Chapter 31

THE ECONOMICS AND ECONOMETRICS OF ACTIVE LABOR MARKET PROGRAMS

JAMES J. HECKMAN*

University of Chicago

ROBERT J. LALONDE*

University of Chicago

JEFFREY A. SMITH*

University of Western Ontario

Contents

* We thank Susanne Ackum Agell for her helpful comments on Scandanavian active labor market programs and Costas Meghir for very helpful comments on Sections 1–7 and Wilbert van der Klaauw for comments on Section 7.4.6.

Handbook of Labor Economics, Volume 3, Edited by A. Ashenfelter and D. Card

Abstract

Policy makers view public sector-sponsored employment and training programs and other active labor market policies as tools for integrating the unemployed and economically disadvantaged into the work force. Few public sector programs have received such intensive scrutiny, and been subjected to so many different evaluation strategies. This chapter examines the impacts of active labor market policies, such as job training, job search assistance, and job subsidies, and the methods used to evaluate their effectiveness. Previous evaluations of policies in OECD countries indicate that these programs usually have at best a modest impact on participants' labor market prospects. But at the same time, they also indicate that there is considerable heterogeneity in the impact of these programs. For some groups, a compelling case can be made that these policies generate high rates of return, while for other groups these policies have had no impact and may have been harmful. Our discussion of the methods used to evaluate these policies has more general interest. We believe that the same issues arise generally in the social sciences and are no easier to address elsewhere. As a result, a major focus of this chapter is on the methodological lessons learned from evaluating these programs. One of the most important of these lessons is that there is no inherent method of choice for conducting program evaluations. The choice between experimental and non-experimental methods

or among alternative econometric estimators should be guided by the underlying economic models, the available data, and the questions being addressed. Too much emphasis has been placed on formulating alternative econometric methods for correcting for selection bias and too little given to the quality of the underlying data. Although it is expensive, obtaining better data is the only way to solve the evaluation problem in a convincing way. However, better data are not synonymous with social experiments.

JEL codes: J24; J31; C50; C93; J64

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 US 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 Block 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 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 US government programs have received such intensive scrutiny, and 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 lifecycle 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 30 years of experience in evaluating government training programs. Most of our lessons come from American studies because the US 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 cannot 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. The understanding that responses to the same measured treatment are heterogenous 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 help a lot. The data available to most analysts have been exceedingly crude. 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 non-parametric 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. 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 non-experimental 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 non-experimental methods is attenuated. Variability in estimates across estimators arises from the fact that different non-experimental 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 non-participants. The current practice in the entire "treatment effect" literature is to ignore the indirect effects of programs on non-participants 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 non-participants, 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 the US 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 evaluation problem cast in terms of mean impacts. These prototypes are generalized throughout the rest of this chapter, but the 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 impacts among persons and to other data generating processes of the underlying econometric model. We also reexamine LaLonde's (1986) evidence on the performance of non-experimental 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 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

Table 1
A classification of government employment and training programs

Classroom training	
Basic education	Provides remedial general education, usually with the goal of high school certification
Classroom training in occupational skills	Provides general skills for a specific occupation or industry; duration usually less than 17 weeks
Wage and employment subsidies	
Wage and employment subsidies to private firms	Provides payments to firms, either as a lump sum per employee or as a fraction of employee wages, for hiring new workers; usually targeted at specific groups
Temporary work experience in the public or non-profit sector	Provides general work skills to youth and economically disadvantaged persons with little past employment
Public service employment	Provides temporary public sector jobs to the unemployed, especially the longterm unemployed
On-the-job training	Provides subsidies to employers to hire and train members of specific groups; when subsidy ends after 3–12 months, the employer may retain the trainee as a regular employee; training content varies from little to some; sometimes coordinated with classroom training
Job search assistance	
Employment service	Provides information on job vacancies and assists in matching workers to jobs
Job readiness training	Provides career counseling, assessment and testing to determine job readiness and to indicate appropriate search strategies; may also recommend training
Job search training and subsidies	Provides counseling, instruction in job search skills and resume preparation, job clubs, and resources such as job listings and free phones to call employers

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 US 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 Mincer, 1962, 1993; Heckman et al., 1997b). For example, in the United States, Mincer has estimated that such training amounts to approximately 4–5% 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 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 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 US and the UK. In decentralized training systems, private firms and local organizations play an important role in providing employment and training services.

Table 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

Table 2
Expenditures on employment and training programs in selected OECD countries as a percentage of GDP, 1994–1995[a]

Country	Adult JSA (%)	Adult CT (%)	Adult OJT (%)	Adult WE (%)	Youth All (%)	Total (%)	Disabled All (%)	Total w/disabled (%)	Income support (%)
Australia	0.20	0.17	0.09	0.13	0.07	0.66	0.07	0.73	1.64
Austria	0.13	0.12	0.02	0.03	0.01	0.31	0.06	0.37	1.44
Canada	0.20	0.34	0.00	0.02	0.02	0.60	0.00	0.60	1.54
Denmark	0.12	1.00	0.12	0.46	0.16	1.86	0.46	2.32	4.56
France	0.16	0.44	0.08	0.13	0.27	1.09	0.08	1.17	1.95
Germany	0.23	0.38	0.09	0.31	0.06	1.07	0.26	1.33	2.14
Ireland	0.14	0.48	0.03	0.25	0.43	1.33	0.15	1.48	3.25
Italy	0.08	0.02	–	–	0.83	0.93	–	0.93	1.03
Japan	0.03	0.03	0.05	–	–	0.11	–	0.11	0.35
Netherlands	0.17	0.16	0.01	0.09	0.09	0.52	0.54	1.06	3.06
Norway	0.17	0.23	0.09	0.14	0.08	0.71	0.64	1.35	1.10
Sweden	0.27	0.78	0.36	0.54	0.23	2.18	0.82	3.00	2.54
United Kingdom	0.21	0.13	0.02	0.01	0.13	0.50	0.03	0.53	1.14
United States	0.07	0.04	0.01	0.01	0.03	0.16	0.04	0.20	0.35

[a] Source: OECD (1996, Table T, pp. 206–212). Figures for Ireland and Italy are from 1991 and 1992, respectively. JSA is defined as public employment services and administration. OJT is defined as subsidies to regular employment in the private sector, or support of unemployed persons starting enterprises; WE is defined as direct job creation in the public or non-profit sector. Youth includes measures for unemployed and disadvantaged youth and support of apprenticeship and related general youth training. Income support includes unemployment compensation and early retirement benefits for labor market reasons.

spends 1% of its GDP on CT for adults, the most of any OECD country. However, this expenditure amounts to only 40% 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 US spend only 0.03% and 0.04% of their GDP, respectively, 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 US has led some researchers to comment on the irony that despite these low expenditures, US 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 US evaluations.[1]

We now consider in detail each type of employment and training service in Table 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 of these programs that we review in Section 10.

The first category listed in Table 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 US, 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 2–4 weeks (Jensen et al., 1993) while in Sweden a duration of 4 months and in the United Kingdom and the United States 3 months is more typical. Per capita expenditures on such training vary substantially, with a training slot costing approximately $7500 in Sweden and between $2000 and $3000 in the United States.[2] The Swedish figures include stipends for participants while the US 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 provided CT in government training centers that used 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 US, the provision of CT is highly decentralized and its content depends on the choices made by local

[1] However, the level of total expenditure in the US is still quite large. Relative total expenditures on active labor market policies can be inferred from Table 2 using the relative sizes of each economy compared with the US. For example, the German economy is somewhat less than one-fourth the size of the US economy, and the French, Italian and British economies are approximately one-sixth the size of the US 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 US (see OECD, 1996, Table 1.1, p. 2).

[2] Unless otherwise indicated all monetary units are expressed in 1997 US dollars.

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 1 is 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 US, the Targeted Jobs Tax Credit is an example of this type of program. Heckman et al. (1997b) discuss the empirical evidence on 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 3–12 months in the public or non-profit 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 US 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 1 is subsidized on-the-job training at private firms. The goal of subsidized OJT programs is to induce employers to provide job-relevant skills, including firm-specific skills, to disadvantaged workers. In the US, employers receive a 50% wage subsidy for up to 6 months; in the UK 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 work experience or wage subsidy programs (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).

[3] The provision of subsidized OJT is particularly hard to monitor both because on-the-job training has proven difficult to measure with survey methods (Barron et al., 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 US, see Barron et al. (1989).

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, and incremental training beyond that provided to unsubsidized workers.

The fourth category of services in Table 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 US, but is now used in the UK, 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, differences have emerged 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 1. For example, classroom training in vocational skills might be followed by job search assistance. In the UK, 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 1 into a single package. For example, in the US 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 datasets 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 US 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

[4] For 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.

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 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 US, 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 US programs, but also has characterized programs in Austria, Denmark, Germany, Norway, and the United Kingdom (Westergard-Nielsen, 1993; Björklund and Regner, 1996; 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 present economic models that describe participation in alterative 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, where $D = 1$ denotes treatment and $D = 0$ denotes non-treatment. 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

[5] In this chapter, 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.

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

$$Y_0 = \mu_0(X) + U_0, \tag{3.1a}$$

$$Y_1 = \mu_1(X) + U_1, \tag{3.1b}$$

where $\mathrm{E}(Y_0 \mid X) = \mu_0(X)$ and $\mathrm{E}(Y_1 \mid X) = \mu_1(X)$. 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 $\mathrm{E}(U_1 \mid X) = \mathrm{E}(U_0 \mid X) = 0$, although as we note at several places in this paper, this is not strictly required. These conditions do *not* imply that $\mathrm{E}(U_1 - U_0 \mid X, D = 1) = 0$. D may depend on U_1, U_0 or $U_1 - U_0$ and X. For many of the estimators that we consider in this chapter we allow for the more general case

$$Y_0 = g_0(X) + U_0, \qquad Y_1 = g_1(X) + U_1,$$

where $\mathrm{E}(U_0 \mid X) \neq 0$ and $\mathrm{E}(U_1 \mid X) \neq 0$. Then $\mu_0(X) = g_0(X) + \mathrm{E}(U_0 \mid X)$ and $\mu_1(X) = g_1(X) + \mathrm{E}(U_1 \mid X)$.[6] Thus X is not necessarily exogenous in the ordinary econometric usage of that term.

Note also that Y may be a vector of outcomes or a time series of potential outcomes: (Y_{0t},Y_{1t}), for $t = 1,\ldots,T$, 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: $\mathrm{E}(U_{0t} \mid X) = 0$, $\mathrm{E}(U_{1t} \mid X) = 0$ where $X = (X_1,\ldots,X_T)$. In defining a sequence of "treatment on the treated" parameters, $\mathrm{E}(Y_{1t} - Y_{0t} \mid X, D = 1)$, $t = 1,\ldots,T$, 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 $Y^{\mathrm{p}} = ((Y_{01},Y_{11}),\ldots,(Y_{0T},Y_{1T}))$ fixed

[6] For example, an exogeneity assumption is not required when using social experiments to identify $\mathrm{E}(Y_1 - Y_0 \mid X, D = 1)$.

where Y^P is the vector of potential outcomes. More precisely, we require that for the conditional density of the data

$$f(X \mid D, Y^P) = f(X \mid Y^P),$$

i.e., we require that the realization of D does not determine X given the vector of potential outcomes. Otherwise, the parameter $E(Y_1 - Y_0 \mid X, D = 1)$ 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 $E(Y_{1t} - Y_{0t} \mid X, D = 1)$ defined on subsets of X, say X_c, simply integrate out $E(Y_{1t} - Y_{0t} \mid X, D)$ against the density $f(\tilde{X}_c \mid D = 1)$ 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

$$\Delta = Y_1 - Y_0. \tag{3.2}$$

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 non-participants and the payments for the

program, for both participants and non-participants, 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 participants and non-participants. Thus a participant may pay taxes to support the program just as persons who do not 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 with each possible state from each possible program.

The traditional program evaluation literature assumes that the valuation of the direct effects of the program boils 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 institu-

tions being evaluated. The classic example is the value of non-market time (Greenberg, 1997). Traditional program evaluations exclude such valuations largely because of the difficulty of imputing the value and quantity of non-market time. By doing this, however, these evaluations value labor supply in the market sector at the market wage, but value labor supply in the non-market sector at a zero wage. By contrast, individuals value labor supply in the non-market 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; Heckman et al., 1997c).

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 non-participation. 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:

$$\Pr(Y_1 > Y_0 \mid D = 1)\Pr(D = 1) = \Pr(\Delta > 0 \mid D = 1)\Pr(D = 1);$$

(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 "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 proportion 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 answer to the question "does the distribution of outcomes for the participants dominate the distribution of outcomes if they did not participate?" (see Heckman et al., 1997c; 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 plays 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 et al., 1997c; Heckman and Smith, 1998a, 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

$$Y = DY_1 + (1 - D)Y_0. \tag{3.3}$$

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:

$$\mathrm{E}(Y_j \mid X) = \mu_j(X) = X\beta_j, \qquad j = 0, 1.$$

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

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

Observe that from the definition of a conditional mean, $\mathrm{E}(U_0 \mid X) = 0$ and $\mathrm{E}(U_1 \mid X) = 0$.

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

$$\mathrm{E}(Y_1 - Y_0 \mid X) = \mathrm{E}(\Delta \mid X).$$

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

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

where the term in braces is the "error."

If the model is specialized so that there are K regressors plus an intercept and $\beta_1 = (\beta_{10},\ldots,\beta_{1K})$ and $\beta_0 = (\beta_{00},\ldots,\beta_{0K})$, where the intercepts occupy the first position, and the

slope coefficients are the same in both regimes:

$$\beta_{1j} = \beta_{0j} = \beta_j, \qquad j = 1, \ldots, K$$

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

$$\mathrm{E}(Y_1 - Y_0 \mid X) = \beta_{10} - \beta_{00} = \alpha. \tag{3.6}$$

The regression model for this special case may be written as

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

It is non-standard 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 $\mathrm{E}[U_0 + D(U_1 - U_0)] = \mathrm{E}(U_1 - U_0 \mid D = 1)\Pr(D = 1)$. If $U_1 - U_0$, or variables statistically dependent on it, help determine D, $\mathrm{E}(U_1 - U_0 \mid D = 1) \neq 0$. Intuitively, if persons who have high gains $(U_1 - U_0)$ are more likely to appear in the program, then this term is positive.

