ’ rather than three or four months before as would be indicated by the operation of eligibility rules. The substantial difference between the mean earnings patterns of JTPA participants and eligibles implies that Ashenfelter’s dip does not result from the mechanical operation of program eligibility rules.⁴⁰

6.5 Administrative Discretion and the Efficiency and Equity of Training Provision

Training participation also often depends on discretionary choices made by program operators. Recent research focuses on how program operators allocate training services among

deremployment” just prior to enrollment. In general, persons became eligible for other CETA programs by having a low income or limited ability in English. Considerable discretion was left to the states and training centers to determine who enrolled in the program. By contrast, the NSW eligibility requirements were quite specific. Adult women had to be on AFDC at the time of enrollment, have received AFDC for 30 of the last 36 months, and have a youngest child age six years or older. Youth in the NSW had to be age 17-20 years with no high school diploma or equivalency degree and have not been in school in the past six months. In addition, fifty percent of youth participants had to have had some contact with the criminal justice system (Hollister, et al., 1984).

³⁹Such data were collected at four of the 16 training centers that participated in the study.

⁴⁰Devine and Heckman (1996) present certain nonstationary family income processes that can generate Ashenfelter’s dip from the application of JTPA eligibility rules. However, in their empirical work they find a dip centered at $k - 3$ or $k - 4$ for adult men and adult women, but no dip for male and female youth.

groups and on how administrative performance standards affect the allocation of these services. The main question that arises in these studies is the potential trade-off between equity and efficiency, and the potential conflict between social objectives and program operators' incentives. An efficiency criterion that seeks to maximize the social return to public training investments, regardless of the implications for income distribution, implies focusing training resources on those groups for whom the impact is largest (per dollar spent). In contrast, equity and redistributive criteria dictate focusing training resources on groups who are most in "need" of services .

These goals of efficiency and equity are written into the U.S. Job Training Partnership Act.⁴¹ Whether or not these twin goals conflict with each other depends on the empirical relationship between initial skill levels and the impact of training. As we discuss in below Section 10, the impact of training appears to vary on the basis of observable characteristics, such as sex, age, race and what practitioners call "barriers to employment" – low schooling, lack of employment experience and so on. These twin goals would be in conflict if the largest social returns resulted from training the most job ready applicants.

In recent years, especially in the United States, policymakers have used administrative performance standards to assess the success of program operators in different training sites. Under JTPA, these standards are based primarily on average employment rates and average wage rates of trainees shortly after they leave training. The target levels for each site are adjusted based on a regression model that attempts to hold constant features of the environment over which the local training site has no control, such as racial composition.⁴² Sites whose performance exceeds these standards may be rewarded with additional funding; those that fall below may be sanctioned. The use of such performance standards, instead of measures of the impact of training, raises the issue of "cream-skimming" by program operators (Bassi, 1984). Program staff concerned solely with their site's performance relative to the standard should admit into the program applicants who are likely to be employed at good wages (the "cream") regardless of whether or not they benefit from the program. By contrast, they should avoid applicants who are less likely to be employed after leaving training or have low expected wages, even if the impact of the training for such persons is likely to be large. The implications of cream-skimming for equity are clear. If it exists, program operators are directing resources away from those most in need. However, its im-

⁴¹A related issue involves differences in the types of services provided to different groups conditional on participation in a program. The U.S. General Accounting Office (1991) finds such differences alarming in the JTPA program. Smith (1992) argues that they result from differences across groups in readiness for immediate employment and in the availability of income support during classroom training.

⁴²See Heckman and Smith (1997d) and the essays in Heckman (1998b) for more detailed descriptions of the JTPA performance standards system. Similar systems based on the JTPA system now form a part of most U.S. training programs.

plications for efficiency depend on the empirical relationship between short-term outcome levels and long-term impacts. If applicants who are likely to be subsequently employed also are those who benefit the most from the program, performance standards indirectly encourage the efficient provision of training services.⁴³

A small literature examines the empirical importance of cream-skimming in JTPA programs. Anderson, et al. (1991) and Anderson, et al.(1993) look for evidence of cream-skimming by comparing the observable characteristics of JTPA participants and individuals eligible for JTPA. They report evidence of cream-skimming defined in their study as the case in which individuals with fewer barriers to employment have differentially higher probabilities of participating in training. However, this finding may result not from cream-skimming by JTPA staff, but because among those in the JTPA eligible population, more employable persons self-select into training.⁴⁴