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

$$\mathrm{E}(\Delta \mid X, D = 1) = \mathrm{E}(Y_1 - Y_0 \mid X, D = 1) = X(\beta_1 - \beta_0) + \mathrm{E}(U_1 - U_0 \mid X, D = 1). \tag{3.8}$$

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

$$\begin{aligned} Y &= \mu_0(X) + D[\mathrm{E}(\Delta \mid X, D = 1)] + \{U_0 + D[(U_1 - U_0) - \mathrm{E}(U_1 - U_0 \mid X, D = 1)]\} \\ &= X\beta_0 + D[X(\beta_1 - \beta_0) + \mathrm{E}(U_1 - U_0 \mid X, D = 1)] \\ &\quad + \{U_0 + D[(U_1 - U_0) - \mathrm{E}(U_1 - U_0 \mid X, D = 1)]\}. \end{aligned} \tag{3.9}$$

$\mathrm{E}(\Delta \mid X, D = 1)$ is a non-standard parameter in conventional econometrics. It combines "structural" parameters $(X(\beta_1 - \beta_0))$ with the means of the unobservables $(\mathrm{E}(U_1 - U_0 \mid X, D = 1))$. 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 non-standard. 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:

$$Y = X\beta + D\alpha + U, \tag{3.10}$$

where $\mathrm{E}(U) = 0$. 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 et al., 1997c, and the evidence presented below).

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

$$\mathrm{E}(U_1 - U_0 \mid X, D = 1) = 0. \tag{3.11}$$

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 arises 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 $\mathrm{E}(U \mid X) = 0$. To see this point take the expectation of the terms in braces in (3.7) and (3.9), respectively, to obtain the following:

$$\mathrm{E}(U_0 + D(U_1 - U_0) \mid X, D) = \mathrm{E}(U_0 \mid X, D)$$

and

$$\mathrm{E}(U_0 + D[(U_1 - U_0) - \mathrm{E}(U_1 - U_0 \mid X, D = 1)] \mid X, D) = \mathrm{E}(U_0 \mid X, D).$$

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:

$$\mathrm{Var}(U_0 + D(U_1 - U_0) \mid X, D) = \mathrm{Var}(U_0 \mid X, D)$$

$$+2\mathrm{Cov}(U_0, U_1 - U_0 \mid X, D)D + \mathrm{Var}(U_1 - U_0 \mid X, D)D.$$

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. Eq. (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.

3.4. Is treatment on the treated an interesting economic parameter?

What economic question does parameter (3.8) 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 (3.8) 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

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

participation and non-participation and a vector of policy variables φ that affect the outcomes in both states and the allocation of all 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 also generate indirect effects. $c(\varphi)$ is the social cost of φ denominated in "1" 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

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

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

$$\Delta(\varphi) = \frac{\partial N_1(\varphi)}{\partial \varphi}[\mathrm{E}(Y_1 \mid D = 1, \varphi) - \mathrm{E}(Y_0 \mid D = 0, \varphi)]$$

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

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 are zero. In this case, the parameters of interest for evaluating the impact of the policy change on GDP are

(i) $\partial N_1(\varphi)/\partial\varphi$; the number of people entering or leaving state 1.

(ii) $\mathrm{E}(Y_1 \mid D = 1, \varphi) - \mathrm{E}(Y_0 \mid D = 0, \varphi)$; the mean output difference between sectors.

(iii) $\partial c(\varphi)/\partial\varphi$; 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) $\mathrm{E}(Y_1 - Y_0 \mid D = 1, \varphi)$ or

(b) $\mathrm{E}(Y_1 \mid \varphi = \bar{\varphi}) - \mathrm{E}(Y_0 \mid \varphi = 0)$ where $\varphi = \bar{\varphi}$ sets $N_1(\bar{\varphi}) = \bar{N}$, 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; 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 ($1(Z \geq 0) = 1$ if $Z \geq 0$; $= 0$ otherwise),

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

and

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

where $f(x)$ is the density of x. Total output is

$$\bar{N}\int_{-\infty}^{\infty} y_1 \int_{-k(\varphi)}^{\infty} f(y_1, x \mid \varphi)\mathrm{d}x\,\mathrm{d}y_1 + \bar{N}\int_{-\infty}^{\infty} y_0 \int_{-\infty}^{-k(\varphi)} f(y_0, x \mid \varphi)\mathrm{d}x\,\mathrm{d}y_0 - c(\varphi).$$

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:

$$\Delta(\varphi) = \bar{N}k'(\varphi)f_x(-k(\varphi))[\mathrm{E}(Y_1 \mid D = 1, x = -k(\varphi), \varphi) - \mathrm{E}(Y_0 \mid D = 0, x = -k(\varphi), \varphi)]$$

$$+\bar{N}[\int_{-\infty}^{\infty} y_1 \int_{-k(\varphi)}^{\infty} \frac{\partial f(y_1, x \mid \varphi)}{\partial \varphi}\mathrm{d}x\,\mathrm{d}y_1 + \int_{-\infty}^{\infty} y_0 \int_{-\infty}^{-k(\varphi)} \frac{\partial f(y_0, x \mid \varphi)}{\partial \varphi}\mathrm{d}x\,\mathrm{d}y_0] - \frac{\partial c(\varphi)}{\partial \varphi}. \tag{3.13}$$

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 7.4.5. 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 ($X = -k(\varphi)$):

$$\mathrm{E}(Y_1 \mid D = 1, X = -k(\varphi), \varphi) - \mathrm{E}(Y_0 \mid D = 0, X = -k(\varphi), \varphi)$$

$$= \mathrm{E}(Y_1 - Y_0 \mid D = 1, X = -k(\varphi), \varphi). \tag{3.14}$$

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,

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

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

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

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

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, so

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

and

$$\mathrm{E}(Y_0 \mid D = 1, \varphi = \tilde{\varphi}) = \mathrm{E}(Y_0 \mid D = 1, \varphi = 0).$$

This condition defines the absence of general equilibrium effects in the base state so the no program state for non-participants is the same as the non-participation state. Assumption (3.15) is what enables analysts to generalize from partial equilibrium to general equili-

[8] Heckman and Smith (1998a) and Heckman (1997) present comprehensive discussions of the Imbens and Angrist (1994) parameter. We discuss this parameter further in Section 7.4.5. 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 them.

brium settings. Recalling that $\bar{N} = N_1(\varphi) + N_0(\varphi)$, when (3.15) holds we have[9]

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

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]

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: $\mathrm{E}(Y_1 - Y_0 \mid X, D = 1)$. All use some form of comparison to construct the required counterfactual $\mathrm{E}(Y_0 \mid X, D = 1)$. Data on $\mathrm{E}(Y_1 \mid X, D = 1)$ 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 non-participants to estimate what they would have earned if they had participated. For that problem the challenge is to find $\mathrm{E}(Y_1 \mid X, D = 0)$ since the data on non-participants enables one to identify $\mathrm{E}(Y_0 \mid X, D = 0)$.

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

[9] Condition (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.

[10] Björklund 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.

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 no-treatment state in the post-program period. In other words, the underlying identifying assumption is

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

If this assumption is valid, the "before–after" estimator is given by

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

where the subscript "1" denotes conditioning on $D = 1$, and the bar 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 lifecycle 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 Figs. 1–6 which show the dip for a variety of programs in different countries. If this decline in earnings is transitory, and earnings follow 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

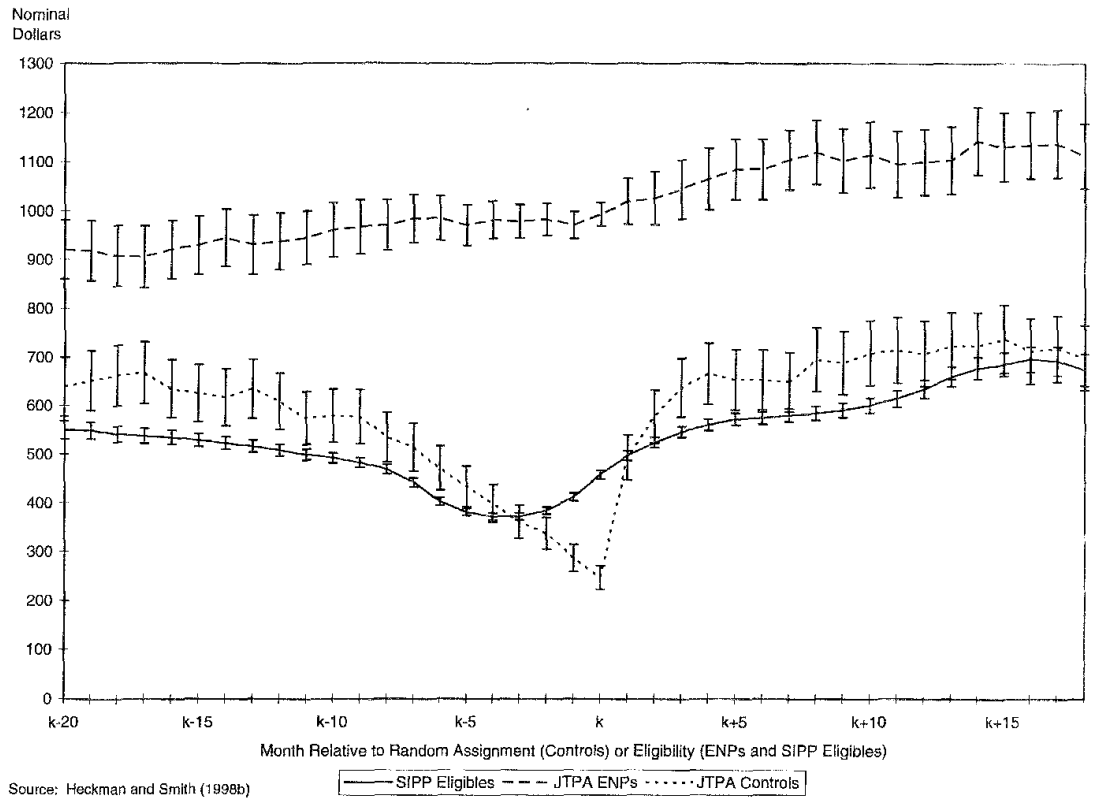

Fig. 1. Mean self-reported monthly earnings: National JTPA Study controls and eligible non-participants (ENPs) and SIPP eligibles (male adults). Source: Heckman and Smith (1999).

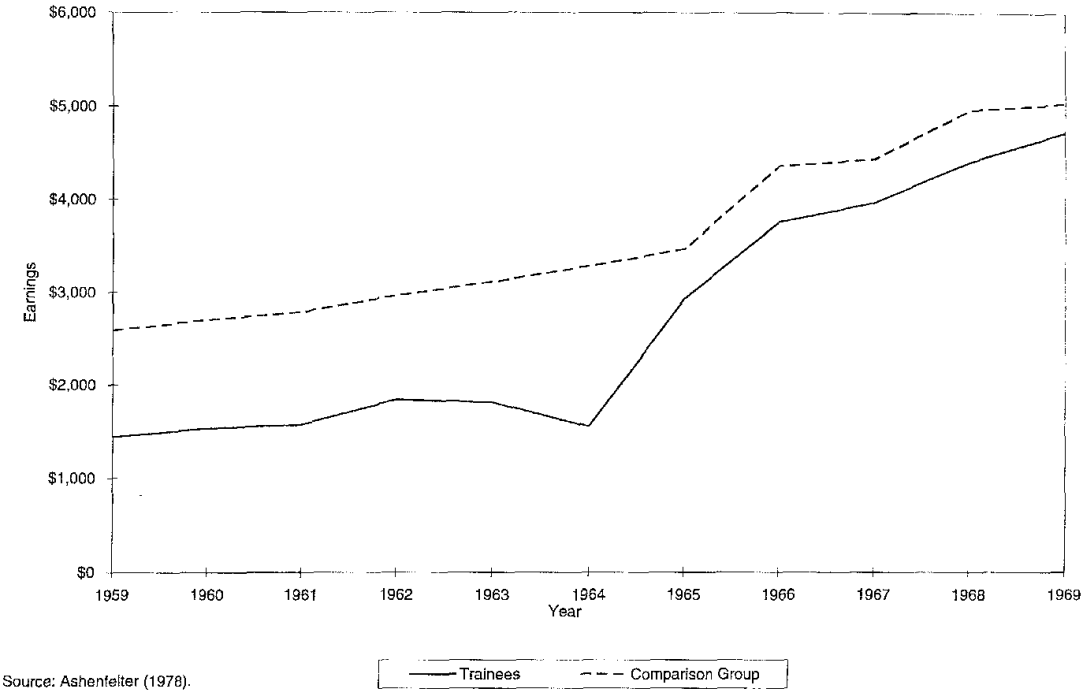

Fig. 2. Mean annual earnings prior, during and subsequent to training for 1964 MDTA classroom trainees and a comparison group (white males).

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 non-participants in periods t and t'. If the mean change in the no-program outcome measures are the same for participants and non-participants i.e., if the following assumption is valid:

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

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

$$(\bar{Y}_{1t} - \bar{Y}_{0t'})_1 - (\bar{Y}_{0t} - \bar{Y}_{0t'})_0, \qquad t > k > t' \tag{4.2}$$

[11] See also Heckman and Robb (1985a, pp. 210–215).

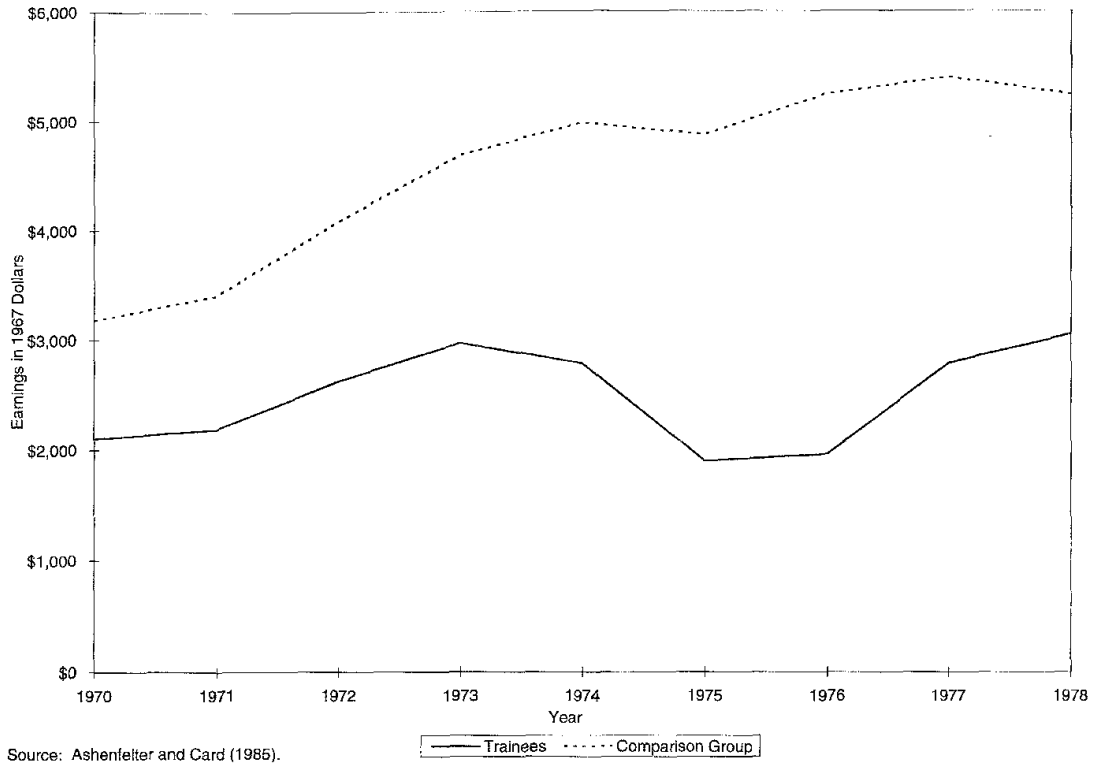

Fig. 3. Mean annual earnings for 1976 CETA trainees and a comparison group (males).

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] 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.

[12] The proof is immediate. Make the following decomposition: $(\bar{Y}_{1t} - \bar{Y}_{0t'})_1 = (\bar{Y}_{1t} - \bar{Y}_{0t})_1 + (\bar{Y}_{0t} - \bar{Y}_{0t'})_1$ The claim follows upon taking expectations.

[13] One 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 et al. (1997c) can be applied. See their paper for details.

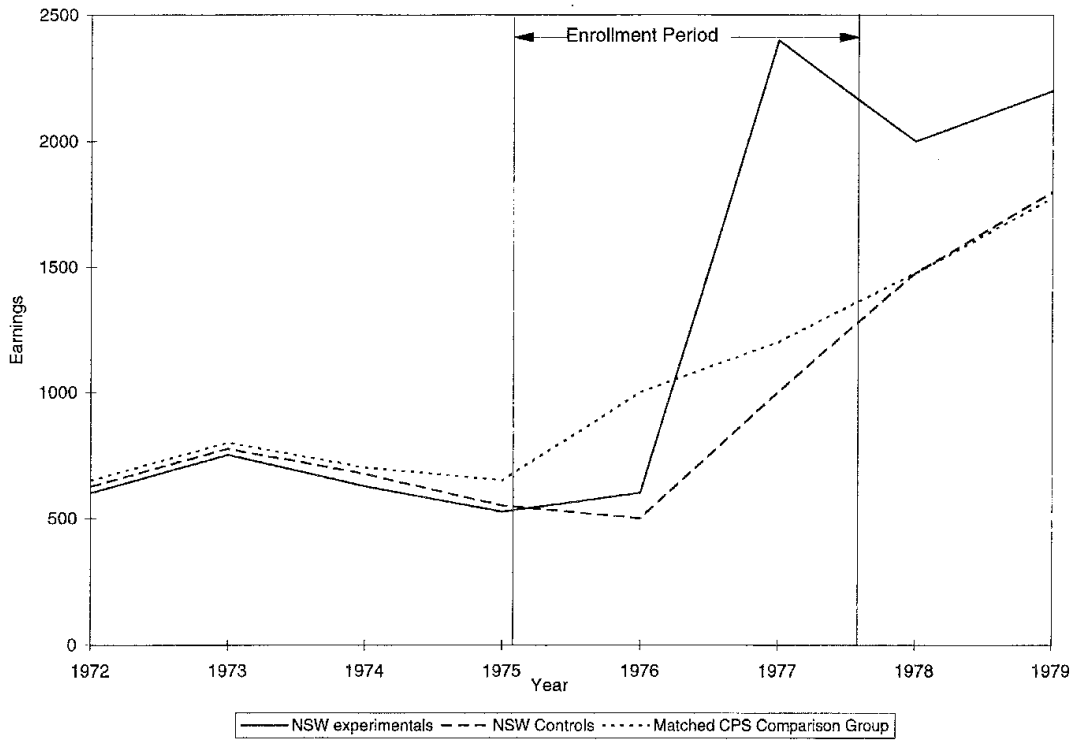

Fig. 4. National supported work (NSW) average annual earnings, treatments, controls and matched CPS comparison group (AFDC recipients).

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-treatment earnings between t' and t will be different between participants and non-participants. 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 non-participants 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

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

i.e., that on average persons who do not participate in the program have the same no-

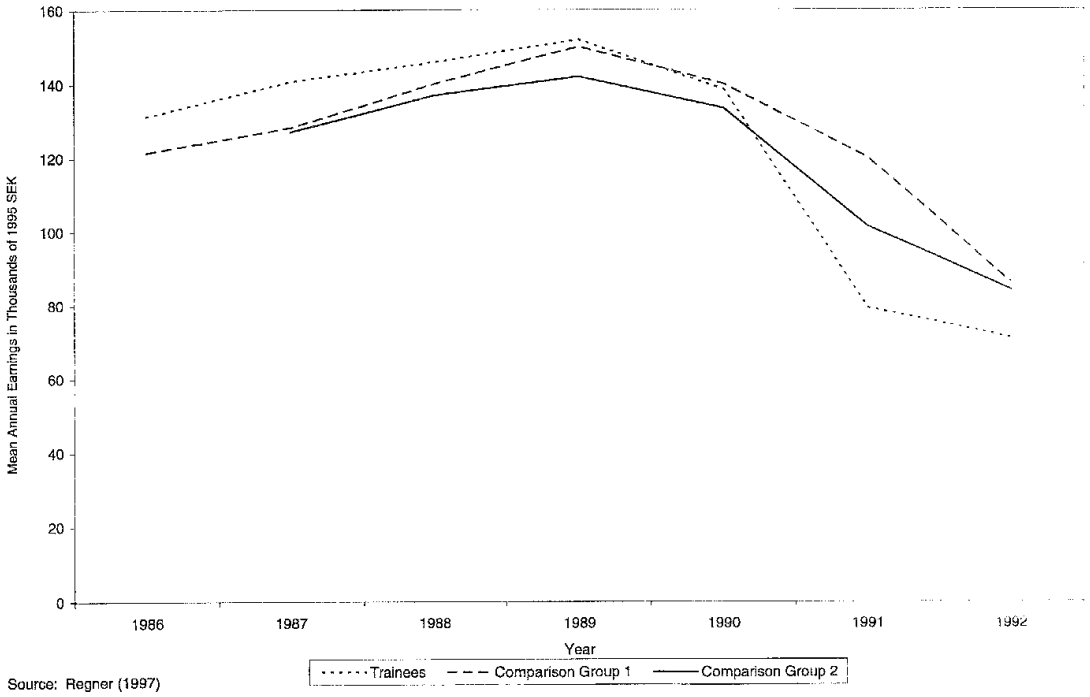

Fig. 5. Earnings of participants in Swedish UI training in 1991 and two comparison groups (adult males aged 26-54).

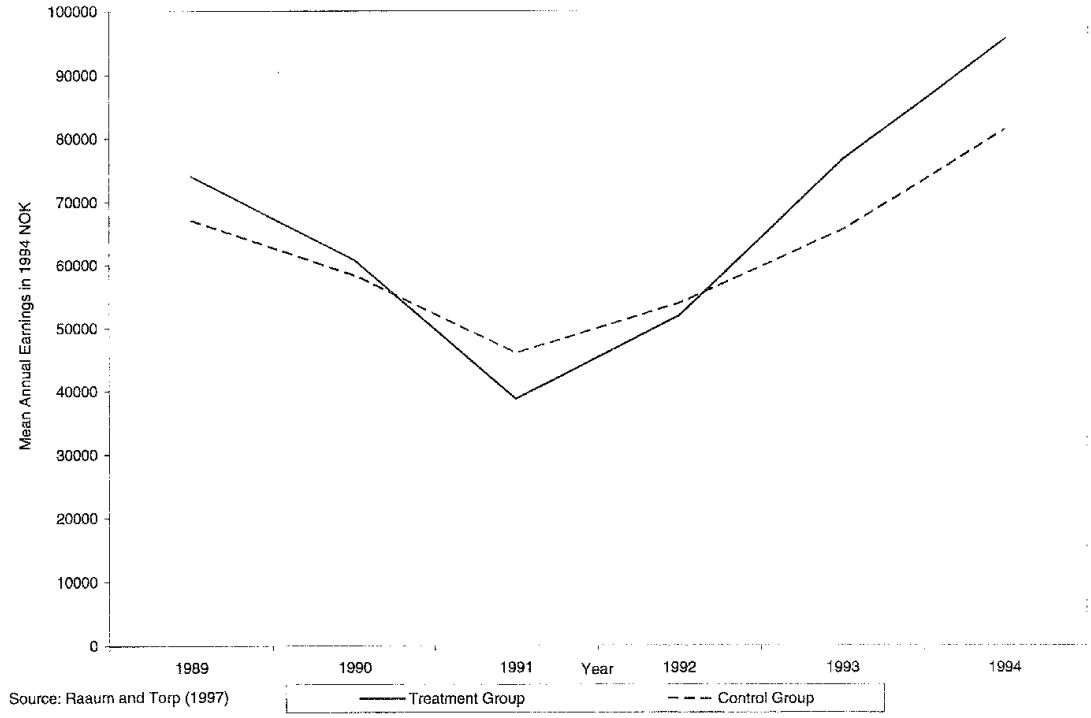

Fig. 6. Earnings of 1991 participants in Norwegian Labor Market Training Program and a randomly assigned control group (all participants).

treatment outcome as those who do participate. If this assumption is valid, then the *cross-section* estimator is given by

$$(\bar{Y}_{1t})_1 - (\bar{Y}_{0t})_0. \tag{4.3}$$

This estimator is valid under assumption (4.A.3) because[14]

$$\mathrm{E}((\bar{Y}_{1t})_1 - (\bar{Y}_{0t})_0) = \mathrm{E}(\Delta_t \mid D = 1).$$

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 non-participants, 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 makes 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 non-participants ($D = 0$), conditioning on X may eliminate systematic differences in outcomes between the two groups. Using modern non-parametric procedures, it is possible to exploit each of the identifying conditions to estimate non-parametric versions of all three estimators. On the other hand, if the difference between participants and non-participants is due to unobservables, conditioning may accentuate, and not eliminate, differences between participants and non-participants 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 non-stationary environments and different values of regressors in the different comparison groups. We first consider social experimentation and how it constructs the counterfactuals used in policy evaluations.

[14] Proof: $(\bar{Y}_{1t})_1 - (\bar{Y}_{0t})_0 = (\bar{Y}_{1t})_1 - (\bar{Y}_{0t})_1 + (\bar{Y}_{0t})_1 - (\bar{Y}_{0t})_0$ and take expectations invoking assumption (4.A.3).

[15] Thus if $\mid \mathrm{E}(Y_0 \mid D = 1) - \mathrm{E}(Y_0 \mid D = 0) \mid = M$, there is no guarantee that $\mid \mathrm{E}(Y_0 \mid D = 1, X) - \mathrm{E}(Y_0 \mid D = 0, X) \mid < M$. 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 US 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

$$\mathrm{E}(Y_1^* - Y_0^* \mid X, D^* = 1) = \mathrm{E}(Y_1 - Y_0 \mid X, D = 1). \tag{5.A.1}$$

A stronger set of conditions, not strictly required, are

$$\mathrm{E}(Y_1^* \mid X, D^* = 1) = \mathrm{E}(Y_1 \mid X, D = 1) \tag{5.A.2a}$$

and

$$\mathrm{E}(Y_0^* \mid X, D^* = 1) = \mathrm{E}(Y_0 \mid X, D = 1). \tag{5.A.2b}$$

Assumption (5.A.1) 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 outcomes of

[16] We discuss this evidence in Section 10.

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

$$\mathrm{E}(Y_1 \mid X, D = 1, R = 1) = \mathrm{E}(Y_1 \mid X, D = 1).$$

Under (5.A2a) 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:

$$\mathrm{E}(Y_0 \mid X, D = 1, R = 0) = \mathrm{E}(Y_0 \mid X, D = 1).$$

Simple mean differences identify

$$\mathrm{E}(\Delta \mid X, D = 1) = \mathrm{E}(Y_1 - Y_0 \mid X, D = 1).$$

Within the context of the model of Eq. (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 Eq. (3.1), under (5.A.2a) and (5.A.2b), $\mathrm{E}(Y \mid X, D = 1, R = 1) = g_1(X) + \mathrm{E}(U_1 \mid X, D = 1)$ and $\mathrm{E}(Y \mid X, D = 1, R = 0) = g_0(X) + \mathrm{E}(U_0 \mid X, D = 1)$.[17]

Rewriting the first conditional mean, we obtain

$$\mathrm{E}(Y \mid X, D = 1, R = 1) = g_1(X) + \mathrm{E}(U_1 - U_0 \mid X, D = 1) + \mathrm{E}(U_0 \mid X, D = 1).$$

Subtracting the second mean from the first eliminates the common selection bias component $\mathrm{E}(U_0 \mid X, D = 1)$ so

$$\mathrm{E}(Y \mid X, D = 1, R = 1) - \mathrm{E}(Y \mid X, D = 1, R = 0) = g_1(X) - g_0(X) + \mathrm{E}(U_1 - U_0 \mid X, D = 1).$$

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 $\mathrm{E}(U_1 \mid X, D = 1)$ or $\mathrm{E}(U_0 \mid X, D = 1)$ equal to zero. Rather, it balances the selection bias in the treatment and control groups.