Two more recent studies address this problem. Using data from the NJS, Heckman and Smith (1998e) decompose the process of participation in JTPA into a series of stages. They find that much of what appears to be cream-skimming in simple comparisons between participants' and eligibles' characteristics is self-selection. For example, high school dropouts are very unlikely to be aware of JTPA and as a result are unlikely ever to apply. To assess the role of cream-skimming, Heckman, Smith and Taber (1996) study a sample of applicants from one of the NJS training centers. They find that program staff at this training center do not cream-skin, and appear instead to favor the hard-to-serve when deciding whom to admit into the program. Such evidence suggests that cream-skimming may not be of major empirical importance, perhaps because the social service orientation of JTPA staff moderates the incentives provided by the performance standards system, or because of local political incentives to serve more disadvantaged groups. For programs in Norway, Aakvik (1998) finds strong evidence of negative selection of participants on outcomes. Heinrich (1998) reports just the opposite for a job training program in the United States. At this stage no universal generalization about bureaucratic behavior regarding cream skimming is possible.

Studies based on the NJS also provide evidence on the implications of cream-skimming, even if it were to exist. Heckman, Smith and Clements (1997) find that except for those who are very unlikely to be employed, the impact of training does not vary with the expected levels of employment or earnings in the absence of training. This finding indicates that the impact on efficiency of cream-skimming (or alternatively the efficiency cost of serving

⁴³Heckman and Smith (1997d) discuss this issue in greater depth. The discussion in the text presumes that the costs of training provided to different groups are roughly equal.

⁴⁴Program staff often have some control over who applies through their decisions about where and how much to publicize the program. However, this control is much less important than their ability to select among program applicants.

the hard-to-serve) is low. Similarly, (1998d) find little empirical relationship between the outcome measures used in the JTPA performance standards system and experimental estimates of the impact of JTPA training. These findings suggest that cream-skimming has little impact on efficiency, and that administrative performance standards, to the extent that they affect who is served, do little to increase either the efficiency or equity of training provision.

6.6 The Conflict Between The Economic Approach to Program Evaluation And The Modern Approach to Social Experiments

We have already noted in Section 5 that under ideal conditions, social experiments identify $E(Y_1 - Y_0|X, D = 1)$. Without further assumptions and econometric manipulation, they do not answer the other evaluation questions posed in Section 3. As a consequence of the self-selected nature of the samples generated by social experiments, the data produced from them are far from ideal for estimating the structural parameters of behavioral models. This makes it difficult to generalize findings across experiments or to use experiments to identify the policy-invariant structural parameters that are required for econometric policy evaluation.

To see this, recall that social experiments balance bias, but they do not eliminate the dependence between U_0 and D or U_1 and D . Thus from the experiments conducted under ideal conditions, we can recover the conditional densities $f(y_0|X, D = 1)$ and $f(y_1|X, D = 1)$. From nonparticipants we can recover $f(y_0|X, D = 0)$. It is the density $f(y_0 | X, D = 1)$ that is the new information produced from social experiments. The other densities are available from observational data. All of these densities condition on choices. Knowledge of the conditional means

$$E(Y_0|X, D = 1) = g_0(X) + E(U_0|X, D = 1)$$

and

$$E(Y_1|X, D = 1) = g_1(X) + E(U_1|X, D = 1)$$

does not allow us to separately identify the structure $(g_0(X), g_1(X))$ from the conditional error terms without invoking the usual assumptions made in the nonexperimental selection literature. Moreover, the error processes for U_0 and U_1 conditional on $D = 1$ are fundamentally different than those in the population at large if participation in the program depends, in part, on U_0 and U_1 .

For these reasons, evidence from social experiments on programs with different participation and eligibility rules do not cumulate in any interpretable way. The estimated treatment effects reported from the experiments combine structure and error in different ways, and the conditional means of the outcomes bear no simple relationship to $g_0(X)$ or $g_1(X)$ ($X\beta_0$ and $X\beta_1$ in a linear regression setting). Thus it is not possible, without conducting a nonexperimental selection study, to relate the conditional means or regression functions obtained from a social experiment to a core set of policy-invariant structural parameters. Ham and LaLonde (1996) present one of the few attempts to recover structural parameters from a randomized experiment, where randomization was administered at the stage where persons applied and were accepted into the program. The complexity of their analysis is revealing about the difficulty of recovering structural parameters from social experiments.

In bypassing the need to specify economic models, many recent social experiments produce evidence that is not informative about them. They generate choice-based, endogenously stratified samples that are difficult to use in addressing any other economic question apart from the narrow question of determining the impact of treatment on the treated for one program with one set of participation and eligibility rules.