Stronger assumptions must be made to identify the distribution of impacts $F(\Delta \mid D = 1)$.[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 and Smith, 1993, 1995, 1998a; Heckman et al., 1997c).

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.

[17] Notice 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 $\mathrm{E}(U_0 \mid X) \neq 0$ and $\mathrm{E}(U_1 \mid X) \neq 0$.

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

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

$$Y_{1i} - Y_{0i} = \Delta_i \equiv \Delta, \qquad \text{for all } i. \tag{5.A.3}$$

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 $\mathrm{E}(\Delta \mid X)$ but not Δ at the time participation decisions are made. Formally, the second condition is

$$\mathrm{E}(\Delta \mid X, D = 1) = \mathrm{E}(\Delta \mid X), \tag{5.A.4}$$

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 $\mathrm{E}(\Delta \mid X,\ D = 1)$.

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

$$\mathrm{E}(\Delta \mid X, D = 1) = \mathrm{E}(Y_1 - Y_0 \mid X, D = 1).$$

Using assumption (4.A.3) in an ordinary non-experimental evaluation, there may be values of X such that $\Pr(D = 1 \mid X) = 1$; that is, there may be values of X with no comparison group members. Randomization avoids this difficulty by balancing the distribution of X 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 $\Pr(D = 1 \mid X) = 0$.

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:

$$\mathrm{E}(\Delta \mid X, D = 1) = \mathrm{E}(Y_1 - Y_0 \mid X, D = 1)$$

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

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

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 US 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 $E(\Delta \mid X)$ using experimental means. One assumption that guarantees compliance is the existence of a "compensation" or "punishment" level c such that

$$\Pr(D = 1 \mid X, c) = 1 \tag{5.A.5a}$$

and

$$E(\Delta \mid X, c) = E(\Delta \mid X). \tag{5.A.5b}$$

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 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 $\Pr(D = 1 \mid X) \neq 0$. Then Heckman (1996) shows that

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

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

Randomization of eligibility produces samples that can be used to identify $E(\Delta \mid X, D = 1)$ and also to recover $\Pr(D = 1 \mid X)$. 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 $E(\Delta \mid X)$.

5.2. *Intention to treat and substitution bias*

The objective of most experimental designs is to estimate the conditional mean impact of training, or $E(\Delta \mid X, D = 1)$. 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 $E(\Delta \mid X, D = 1)$ 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 members indicating whether or not they would have actually received training, had they been in the treatment group. Define

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

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 $E(\Delta \mid X, D = 1)$ 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 receive 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

$$E(Y \mid X, D = 1, R = 1, T = 0) = E(Y \mid X, D = 1, R = 0, T^* = 0). \tag{5.A.6}$$

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

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

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

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 et al. (1998f) 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 $\mathrm{E}(Y_0 \mid X, D=1)$ 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.

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 et al. (1998a) 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

$$Y_{1i} - Y_{0i} = Y_{2i} - Y_{0i} = \Delta_i \equiv \Delta, \qquad \text{for all } i. \tag{5.A.3$'$}$$

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

[21] See, e.g., Mallar (1978), Bloom (1984) and Heckman et al. (1998f).

same way to the program (a common effect assumption). The second assumption is a generalized version of (5.A.4), where

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

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

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

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 et al., 1998a). 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 US 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 9–18 months of guaranteed work experience in unskilled occupations

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

on groups of longterm AFDC (welfare) recipients, ex-drug addicts, ex-criminal offenders, and economically disadvantaged youths in 10 sites across the US. 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 in selecting the training centers at which random assignment would take place. Initially, evaluators planned to use a random sample of the nearly 600 US 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 services provided by JTPA. This program offers a personalized mix of employment and training services including all those listed in Table 1 with the exception of public service employment.

[23] Very 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 US$1 million in payments to cover the cost of participating in the experiment.

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), but not on-the-job training (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, followup data on earnings, employment and other outcomes were obtained from both surveys and multiple 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 non-participants engage. Experiments exacerbate this problem by creating a pool of persons who attempt to take training who then flock to substitute programs when they are placed in an experimental control group.

Table 3 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 3 shows that as many as 40% 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 fractions 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 treat-

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

ment 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 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 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 followup period, while application of the IV estimator that uses sample moments in place of (5.2) yields an adjusted impact estimate of $1100 (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

[25] For the NJS, Table 3 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-reported data on controls but the combined data for the treatment group will likely overstate the difference in service receipt levels between the two groups.

Table 3
Treatment group dropout and control group substitution in experimental evaluations of active labor market policies (fraction of experimental treatment and control groups receiving services)[a]

Study	Authors/time period	Target group(s)	Fraction of treatments receiving services	Fraction of controls receiving services
1. NSW*	Hollister et al. (1984)	Longterm AFDC women	0.95	0.11
	(9 months after RA)	Ex-addicts	NA	0.03
		17–20 year old HS dropouts	NA	0.04
2. SWIM	Friedlander and Hamilton (1993)	AFDC women: applicants and recipients		
	(Time period not reported)	a. Job search assistance	0.54	0.01
		b. Work experience	0.21	0.01
		c. Classroom training/OJT	0.39	0.21
		d. Any activity	0.69	0.30
		AFDC-U unemployed fathers		
		a. Job search assistance	0.60	0.01
		b. Work experience	0.21	0.01
		c. Classroom training/OJT	0.34	0.22
		d. Any activity	0.70	0.23
3. JOBSTART	Cave et al. (1993)	Youth HS dropouts		
	(12 months after RA)	Classroom training/OJT	0.90	0.26
4. Project Independence	Kemple et al. (1995)	AFDC women: applicants and recipients		
	(24 months after RA)	a. Job search assistance	0.43	0.19
		b. Classroom training/OJT	0.42	0.31
		c. Any activity	0.64	0.40

Table 3 (*continued*)

Study	Authors/time period	Target group(s)	Fraction of treatments receiving services	Fraction of controls receiving services
5. New Chance	Quint et al. (1994) (18 months after RA)	Teenage single mothers		
		Any education services	0.82	0.48
		Any training services	0.26	0.15
		Any education or training	0.87	0.55
6. NJS	Heckman and Smith (1998a,b) (18 months after RA)	Self-reported from Survey Data		
		Adult males	0.38	0.24
		Adult females	0.51	0.33
		Male youth	0.50	0.32
		Female youth	0.58	0.41
		Combined Administrative and Survey Data		
		Adult males	0.74	0.25
		Adult females	0.78	0.34
		Male youth	0.81	0.34
		Female youth	0.81	0.42

[a] Sources: Masters and Maynard (1981, p. 148, Table A.15); Maynard (1980, p. 169, Table A14); Friedlander and Hamilton (1993, p. 22, Table 3.1); Cave et al. (1993, p. 95, Table 4-1); Kemple et al. (1995, p. 58, Table 3.5); Quint et al. (1994, p. 110, Table 4.9); Heckman and Smith (1998a,b) and calculations by the authors. Notes: RA, random assignment; HS, high school. Service receipt includes any employment and training services. The services received by the controls in the NSW study are CETA and WIN jobs. For the longterm AFDC women, this measure also includes regular public sector employment during the period.

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 impact of training on 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 distributions of observed wages in the treatment and control groups that result from the effect of the program on wage rates from those that result from the effect of the program on selection into employment. Under these

[26] Alternatively, 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.

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 analysts 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 the NJS presented in the two official experimental impact reports, Bloom et al. (1993) and Orr et al. (1994).[27] As shown in Table 4, 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

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

Table 4
Variability in experimental impact estimates for adult women in the NJS (mean difference in earnings between the experimental treatment and control groups during the 18 months after random assignment)[a]

Treatment stream	Follow-up report ($)	
	18 month report Bloom et al. (1993)	30 month report Orr et al. (1994)
Recommended for classroom training		
1–6 months	−65	−121
7–18 months	463	312
Sample size	2847	2343
Recommended for on-the-job training		
1–6 months	225	255
7–18 months	518	418
Sample size	2287	2284
Recommended for other services		
1–6 months	171	238
7–18 months	286	879
Sample size	1340	1475

[a] Sources: Bloom et al. (1993, pp. 106, Exhibit 4.12); Orr et al. (1994, pp. 121, 129, 131, Exhibits 5.1, 5.5, and 5.7). Notes: Orr et al. (1994) report the impact per enrollee obtained using the Bloom (1984) estimator rather than the impact per treatment group member. To make the figures in the two columns comparable, we adjusted the impacts per enrollee by the fraction of the treatment group in each recommended service category who enrolled in JTPA. The fraction enrolling among those recommended for classroom training is 0.719, among those recommended for on-the-job training it is 0.532, and among those recommended for other services it is 0.499.

from the evaluation sample by the 18 month followup interview was 10% for the adult women, but more than 30% for the male participants. In the NJS study, sample attrition by the 18 month followup was 12% for the adult women and approximately 20% for 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 followup 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 (1999) and Heckman et al. (1998b).

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":

$$Y_{0j}, \qquad j = 1, \ldots, k.$$

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

$$(Y_{0j}, Y_{1j}), \qquad j = k+1, \ldots, T,$$

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

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

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

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

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 training 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

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

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; 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., $\mathrm{E}(Y_{1t} - Y_{0t} \mid X, D = 1) = \mathrm{E}(Y_{1t} - Y_{0t} \mid X)$ 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

$$D = 1, \qquad \text{if } \mathrm{E}\left(\sum_{j=1}^{T-k} \frac{\alpha_{k+j}}{(1+r)^j} - c - Y_{0k} \mid I_k\right) \geq 0,$$

$$D = 0, \qquad \text{otherwise.} \tag{6.3}$$

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

$$D = 1, \qquad \text{if } \mathrm{E}\left(\frac{\alpha}{r} - c - Y_{0k} \mid I_k\right) \geq 0,$$

$$D = 0, \qquad \text{otherwise.} \tag{6.4}$$

[28] A sharp contrast between a model of perfect certainly 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.

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

$$\mathrm{E}\left[\left(\sum_{j=1}^{T-k}\frac{\alpha_{k+j}}{(1+r)^j}\right) \mid I_k\right].$$

Future rewards are the same for everyone of the same age and with the same discount rate. Future values of Y_{0t} do not directly determine participation given Y_{0k}. The link between D and 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 non-participants. 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

$$\Pr(D = 1 \mid X) = \Pr(V < H(X) \mid X). \tag{6.5}$$

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

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

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

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

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

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

Although these functional forms are traditional, they are restrictive and are not required. Conditions for non-parametric identifiability of $\Pr(D = 1 \mid X)$ 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 non-parametric 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 $\Pr(D = 1 \mid X)$ 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 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 $[\mathrm{Var}(U_1 - U_0)]^{1/2}$. This is a standard result in discrete choice theory. Thus in the Roy model it is possible to recover $\mathrm{E}(Y_1 - Y_0 \mid X)$ 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 $F(y_0, y_1 \mid X)$ and to answer *all* of the evaluation questions about

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

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 (4.3) – which is valid if assumption (4.A.3) is correct. Given decision rule (6.3), under what conditions is it plausible to assume that

$$\mathrm{E}(Y_{0t} \mid D = 1) = \mathrm{E}(Y_{0t} \mid D = 0), \qquad t > k \tag{4.A.3}$$

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

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

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 $t > k + l$, assumption (4.A.3) would be satisfied in such periods.

Considerable evidence indicates that earnings have an autoregressive component (see, e.g., Ashenfelter, 1978; MaCurdy, 1982; Ashenfelter and Card, 1985; 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 (4.A.3) will be satisfied if it is revised to condition on observables X:

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

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

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

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

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

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

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

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

and

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

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.

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 non-participants.[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

$$Y_{0t} = \beta_t + \varphi + U_{0t}, \qquad \text{for all } t \tag{6.7}$$

and $\mathrm{E}(U_{0t} \mid \varphi) = 0$ 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

$$\mathrm{E}(Y_{0t} - Y_{0t'} \mid D = 1) = \beta_t - \beta_{t'},$$

since $\mathrm{E}(U_{0t} \mid D = 1) - \mathrm{E}(U_{0t'} \mid D = 1) = 0$.

From the assumption of time homogeneity, $\beta_t = \beta_{t'}$. Thus assumption (4.A.1) is satis-

[32] This 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.

fied and the before–after estimator identifies α_t. It is clearly not necessary to assume that the U_{0t} are mutually independent, just that

$$\mathrm{E}(U_{0t} - U_{0t'} \mid D = 1) = 0, \tag{6.8}$$

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 $t = k + l$ and $t' = k - l$) and conditional expectations forward and backward are symmetric, so that

$$\mathrm{E}(U_{0t} \mid c + \beta_k + U_{0k}) = \mathrm{E}(U_{0t'} \mid c + \beta_k + U_{0k}), \tag{6.9}$$

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

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

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 $\mathrm{E}(\eta) = 0$. Suppose c is independent of all the ν_j and η, and

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

and

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

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 $\mathrm{E}(\nu_{t-j}) = 0$ at all leads and lags, and are uncorrelated with η, and 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 (1982, 1985a, 1986a), 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 non-participants. 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 lifecycle 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 non-participants will be inappropriate and the difference-in-differences estimator will not identify true program impacts.

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}:

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

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

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 $\mathrm{E}(U_{1t} \mid X) \neq 0$ and $\mathrm{E}(U_{0t} \mid X) \neq 0$. The separability enables us to isolate the effect of self selection, as it operates through the "error term", from the structural outcome equation:

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

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

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 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

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

and

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

[34] The 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 $f(U_{0t}, V \mid X)$. From (P.2) $f(U_{0t}, V \mid X) = f(U_{0t}, V)$. Thus

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

Now

$$\Pr(D = 1 \mid X) = \int_{\infty}^{H(X)} f(V)\mathrm{d}V.$$

Inverting, we obtain $H(X) = F_V^{-1}(\Pr(D = 1 \mid X))$. Thus

$$\mathrm{E}(U_{0t} \mid X, D = 1) = \frac{\int_{-\infty}^{\infty} U_{0t} \int_{-\infty}^{F_V^{-1}(\Pr(D=1|X))} f(U_{0t}, V)\mathrm{d}V \, \mathrm{d}U_{0t}}{\Pr(D = 1 \mid X)} \overset{def}{=} K_{0t}(\Pr(D = 1 \mid X)).$$

The mean error term is a function of P, the probability of participation in the program. This special case receives empirical support in Heckman et al. (1997a, 1998b). It enables analysts to characterize the dependence between U_{0t} and X by the dependence of U_{0t} on $\Pr(D = 1 \mid X)$ 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 $\Pr(D = 1 \mid X)$. 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 it is the basis for the "control function" estimators developed in Section 7.

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 $\mathrm{E}(U_{0t} \mid X,\ D = 1)$. 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,

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

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. $\mathrm{E}(\alpha_t \mid I_k)$ 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 non-participants would have been had they participated, then the two parameters are the same. This would arise if both participants and non-participants have the same ex ante expected gains

$$\mathrm{E}(\alpha_t \mid D = 1, I_k) = \mathrm{E}(\alpha_t \mid D = 0, I_k) = \mathrm{E}(\alpha_t \mid I_k),$$

and if

$$\mathrm{E}[\mathrm{E}(\alpha_t \mid I_k) \mid X, D = 1] = \mathrm{E}[\mathrm{E}(\alpha_t \mid I_k) \mid X, D = 0],$$

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 non-participants. 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

$$\mathrm{E}[\mathrm{E}(\alpha_t \mid I_k, D = 1) \mid X, D = 1] = \mathrm{E}[\mathrm{E}(\alpha_t \mid I_k, D = 0) \mid X, D = 0].$$

As long as the ex-post objective expectation of the subjective expectations is the same, the two parameters $\mathrm{E}(\alpha_t \mid X,\ D = 1)$ and $\mathrm{E}(\alpha_t \mid X)$ 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.,

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

(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

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

So

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

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 $\mathrm{E}(Y_{1t} - Y_{0t} \mid X)$ and the "treatment on the treated" parameter is the gain in the unobservables going from one state to the next:

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

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

Unlike the other estimators, the before–after estimators are non-robust to time effects that are common across participants and non-participants. The difference-in-differences estimators and the cross-section estimators are unbiased under different conditions. The cross-section estimators 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 non-participants. 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 an 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:

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

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]

Consider the permanent-transitory model of Eq. (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

[35] Due to the non-linearity of G, there are wealth effects in the decision to take training.

satisfied in a perfect foresight environment when there are credit constraints, or in an environment in which the U_{0t} can be partially forecast,[36] because for $t > k > t'$

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

even though

$$\mathrm{E}(U_{0t'} \mid X, D = 1) = 0$$

so

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

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-differences 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 (1999) 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 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 US CETA training parti-

[36] "Partially forecastable" means that some component of U_{0t} resides in the information set I_k. That is, letting $f(y \mid x)$ be the density of Y given X, $f(U_{0t} \mid I_k) \neq f(U_{0t})$ 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 $\mathrm{E}(\varphi(U_{0t}) \mid \mathrm{I}_k) \neq \mathrm{E}(\varphi(U_{0t}))$, 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.

cipants' 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 (1999) 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 program participation 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, 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 (1999).

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 contributions to labor market prospects: (a) it increases 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

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

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 non-market 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 non-market 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, 1985). Under our assumptions, being in the non-market 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

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

The first term, $(N - c_j)$, is the value of non-market 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 jth 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}:

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

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 ($1/1 + r$) 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 non-market 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

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

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

$$V_{ju} = \frac{(1 + r)(N - c_j) + \lambda_j \mathrm{E}_j(V_{je} \mid V_{je} > V_{ju})\Pr(Y_j > V_{ju}(r/(1 + r)))}{r + \lambda_j \Pr(Y_j > V_{ju}(r/(1 + r)))}.$$

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

$$\hat{j} = \underset{j}{\operatorname{argmax}} \{V_{ju}\}$$

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 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 non-linear in the basic parameters. Because of this non-linearity, 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% 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 (1999) that persons enter training as a consequence of displacement from both the market and non-market sector.

The full details of this model remain to be developed. 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 (6.2) 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 the 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 6.3. 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

$$\alpha/r - Y_{ik} - c_i > 0 \tag{6.15}$$

and the eligibility rule $Y_{i,k-1} < K$ 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 has been confined to the economically disadvantaged (defined by low family income over the past 6 months, participation in a cash welfare program or Food Stamps or being a foster child or disabled). There is also a 10% "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 US Survey of Income and Program Participation (SIPP). For most people, eligibility is determined by family earnings over the past 6 months. The mean monthly earnings of adult males appear in Fig. 1 aligned relative to month k, the month when eligibility is measured. The figure reveals a dip in the mean earnings of adult

[38] Eligibility 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 "underemployment" 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 6 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 6 months. In addition, 50% of youth participants had to have had some contact with the criminal justice system (Hollister et al., 1984).

male *eligibles* centered in the middle of the six month window over which family income is measured when determining JTPA eligibility.

Fig. 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 to 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 3 or 4 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 groups and on how administrative performance standards affect the allocation of these services. The main question that arises in these studies is the potential tradeoff 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 US 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 below in 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

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

[40] Devine and Heckman (1996) present certain non-stationary 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.

[41] A related issue involves differences in the types of services provided to different groups conditional on participation in a program. The US 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.

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 implications for efficiency depend on the empirical relationship between shortterm outcome levels and longterm 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, 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 arise 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 (1998d) 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 et al. (1996c) study a sample of applicants from one of the NJS training centers. They find that program staff at this training center do not cream-skim, 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

[42] See Heckman and Smith (1998c) 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 US training programs.

[43] Heckman and Smith (1998c) discuss this issue in greater depth. The discussion in the text presumes that the costs of training provided to different groups are roughly equal.

[44] Program 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.

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. Heckman et al. (1997c) 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 the hard-to-serve) is low. Similarly, Heckman et al. (1996c) 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 $\mathrm{E}(Y_1 - Y_0 \mid X, D = 1)$. 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 $f(y_0 \mid X, D = 1)$ and $f(y_1 \mid X, D = 1)$. From non-participants we can recover $f(y_0 \mid X, D = 0)$. It is the density $f(y_0 \mid X, D = 1)$ 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

$$\mathrm{E}(Y_0 \mid X, D = 1) = g_0(X) + \mathrm{E}(U_0 \mid X, D = 1)$$

and

$$\mathrm{E}(Y_1 \mid X, D = 1) = g_1(X) + \mathrm{E}(U_1 \mid X, D = 1)$$

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 non-experimental selec-

tion 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 does 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 non-experimental 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 data generated by 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.

7. Non-experimental evaluations

7.1. The problem of causal inference in non-experimental evaluations

Without invoking the very non-experimental methods they seek to avoid, social experiments cannot address many questions of interest to researchers and policymakers. Even if they could, such data are generally not available. As a result, analysts must rely on "observational" or non-experimental methods to address the problem of selection bias resulting from non-random participation of individuals in employment and training programs.

In an experimental evaluation, information from the control group is used to fill in missing counterfactual data for the treatments. As we have seen, under the assumptions specified in Section 5, an experiment is most successful in generating certain counterfactual means. In a non-experimental evaluation, analysts must replace these missing data with data on non-participants along with assumptions different from those invoked when using the method of social experiments.

To illustrate this point and to highlight an important distinction between experimental and non-experimental solutions to the evaluation problem, consider Fig. 7. It presents a model of potential outcomes in which each outcome takes on one of two possible values. For training participants, Y_1 equals one if the individual is employed after completing

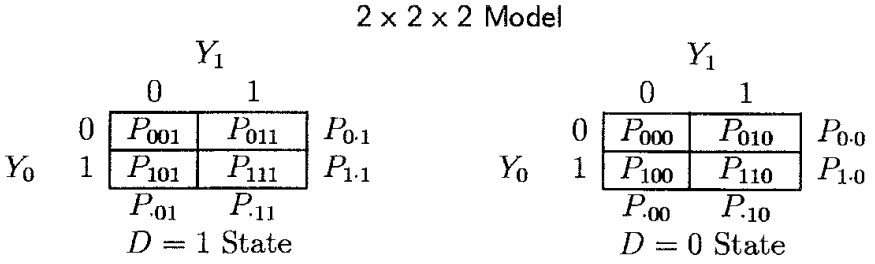

Fig. 7. $2 \times 2 \times 2$ model. Y_1 is an indicator variable for whether or not a person would be employed if trained; Y_0 is an indicator of employment without training. P_{abc} is the probability that $Y_0 = a$, $Y_1 = b$ and $D = c$.

training and equals zero otherwise. For non-participants Y_0 is defined similarly. As before, $D = 1$ for persons who select into training (but who may be excluded in an experimental evaluation) and $D = 0$ otherwise. When program evaluators have access to experimental data, they observe both Y_1 and Y_0 (but never both at the same time for the same person) for persons who select into training. That is, they observe the row and column totals for the $D = 1$ table, but not the proportion of persons for whom $D = 1$ who are in each individual cell. For example, the experimental controls enable the analyst to estimate the proportion of the persons selecting into training ($D = 1$) who would not have been employed in the absence of training, denoted $P_{0\cdot1}$, but not the proportion of persons selecting into training who would not have been employed either with or without training, denoted P_{001}. In order to estimate this proportion, we require another assumption, such as that training did not cause anyone to be non-employed who otherwise would have been employed. This "monotonicity" assumption (training can only make people better off), first invoked in Heckman and Smith (1993), allows us to set $P_{101} = 0$. In that case we can fill in the remaining elements of the table using the row and column totals. The proportion of trainees whose employment status changes as a result of training is now given by P_{011}. When the monotonicity assumption is imposed onto the data from experimental evaluations of training, P_{011} is typically relatively small (see, e.g., Heckman and Smith, 1993). Training causes a relatively small proportion of trainees to switch from the non-employment state to the employment state.

Analysts who have access only to non-experimental data observe only the column totals in the $D = 1$ table and the row totals in the $D = 0$ table. In addition, the proportion of people who take training is known. This can be determined from an experiment that randomizes eligibility but not from an experiment that randomizes among those who apply and are accepted into the program. The remaining elements of both tables, including the other row and column totals, are unknown. The task in observational studies is to find a set of conditioning variables and to impose an appropriate set of assumptions so that the row totals in the $D = 0$ table can be used to estimate the missing row totals in the $D = 1$ table. Regardless of the conditioning variables used or assumptions imposed, there always exists a set of minimal assumptions necessary to identify the impact of training that cannot be tested with the data. The same is true for the analysis of experimental data; the assumptions of no randomization bias or the unimportance of sample attrition cannot be

tested with the data typically generated from experimental evaluations. Both experimental and non-experimental approaches require assumptions that cannot be tested without collecting data specifically designed to test the assumptions of the model.

7.2. Constructing a comparison group

All evaluations are based on comparisons between treated and untreated persons. The comparisons may be constructed using the same persons in the treated and untreated states as in the before–after estimator. More commonly, different persons are compared.

The evaluation literature makes an artificial distinction between the task of creating a comparison group and the task of selecting an econometric estimator to apply to that comparison group. In truth, all estimators define an appropriate comparison group and the choice of a comparison group affects the properties of an estimator. The act of constructing or selecting a valid estimator entails assumptions about the samples on which it should be applied.

This simple point is usually overlooked in the empirical literature on program evaluation. It is common to observe analysts first constructing a comparison group on the intuitive principle of making the comparison group "comparable" in some way or other to the treatment group, and then to debate the choice of an estimator as if all estimators defined for random samples of the population can be applied to a comparison group so constructed. Many econometric estimators are only valid for random samples of the population. When non-random samples are generated, the estimators are sometimes no longer valid and have to be modified to account for the impact of the sampling rule used to generate the comparison samples.

The most common instance of this point arises in oversampling participants compared to non-participants. Program records are often abundant for participants; comparison samples often have to be collected at considerable cost. The ratio of program records to comparison group records is usually much larger than one. Simply pooling the two samples misrepresents the population proportion of persons taking training. In order to use the many conventional econometric methods that assume random sampling on such data, the samples have to be reweighted (see the discussion in Heckman and Robb, 1985a, 1986a). A special class of "control function" estimators that we define below does not have to be reweighted. However, instrumental variables estimators have to be reweighted in this case. Different classes of estimators exhibit different degrees of sensitivity to departures from random sampling in constructing comparison groups.

A second example is contamination bias, which we discuss in detail in Section 7.7. Many comparison groups include persons who have actually participated in the program but who have not been recorded as having done so. Again, estimators suitable to random samples without such measurement error on treatment status have to be modified for contaminated samples (Heckman and Robb, 1985a; Imbens and Lancaster, 1996).

A third example concerns the widespread practice of "matching" treatment and comparison group members on dimensions such as pre-program earnings. The literature

often distinguishes between "screening" on characteristics and matching. Screening usually refers to the application of certain broad rules (e.g., income below a certain level) to select observations from a source sample into a comparison sample; matching refers to alignment of trainees and comparison group members over narrower intervals. Both are a form of matching as we define it below and the distinction between them is of no practical value.

More serious are the consequences of this type of matching on the performance of econometric estimators. Matching on variables that are stochastically dependent on the errors of the model sometimes alters the stochastic structure of the errors. Econometric estimators that are valid for random samples can be invalid when applied to the samples generated by matching procedures.

To illustrate the foregoing point, consider the common-coefficient autoregressive estimator introduced into the econometric evaluation literature in Heckman and Wolpin (1976). Using decision rule (6.3) and assuming that agents make their decisions in an environment of perfect certainty and that enrollment into the program only occurs in period k,

$$Y_t = \beta + \alpha D + U_t, \qquad \text{for } t > k, \tag{7.1a}$$

$$Y_t = \beta + U_t, \qquad \text{for } t \leq k, \tag{7.1b}$$

$$U_t = \rho U_{t-1} + \varepsilon_t, \tag{7.1c}$$

where ε_t is an independently and identically distributed error with mean zero. In terms of the model of potential outcomes introduced in Section 3, $Y_t = DY_{1t} + (1 - D)Y_{0t}$ and $Y_{1t} - Y_{0t} = \alpha$, the parameter of interest. The model is in the form of Eq. (3.10) with an autoregressive error. The assumptions about the error terms are typically invoked about random samples of the population. Selection bias in this model arises because of the covariance between D and U_t. In a model with perfect capital markets, only if $\rho=0$ would there be no selection bias.[45]

If we have access to panel data, we can use two post-program observations to estimate α.[46] Write

$$Y_{t-1} = \beta + \alpha D + U_{t-1},$$

where $t - 1 > k$, so that

$$U_{t-1} = Y_{t-1} - \beta - \alpha D.$$

Substituting into (7.1c) and collecting terms, we may rewrite (7.1a) as

[45] However, this result crucially depends on the perfect capital market assumption as we noted in Section 6.3.4.

[46] As noted in Heckman and Robb (1985a, 1986a) and below, this estimator can also be applied to repeated cross-section data.

$$Y_t = \beta(1 - \rho) + \alpha(1 - \rho)D + \rho Y_{t-1} + \varepsilon_t. \tag{7.2}$$

Under decision rule (6.3), D is orthogonal to ε_t even though agents are making their participation decisions under perfect certainty. Least squares applied to (7.2) identifies ρ, and hence α and β. This estimator can be applied to training programs or schooling. Its great advantage is that it can be implemented using only post-program outcome measures provided $\rho \neq 1$. Properties of this estimator are presented in Section 7.6.

Another way to identify α is to use instrumental variables or classic selection bias estimators which we describe in detail below. Assuming random sampling, both of these estimators identify α.

Suppose, however, that we first "match" on pre-training earnings, $Y_{t'}$, $t' < k$, in order to construct a comparison sample of non-participants. Consider a simple screening rule: select observations into the sample if $Y_{t'} < l$. This rule is widely used in constructing comparison samples. How are the error structure (7.1c) and the properties of the three estimators just discussed affected by the application of such screening rules? The auto-regressive estimator just presented using post-program observations is unaffected by these sampling rules. It continues to identify α and ρ. This is immediately seen because $\mathrm{E}(\varepsilon_t \mid D = 1, Y_{t-1}, Y_{t'} < l) = 0$ since ε_t is independent of $Y_{t'}, t' < k$, and Y_k.

However, matching affects the distribution of the errors. This makes a sample selection model based on a distributional assumption appropriate to a random sample inappropriate when applied to a matched sample. In this case, two selection rules generate the outcomes; classical selection estimators that only account for agent self-selection do not account for the selection bias induced by the analysts' matching procedure. Instrumental variables methods appropriate to random samples in general become inconsistent when applied to matched samples for reasons exposited in Section 7.7.

Another strategy for defining a comparison group is to use program applicants who drop out of the application and enrollment process before receiving training. Such comparison groups include persons who applied and were rejected from the program, those who were admitted but never showed up for training ("no-shows"), or early program dropouts. (No-shows are used in, e.g., Cooley et al., 1979; LaLonde, 1984; Bell et al., 1995; Heckman et al., 1997a). In samples based on no-shows, two decision rules – whether or not to apply to the program and whether or not to stay in the program if accepted – determine which non-participants end up in the comparison group sample. The properties of econometric estimators have to be examined to see if they are robust to such sample selection rules. Analytically, this is the same problem as arises in the construction of matched samples, except that in this case the decision rules of agents govern the construction of samples. Estimators valid for samples generated by one decision rule need not be valid for another.

A brief summary of the screening and matching criteria used in several major evaluations is presented in the last row of Tables 5 and 6. Table 7, based on Barnow (1987), presents a more exhaustive list of characteristics used to match and control for differences in evaluations of the US CETA program, the immediate predecessor of JTPA. Combining

matching and different non-experimental evaluation methods that break down when applied to matched samples constitutes an important source of variability across these studies, one that has more to do with the properties of the estimators selected than with the properties of the programs being studied.

In the literature, the act of specifying a comparison group and then making conditional mean comparisons between participants and comparisons is equivalent to defining a matching estimator. The matching estimator may be embellished by further adjustments as we note below. A different comparison group might be specified for each treatment observation. The potential sample from which the comparison group is taken includes all persons who do not take treatment. Further restrictions on this universe define different matching rules.

7.3. Econometric evaluation estimators

All evaluation estimators are based on the three basic estimation principles introduced in Section 4. They entail making some comparison of treated individuals with the untreated. The comparison may be between treated and untreated persons at a point in time as in the cross-section estimator; it may be between the same persons in the treated and untreated states as in the before–after estimator; or it may be a hybrid of the two principles as in the difference-in-differences estimator. In this section, we extend these basic estimators to allow for conditioning variables and to exploit knowledge of the serial correlation properties of error terms.

The estimators within each class differ in the way they adjust, condition or transform the data in order to construct the counterfactual $\mathrm{E}(Y_{0t} \mid X, D = 1)$. Throughout the rest of this section, we consider how the various estimators construct the counterfactual and what assumptions they make about individual decision processes that determine program participation. We motivate this discussion using the simple decision and outcome models of Section 6.3. The first class of estimators that we consider are cross-section estimators based on matching methods. These estimators are frequently used in studies by consulting firms because they are relatively easy to explain to their clients. A disadvantage of this approach is that it requires strong underlying assumptions about the selection process into training. Although the method is usually applied in a cross-sectional setting, matching can be generalized to apply to panel settings as in Heckman et al. (1997a, 1998c). The second class of cross-section methods we consider are selection bias correction methods developed in Heckman (1976, 1979) or Heckman and Robb (1985a, 1986a). This approach is often used in studies of European training programs. It too can be extended to apply to panel data, but is most frequently applied in a cross-sectional setting.

Program evaluations by academic labor economists in the United States have relied almost exclusively on a third class of estimators: longitudinal methods that extend the before–after and difference-in-differences estimators. An implicit belief shared by the authors of these studies is that longitudinal methods are more robust than cross-section selection bias correction methods, which are sometimes dismissed as being "functional

Table 5
Explanatory variables used in previous studies[a]

	MDTA data		CLMS-based studies[b]	
	Ashenfelter (1978)	Ashenfelter and Card (1985)	Dickinson et al. (1987)	Bryant and Rupp (1987)
Program, year, outcome variable	MDTA classroom trainees enrolled in first 3 months of 1964; 1965–1969 annual social security earnings	1976 CETA trainees. 1977 and 1978 annual social security earnings	CETA trainees enrolled in 1976; 1978 annual social security earnings	Two cohorts of CETA trainees; 1977 and 1978 annual social security earnings
Local labor market information	No	No	No	No
Age, race, sex	Yes	Yes	Yes	Yes
Education	No	Yes	Yes	Yes
Training history	No	No	No	No
Children	No	No	No	Yes
Employment histories	No	No	Yes (recent)	Yes (recent)
Hours worked	No	Yes	Yes	Yes
Unemployment histories	No	No	Yes (recent)	Yes (recent)
Welfare receipt	No	No	Yes	No

Earnings histories	5 years pre-program 5 years post-program	2 years pre-program 2 years post-program	Same as Ashenfelter and Card (1985)	4 years pre-enrollment
Matching criteria[c]	None specified	(a) 1975 earnings $\leq$ \$20K and household income $\leq$ \$30K (b) In labor force in March, 1976 (c) Age greater than 20	Matching based on a metric over a vector of predictors of 1978 earnings including lagged earnings (1970–1975), unemployment in 1975, worked in public sector, in labor force in March 1976 and demographics.	Matching on 1976 earnings, change in earnings 1975–1976, 1975 labor force status, family income, in labor force in 1975 or at March 1976 interview and demographics.

[a] Source: Heckman et al. (1997a, Table 1). Notes: "Used" means employed in the matching process and/or included in the outcome equation.

[b] CLMS, CETA Longitudinal Manpower Survey. The CLMS data matched Social Security longitudinal earnings records to CETA participants and CPS comparison group members from the March 1976 and 1977 CPS. All of the CLMS-based studies use the social security earnings data except for Bassi (1983), which also uses the CPS earnings data. All of the personal and family information available in the CPS, including shortterm employment and labor force participation histories, are available in the CLMS but not necessarily used in the analyses based upon it.

[c] "Matching Criteria" indicate the criteria for membership in the comparison group. This is sometimes referred to as "screening" in the literature.

Table 6
Explanatory variables used in previous studies[a]

	CLMS-based studies[b]	NSW (Supported Work) data		NJS data
	Bassi (1983, 1984)	LaLonde (1986)	Fraker and Maynard (1987)	Heckman et al. (1998b)
Program and outcome variable	CETA, 1977 and 1978 annual social security earnings	NSW, 1979 annual social security and PSID survey earnings	NSW, 1977, 1978 and 1979 annual earnings for AFDC women and youth	NJS, monthly survey earnings in the 18 months after random assignment or measured eligibility
Local labor market information	No	No	No	Yes
Age, race, sex	Yes	Yes	Yes	Yes
Education	?[d]	Yes	Yes	Yes
Training history	No	No	No	Yes (partial)
Children	?	Yes	Yes	Yes
Employment histories	?	No	No	Yes
Hours worked	?	No	No	Yes
Unemployment histories	?	No	No	Yes (1 year)
Welfare receipt	?	Yes	Yes	Yes (1 year)
Earnings histories	Same as Bryant and Rupp (1987)	Two years pre-program Two years post-program	Two years pre-program Two years post-program	Five years pre-program Two years post-program

Matching criteria[c] (criteria for membership in comparison sample)	Same as Bryant and Rupp (1987); also uses a random sample from the March 1976 CPS	PSID: household head in 1975–1979; CPS: March 1976 CPS earnings matched to SSA earnings; screens based on 1976 personal and household income	Three samples: I: eligible in sample period; screen out in-school youth; AFDC women match on age of youngest child and welfare receipt II: eligible in sample period; cell matching based on predictors of 1979 SSA earnings including prior earnings, change in earnings, education, family income and demographics III: eligible in sample period; match on earnings estimated on eligible non-participant sample, age and sex	Within age and sex groups, match on propensity score based on site, race, age, schooling, marital status labor force status history, number of recent jobs, training history, household size and recent earnings.

[a] Source: Heckman et al. (1997a, Table 1). Notes: "Used" means employed in the matching process and/or included in the outcome equation.

[b] CLMS, CETA Longitudinal Manpower Survey. The CLMS data matched Social Security longitudinal earnings records to CETA participants and CPS comparison group members from the March 1976 and 1977 CPS. All of the CLMS-based studies use the social security earnings data except for Bassi (1983), which also uses the CPS earnings data. All of the personal and family information available in the CPS, including shortterm employment and labor force participation histories, are available in the CLMS but not necessarily used in the analyses based upon it.

[c] "Matching Criteria" indicate the criteria for membership in the comparison group. This is sometimes referred to as "screening" in the literature.

[d] "?" indicates that the study does not specify the variables used.

Table 7
Matching characteristics used in CETA evaluation studies[a]

	Westat (1981)	Westat (1984)	Bassi (1983)	Bassi et al. (1984)	Bloom and McLaughin (1982)	Dickinson et al. (1984)	Geraci (1984)
Program entry	7/75–6/76	7/75–6/76 (A) 7/75–6/77 (B)	7/75–6/76	7/76–9/77	1/75–6/76	1/76–12/76	7/75–7/76
Post-program period	1977	1977 (A) 1978 (B)	1977, 1978	1978, 1979	1976, 1977, 1978	1978	1977–1979 average
CETA participants included in analysis	Ages 14–16 and enrolled in CT, PSE, OJT, WE, or multiple training types; over 7 days in program; prior year earnings less than 20,000; prior year family income less than 30,000; terminated from program by 12/76; valid SSA match on 3 of 5 criteria	Same as Westat (1981) except family income excludes participant's earnings	Same as Westat (1981)	Welfare recipients and other economically disadvantaged persons ages 18–65 and youth ages 13–22; no other restrictions	Ages 25–60 and enrolled in CT, OJT, or WE only. Must have over 7 days in program	Ages 16–64 and not in summer youth program; must be complete or close SSA match and not in program in 1978	Same as Westat (1981) except only persons over age 22
CPS individuals eligible for comparison group	Same age, earnings, income and SSA match; in labor force in 3/76 or worked in 1975	Same as Westat (1981).	Same as Westat (1981).	For ages 18–65, must be on welfare or economically disadvantaged; for youths 13–22, same as Westat (1984b) (B)	Ages 25–60; earned less than SSA maximum from 70–75; 1975 family income $<$ $30,000	Adults in labor force in 3/76; youth in labor force in 3/76 or worked in 1975	Same as Westat (1984) (A) group

Matching procedure	Cell matching for 1972–1974 earnings groups; for low earners: exact match on sex, race, and age; collapsing permitted on education, family income, labor force experience, family head status, 1975 SSA earnings, change in SSA earnings 74–75, change in SSA earnings 73–74, poverty status and private sector employment; for intermediate earners: exact match on sex, race, 1975 SSA earnings, change in SSA earnings; collapsing permitted on other variables: for high earners: same as intermediate earners except family income	Cell matching for each activity; exact match on sex, 1975 SSA earnings for (A) or 1976 SSA earnings for (B), change in SSA earnings for two previous years (73–74 and 74–75 or 74–75 and 75–76) and race; collapsing permitted on age, education, family income, prior year labor force experience, family head status, and poverty status	Same as Westat (1981)	All economically disadvantaged persons and welfare recipients ages 18–65 included in adult study. For youth ages 13–22, used Westat (1981) youth matched comparison groups	All CPS individuals who met the above criteria were included	Weighted nearest-neighbor match based on SSA earnings in 1970–1975, square of 1975 SSA earnings, race and ethnicity, age, age^2, age^3, family head status, occupational categories, public sector employment, poverty status, AFDC receipt, UI receipt, percent of time worked in 1975, percent of time worked in 1974, CPS reported earnings for those at SSA maximum, and 15 interaction variables; match groups formed overall and by activity	Same as Westat (1981) (A)

Table 7 (*continued*)

	Westat (1981)	Westat (1984)	Bassi (1983)	Bassi et al. (1984)	Bloom and McLaughin (1982)	Dickinson et al. (1984)	Geraci (1984)
	given less priority in cell collapsing						
Regression procedure	Weighted least squares with separate regressions for each race-sex-earnings group	Weighted least squares with separate regressions for each activity	Difference-in-differences with separate regressions by sex-race group	Difference-in-differences with separate regressions by sex-race-welfare group	OLS with fixed effects, individual time trends and correction for earnings drop for participants	Ordinary least squares with separate regressions by age-sex-activity group	Two-step procedure: (1) probit for positive earnings; (2) weighted least squares for positive earners with separate analysis by sex
Regressors	Family head status, education, prior work in private sector, 1973 SSA earnings, 1974 SSA earnings, proxy for cyclical unemployment, family income, prior labor force status, age, educational disadvantage and status (ages 16–18 only) veteran	Same as Westat (1981)	Age, age^2	Age, age^2	Age, age^2, education, education^2, family size, minority status, head of household status, current marital status, past marital status, presence of children under age 4, presence of children ages 7–18	Same regressors as used for matching	Age, age^2, education, marital status, head of household status, economically disadvantaged status, minority status, presence of children under age 6 (females only), presence of children ages 6–18 (females only), interaction terms for experience

status (males only), presence of children under age 6 (females only), presence of children ages 6–18 (females only)		and education

[a] Source: Barnow (1987, Table 2).

form dependent." However, we demonstrate below that as currently utilized in the applied evaluation literature, longitudinal estimators depend on functional form assumptions. Moreover, longitudinal estimators are often much less robust to choice-based sampling and other matching and screening procedures used to produce comparison samples in the empirical literature than are cross-section sample selection estimators. In the remainder of this section, we discuss the identifying assumptions that underlie the main methods used in evaluation research, and sketch out how they are implemented to produce practical estimators.

We remind the reader that throughout this chapter, we use X variables that are not determined by D. Letting X be the vector of conditioning variables and Y^P a vector of potential outcomes, we write $Y_t^P = (Y_{0t}, Y_{1t})$, and $Y^P = (Y_1^P, \ldots, Y_T^P)$, $X = (X_1, \ldots, X_T)$. We define the admissible X on which we condition to define parameters as those X that satisfy

$$f(X \mid D, Y^P) = f(X \mid Y^P). \tag{7.A.1}$$

where $f(X \mid D, Y^P)$ is the density of X given D and Y^P and $f(X \mid Y^P)$ is the density of X given Y^P.[47] This assumption says that given the potential outcomes in both states, the actual occurrence of D provides no more information on X ("Does not cause X"). We maintain this assumption in order to avoid masking the effects of D on outcomes by conditioning on variables that are determined by D. Other definitions are possible but we maintain this one to make our analysis interpretable and to avoid certain technical problems in making forecasts with our parameters. Heckman (1998a) presents a more extensive discussion of this condition and relates it to definitions of causality and exogeneity in the econometric time series literature.

7.4. Identification assumptions for cross-section estimators

When participation in training is voluntary, and evaluators have access to cross-sectional data, they can construct the distribution of outcomes for participants, $F(Y_1 \mid X, D = 1)$, and for non-participants, $F(Y_0 \mid X, D = 0)$. They use $F(Y_0 \mid X, D = 0)$ to approximate $F(Y_0 \mid X, D = 1)$, which runs the risk of selection bias. When using this approximation, the bias in estimating $\mathrm{E}(Y_1 - Y_0 \mid X, D = 1)$ is given by

$$B(X) = \mathrm{E}(Y_0 \mid X, D = 1) - \mathrm{E}(Y_0 \mid X, D = 0). \tag{7.3}$$

Many schemes have been proposed to circumvent this bias. We begin by considering the intuitively-appealing method of matching.

7.4.1. The method of matching

The method of matching assumes that analysts have access to a set of conditioning variables, X, such that, within each "stratum" defined by X, the counterfactual outcome distribution of the participants is the same as the observed outcome distribution of the non-

[47] Heckman and Borjas (1980) develop this non-causality condition.

participants.[48] The statistical matching literature assumes access to a set of X variables such that

$$(Y_0, Y_1) \perp\!\!\!\perp D \mid X, \tag{7.4}$$

where "$\perp\!\!\!\perp$" denotes independence and X denotes variables on which conditioning is conducted. As a consequence of (7.4), the distributions of outcomes $F(Y_0 \mid D = 1, X) = F(Y_0 \mid D = 0, X) = F(Y_0 \mid X)$ and $F(Y_1 \mid D = 1, X) = F(Y_1 \mid D = 0, X) = F(Y_1 \mid X)$. The method appeals to the intuitive principle that it is possible to "adjust away" differences between participants and non-participants using the available regressors.

If assumption (7.4) is valid we can use non-participants to measure what participants would have earned had they not participated, provided we condition on the variables X. To ensure that this assumption has empirical content, it also is necessary to assume that there are participants and non-participants for each X for which we seek to make a comparison. More formally, this means that

$$0 < \Pr(D = 1 \mid X) < 1 \tag{7.5}$$

over the set of X values where we seek to make a comparison. To satisfy this condition, at least in large samples, there must be both participants and non-participants for each X. In a finite sample of any size, we replace this condition by the empirical probability.[49] This condition ensures that the distributions in (7.4) are defined for all X that satisfy it. As we demonstrate below in Section 8.2, this assumption has important practical consequences for training evaluations. Failure to satisfy this condition appears to be one major reason why matching methods produce biased estimates of the impact of training in the NJS study. The treatment parameter $\mathrm{E}(Y_1 - Y_0 \mid X, D = 1)$ cannot be identified for values of X where (7.5) is violated.

Under assumptions (7.4) and (7.5), matching produces a comparison group that resembles an experimental control group in one key respect: conditional on X, the distribution of the counterfactual outcome, Y_0, for the participants is the same as the observed distribution of Y_0 for the comparison group. In particular, as long as the means exist, assumptions (7.4) and (7.5) imply that

$$\mathrm{E}(Y_0 \mid X, D = 1) = \mathrm{E}(Y_0 \mid X, D = 0), \tag{7.6a}$$

and that

$$\mathrm{E}(Y_1 \mid X, D = 1) = \mathrm{E}(Y_1 \mid X, D = 0). \tag{7.6b}$$

Therefore, for each point in X, the bias $B(X) = 0$. However, this assumption does not imply no selection bias, i.e., that $\mathrm{E}(U_0 \mid X, D = 1) = 0$. Instead, like experiments, match-

[48] The first published instance of the use of this method of which we are aware is Fechner (1860).

[49] The support of X consists of those values of X with positive density. Assumptions (7.4) and (7.5) are called "strong ignorability" by Rosenbaum and Rubin (1983).

ing balances the bias:

$$\mathrm{E}(U_0 \mid X, D = 1) = \mathrm{E}(U_0 \mid X, D = 0) = \mathrm{E}(U_0 \mid X). \tag{7.7}$$

In an ideal experiment, we obtain a comparison group via randomization among persons for whom $D = 1$. Matching emulates an experiment by replacing randomization with conditioning on a set of X variables. Conditional on those values, persons randomly select into the program. There are no selective differences in Y_0 outcomes between participants and non-participants given X. Randomization at the stage where persons enter the program also may be thought of as a form of conditioning (Heckman, 1996). It operates conditional on $D = 1$. Under the conditions that justify it, randomization generates a control group for each X in the participant population. Similarly, under assumption (7.4), matching generates a comparison group, but only for these X values that satisfy (7.5), which in practice is often a much smaller set of values than would be the case with randomization.

In Section 8.2 below, we draw on the work of Heckman et al. (1998b) and demonstrate that the reduction in the set of X for which the parameter of interest is defined can be substantial. Further, because the impact parameter may depend on X, the parameter estimated by an experimental evaluation and the parameter estimated by matching may be different.

When the Rosenbaum–Rubin assumptions (7.4) and (7.5) are invoked, it is possible to construct both the "treatment on the treated" parameter $\mathrm{E}(Y_1 - Y_0 \mid X, D = 1)$ and the effect of "non-treatment on the non-treated" $\mathrm{E}(Y_0 - Y_1 \mid X, D = 0)$. Only assumption (7.6a) is required if we are interested in the mean effect of treatment on the treated. It permits agents to select into the program on the basis of U_1 but not U_0. Assuming that $\mathrm{E}(U_0 \mid X) = 0$, it implicitly defines the parameter "treatment on the treated" in an asymmetric way:

$$\mathrm{E}(Y_1 - Y_0 \mid X, D = 1) = \mu_1(X) - \mu_0(X) + \mathrm{E}(U_1 \mid X, D = 1)$$

because $\mathrm{E}(U_0 \mid X, D = 1) = \mathrm{E}(U_0 \mid X) = 0$. This parameter no longer equals the effect of treatment on a randomly selected person as it would if (7.4) held. Assumption (7.6b) allows us to identify the mean effect of non-treatment on the non-treated.

Using representation (3.1a) and (3.1b), (7.4) and (7.5) imply that $\mathrm{E}(U_0 \mid X, D = 1) = \mathrm{E}(U_0 \mid X, D = 0) = \mathrm{E}(U_0 \mid X) = 0$ and $\mathrm{E}(U_1 \mid X, D = 1) = \mathrm{E}(U_1 \mid X, D = 0) = \mathrm{E}(U_1 \mid X) = 0$. Thus conditioning on X, the two parameters "treatment on the treated" and "the effect of randomly assigning a person with characteristics X to the program" are the same.[50] From an economic standpoint, assumption (7.4) rules out selection into the program on the basis of unobservables (U_0, U_1) that may be partially known to people taking training but are unknown to the observing economist. In terms of the random coefficient model of Section 3, it rules out correlation between D and the difference in unobserved components, $(U_1 - U_0)$. It defines an implicit economic model that assumes that agents do not enter the program on the basis of gains unobserved by analysts. Thus it is

[50] This is also true if $Y_1 = g_1(X) + U_1$ and $Y_0 = g_0(X) + U_0$ and $\mathrm{E}(U_1 \mid X) \neq 0$ and $\mathrm{E}(U_0 \mid X) \neq 0$. In that case, $\mathrm{E}(Y_1 - Y_0 \mid X, D = 1) = g_1(X) - g_0(X) + \mathrm{E}(U_1 - U_0 \mid X)$ so that the two parameters are the same.

a method congenial with the assumption that α in (6.3) is a common coefficient, or that if α varies among persons with identical X, then participation in the program is not based on this variation. In the context of that model, the "cost of participation" or any of the variables generating participation, but not outcomes, are valid conditioning variables. Thus, if the costs of participation are distributed independently of all other variables and if Y_{0k} is independent of $Y_{0\ell}$, then conditioning on c or Y_{0k} will satisfy the conditions required to justify the matching estimator. However, as we explained in Section 6.3.1, if we condition on both cost of participation and Y_{0k}, we violate condition (7.5). Matching breaks down if there is too much information and other methods must be used to evaluate the program.[51]

To operationalize the method of matching, assume two samples: "t" for treatment and "c" for comparison group. Unless otherwise noted, observations are statistically independent. Simple matching methods are based on the following idea: For each person i in the treatment group, we find some group of "comparable" persons. The same individual may be in both groups if that person is treated at one time and untreated at another. We denote outcomes in the treatment group by Y_i^t and we match these to the outcomes of a subsample of persons in the comparison group to estimate a treatment effect. In principle, we can use a different subsample as a comparison group for each person.

In practice, we can construct matches on the basis of a neighborhood $C(X_i)$, where X_i is a vector of characteristics for person i. Neighbors to treated person i are persons in the comparison sample whose characteristics are in neighborhood $C(X_i)$. Suppose that there are N_c persons in the comparison sample and N_t in the treatment sample. Thus the persons in the comparison sample who are neighbors to i, are persons j for whom $X_j \in C(X_i)$, i.e., the set of persons $A_i = \{j \mid X_j \in C(X_i)\}$. Let $W(i,j)$ be the weight placed on observation j in forming a comparison with observation i and further assume that the weights sum to one,

$$\sum_{j=1}^{N_c} W(i,j) = 1,$$

and that $0 \leq W(i,j) \leq 1$. Then we form a weighted comparison group mean for person i, given by

$$\bar{Y}_i^c = \sum_{j=1}^{N_c} W(i,j) Y_j^c, \tag{7.8}$$

and the estimated treatment effect for person i is $Y_i - \bar{Y}_i^c$.

Heckman et al. (1997a) survey a variety of alternative matching schemes proposed in the literature. Here we briefly introduce two widely used methods. The nearest-neighbor

[51] The regression discontinuity design estimator discussed in Section 7.4.6 can be applied here as a limit form of the matching estimator that identifies $E(Y_1 - Y_0 \mid X, D = 1)$ at one point.

matching estimator defines A_i such that only one j is selected so that it is closest to X_i in some metric:

$$A_i = \{j \mid \underset{j \in \{1,\ldots,N_c\}}{\text{Min}} \| X_i - X_j \| \},$$

where $\| \ \|$ is a metric measuring distance in the X characteristics space. The Mahalanobis metric is one widely used metric for implementing the nearest-neighbor matching estimator. The metric used to define neighborhoods for i is

$$\| \ \| = (X_i - X_j)' \Sigma_c^{-1} (X_i - X_j),$$

where Σ_c is the covariance matrix in the comparison sample. The weighting scheme for the nearest neighbor matching estimator is

$$W(i,j) = \begin{cases} 1 & \text{if } j \in A_i, \\ 0 & \text{otherwise.} \end{cases}$$

A version of nearest-neighbor matching, called "caliper" matching (Cochran and Rubin, 1973), makes matches to person i only if

$$\| X_i - X_j \| < \varepsilon,$$

where ε is a pre-specified tolerance. Otherwise person i is bypassed and no match is made to him or her.

Kernel matching uses the entire comparison sample, so that $A_i = \{1, \ldots, N_c\}$, and sets

$$W(i,j) = \frac{K(X_j - X_i)}{\sum_{j=1}^{N_c} K(X_j - X_i)}, \tag{7.9}$$

where K is a kernel. In practice, kernels are typically a standard distribution function such as that for the normal. Kernel matching is a smooth method that reuses and weights the comparison group sample observations differently for each person i in the treatment group with a different X_i. Kernel matching can be defined pointwise at each sample point X_i or for broader intervals.

The impact of treatment on the treated is estimated by forming the mean difference across the i

$$m = \frac{1}{N_t} \sum_{i=1}^{N_t} (Y_i^t - \bar{Y}_i^c) = \frac{1}{N_t} \sum_{i=1}^{N_t} \left(Y_i^t - \sum_{j=1}^{N_c} W(i,j) Y_j^c\right). \tag{7.10}$$

We can define this mean for various subsets of the treatment sample defined in various ways. More efficient estimators weight the observations accounting for the variance (Heckman et al., 1997a, 1998c; Heckman, 1998a).

Regression-adjusted matching, proposed by Rubin (1979) and clarified in Heckman et

al. (1997a, 1998c), uses regression-adjusted Y_i, denoted by $A(Y_i) = Y_i - X_i\beta$, in place of Y_i in the preceding calculations. (See the cited papers for the econometric details of the procedure). Regression-adjusted matching methods were widely used in the controversial CETA evaluations conducted in the early 1980s, which we discuss below.

The essence of the idea justifying matching is that conditioning on X eliminates selection bias. Like social experiments, the method requires no functional form assumptions for outcome equations. If, however, a functional form assumption is maintained, as in the econometric procedure proposed by Barnow et al. (1980), it is possible to implement the matching assumption using regression analysis. Suppose that Y_0 is linearly related to observables X and an unobservable U_0, so that $\mathrm{E}(Y_0 \mid X,\ D = 0) = X\beta + \mathrm{E}(U_0 \mid X,\ D = 0)$, and $\mathrm{E}(U_0 \mid X,\ D = 0) = \mathrm{E}(U_0 \mid X)$ is linear in X. Under these assumptions, controlling for X via linear regression allows one to identify $\mathrm{E}(Y_0 \mid X,\ D = 1)$ from the data on non-participants. Such functional form assumptions are not strictly required to implement the method of matching. Moreover, in practice, users of the method of Barnow et al. (1980) do not impose the common support condition (7.5) for the distribution of X when generating estimates of the training effect. The distribution of X may be very different in the trainee ($D = 1$) and comparison group ($D = 0$) samples, so that comparability is only achieved by imposing linearity in the parameters and extrapolating over different regions.

One advantage of the method of Barnow et al. (1980) is that it uses data parsimoniously. If the X are high dimensional, the number of observations in each cell when matching can get very small. Another solution to this problem that reduces the dimension of the matching problem without imposing arbitrary linearity assumptions is based on the probability of participation or the "propensity score," $P(X) = \Pr(D = 1 \mid X)$. Rosenbaum and Rubin (1983) demonstrate that if assumptions (7.4) and (7.5) hold, then

$$(Y_1, Y_0) \perp\!\!\!\perp D \mid P(X) \text{ for } X \in \chi_c, \tag{7.11}$$

for some set χ_c where it is assumed that (7.5) holds in the set. Conditioning on $P(X)$ rather than on X produces conditional independence. Condition (7.11) has the important implication that to construct the desired counterfactual conditional mean $\mathrm{E}(Y_0 \mid P(X), D = 1)$, we require only that

$$B(P(X)) = \mathrm{E}(Y_0 \mid P(X), D = 1) - \mathrm{E}(Y_0 \mid P(X), D = 0) = 0. \tag{7.12}$$

We also could invoke (7.12) in place of (7.11) to define the conditions required to justify matching to estimate mean impacts. Conditioning on $P(X)$ sets $B(P(X)) = 0$ and reduces the dimension of the matching problem down to matching on the scalar $P(X)$. The analysis of Rosenbaum and Rubin (1983) assumes that $P(X)$ is known rather than estimated. Heckman et al. (1998c) present the asymptotic distribution theory for the kernel matching estimator in the cases in which $P(X)$ is known and in which it is estimated both parametrically and non-parametrically. They also answer the question, "If $P(X)$ were known would we match on it or on X?" Using the variance of the estimated average impacts as the choice criterion, the answer is "it depends".

A major advantage of the method of randomized trials over the method of matching is

that randomization works for any choice of X. In the method of matching, there is the same uncertainty about which X to use as there is in the specification of conventional econometric models. Even if one set of X values satisfies condition (7.11) or (7.12), an augmented or reduced version of this set may not. Heckman et al. (1997a) discuss tests that can help determine the appropriate choice of X variables. Any convincing application of the method of matching requires a demonstration that an adequate model for $P(X)$ has been selected. Heckman et al. (1998b) discuss this problem in depth. In the statistics literature, there is no discussion of the choice of X. Implicitly, the advice given there is to use all available regressors. One general rule, already noted in the introduction to this section, is to include in X only variables that are not caused by D given the unobservables. Intuitively, conditioning on variables caused by D masks the true effect of D on outcomes.

The method of matching is sometimes used to estimate $\mathrm{E}(Y_1 - Y_0 \mid X,\ D = 1)$ at points of $X = x$. More commonly, an averaged version of this parameter is estimated over a set $S(X)$:

$$\mathrm{E}(Y_1 - Y_0 \mid D = 1) = \frac{\int_{S(X)} \mathrm{E}(Y_1 - Y_0 \mid X, D = 1)\mathrm{d}F(X \mid D = 1)}{\int_{S(X)} \mathrm{d}F(X \mid D = 1)}. \tag{7.13}$$

The distinction between the average parameter and the pointwise parameter is an important one. Even though the behavioral motivation and the identifying assumptions are different, it turns out that both the matching estimator and the classical selection estimator can identify (7.13) under very different behavioral assumptions. We now turn to consider the classical selection estimator.

7.4.2. Index sufficient methods and the classical econometric selection model

The most troubling feature of the method of matching is the assumption that selection into a program does not occur on the basis of unobservable (to the economist) gains from the program (U_0 if (7.6a) is assumed; $U_1 - U_0$ if (7.4) is assumed). Depending on the quality of the data at the analyst's disposal, it may or may not be attractive to assume that the analyst knows as much as the people being studied. The method of matching is not robust to violations of this assumption.

The traditional econometric approach to the selection problem adopts a more conservative approach and allows for selection on unobservables. As currently formulated, it assumes an additively separable model relating outcomes to regressors and additive errors, but does not require the strong behavioral assumptions that justify matching. Thus it trades a behavioral assumption for an additive separability assumption. It allows for selection into the program on the basis of unobserved components of outcomes. This approach is in the spirit of much econometric work that builds models to estimate a variety of counterfactual states, rather than just the single counterfactual state required to estimate the mean impact of treatment on the treated, which is the parameter of interest in most evaluations based on the methods of matching or random assignment.

In the simplest econometric approach, two functions are postulated: $Y_1 = g_1(X, U_1)$ and

$Y_0 = g_0(X, U_0)$, where U_0 and U_1 are unobservables. A selection equation is specified to determine which outcome is observed. Separability between X and (U_0,U_1) is assumed, so that

$$Y_1 = g_1(X) + U_1 \qquad \text{and} \qquad Y_0 = g_0(X) + U_0,$$

where for simplicity we assume that $\mathrm{E}(U_1 \mid X) = \mathrm{E}(U_0 \mid X) = 0$ so that $g_1(X) = \mu_1(X)$ and $g_0(X) = \mu_0(X)$. These exogeneity assumptions are not strictly required but for simplicity we maintain them.[52] This assumption *defines* functions called "structural equations" that do not depend on unobserved variables. In this notation, the treatment on the treated parameter is

$$\mathrm{E}(Y_1 - Y_0 \mid X, D = 1) = g_1(X) - g_0(X) + \mathrm{E}(U_1 - U_0 \mid X, D = 1),$$

which combines "structure" and "error" in a somewhat unusual way.

Much applied econometric work is devoted to eliminating the mean effect of unobservables on estimates of functions like g_0 and g_1. However, as previously noted, the mean difference in unobservables is an essential component of the definition of the parameter of interest in evaluating social programs. In the conventional framework, the selection bias that arises from using a non-experimental comparison group is given by

$$B(X) = \mathrm{E}(U_0 \mid X, D = 1) - \mathrm{E}(U_0 \mid X, D = 0).$$

In the standard evaluation problem, the goal is to set $B(X) = 0$, *not* to eliminate dependence between (U_0,U_1) and X and D.

The conventional econometric approach for addressing selection bias partitions the observed variables X into two not necessarily disjoint sets (Q,Z) corresponding to those variables in the outcome equations and those variables in the participation equation, and then postulates exclusion restrictions. It assumes that certain variables appear in Z but not in Q. The conventional approach further restricts the model so that the bias $B(X)$ only depends on Z through a scalar index. Recall that such exclusion restrictions are not required to justify matching as an estimator.[53]

The latent index model of program participation introduced in Section 6 motivates the characterization of selection bias as a function of a scalar index. In that model, we defined the index $IN = H(Z) - V$, where $H(Z)$ is the mean difference in utilities or discounted earnings between the training and non-training states, and V is assumed to be independent of Z. The training indicator, D, then equals one when $IN > 0$ and equals zero otherwise, resulting in $\Pr(D = 1 \mid Z) = F_V(H(Z))$. The conventional econometric selection model further assumes that the dependence of D and the unobservables U_0 and U_1 arises only

[52] Thus we could instead postulate instruments Z such that $\mathrm{E}(U_1 \mid X) \neq 0$ and $\mathrm{E}(U_0 \mid X) \neq 0$ but $\mathrm{E}(U_1 \mid X, Z) = 0$ and $\mathrm{E}(U_0 \mid X, Z) = 0$ in order to define the g_0 and g_1 functions.

[53] Heckman et al. (1997a, 1998c) extend the theory of matching to consider separable models with exclusion restrictions and discuss the efficiency gains from using such restrictions. Exclusion restrictions are natural in the context of panel data models where the variables in the outcome equation are measured in periods after the decision to participate in the program is made.

through V and that Q and Z are independent of U_0 and U_1. These assumptions imply the following:

$$\mathrm{E}(U_0 \mid Z, Q, D = 1) = \mathrm{E}(U_0 \mid V < H(Z)),$$

$$\mathrm{E}(U_0 \mid Z, Q, D = 0) = \mathrm{E}(U_0 \mid V \geq H(Z)),$$

$$\mathrm{E}(U_1 \mid Z, Q, D = 1) = \mathrm{E}(U_1 \mid V < H(Z)),$$

and

$$\mathrm{E}(U_1 \mid Z, Q, D = 0) = \mathrm{E}(U_1 \mid V \geq H(Z)),$$

We could just as well postulate this representation as the starting point for our analysis of the selection estimator. Both the bias, $B(Z)$ and the mean gain of the unobservables, $\mathrm{E}(U_1 - U_0 \mid Z, Q, D = 1)$, depend on Z only through the index $H(Z)$. When F_V is assumed to be strictly monotonic almost everywhere, we may write $H(Z) = F_V^{-1}(\Pr(D = 1 \mid Z))$ and the bias and mean gain terms depend on Z solely through the probability of participation P. The bias is now given by

$$B(P(Z)) = \mathrm{E}(U_0 \mid P(Z), D = 1) - \mathrm{E}(U_0 \mid P(Z), D = 0). \tag{7.14}$$

This is the "index sufficient" representation where $P(Z)$, or equivalently $H(Z)$, is the index.[54] An important question in the program evaluation literature is whether the selection bias can be characterized solely as a function of $P(Z)$ for different sets of Z, or if a more general conditioning set (Q,Z) is required to characterize this bias. In terms of the behavioral model of program participation and program outcomes presented in Section 6.2, the cost of participation, c, may play the role of V assuming that it is independent of other variables. Y_{0k} also could play that role provided that we condition on observed variables in forming the probability, and that the residual from this conditioning is independent of all the explanatory variables in the model.

Conventional econometric selection models (e.g., Amemiya, 1985) assume that the latent variables V, U_0, U_1 are symmetrically distributed around zero. The assumption of symmetry for U_0 and V implies that the bias $B(P(Z))$ is symmetric around $P(Z)$ equal to one-half. As shown by Fig. 8, in the normal selection model, if $P(Z)$ is symmetrically distributed around one-half, the average bias over symmetric intervals around that value is zero even though the pointwise bias is non-zero. If the values of $P(Z)$ for a sample of trainees were symmetrically distributed around one-half, the pointwise bias would be non-zero and the assumption justifying matching would not hold. Nonetheless, the selection bias would still average out to zero over any symmetric intervals of $P(Z)$ constructed around $P(Z) = 1/2$. Hence, the classical selection model justifies matching as a consistent estimator of (7.13) when it is defined over intervals of $P(Z)$ where the bias cancels out,

[54] See Heckman (1980) for the first derivation of this representation or Heckman and Robb (1985a, 1986a). Multiple decision rules for admission into a program require a multiple index model (Heckman and Robb, 1985a).

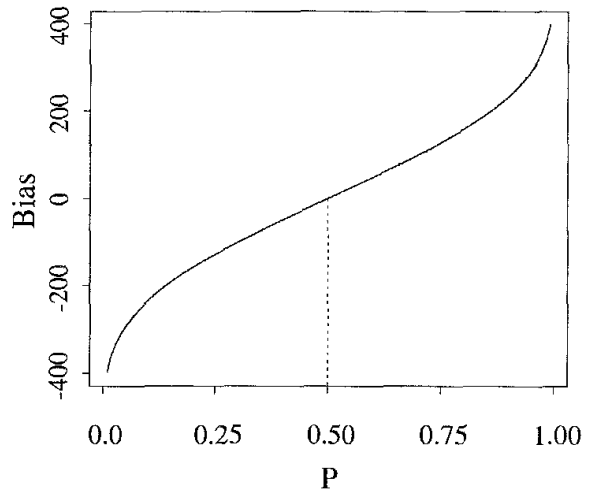

Fig. 8. Prototypical selection model, normal example: $B(P(X)) = \mathrm{E}(U_0 \mid P(X), D = 1) - \mathrm{E}(U_0 \mid P(X), D = 0)$. This is the index model where V and U_0 are assumed to be normal and $\sigma_V = 1$, $\sigma_{U_0} = 375$ and $\rho = \mathrm{cov}(U_0, V)/\sigma_{U_0} = 0.16$.

even though it would *not* justify matching for $\mathrm{E}(Y_1 - Y_0 \mid X, D = 1)$ defined pointwise for any points except those where the bias is zero.

To estimate the mean effect of treatment on the treated in the classic econometric selection model, we form the following regression based on Eq. (3.3):

$$\begin{aligned}\mathrm{E}(Y \mid Q, P(Z), D) &= \mathrm{E}(Y_1 D + Y_0(1 - D) \mid Q, P(Z), D)\\ &= g_0(Q) + D(g_1(Q) - g_0(Q)) + D(\mathrm{E}(U_1 \mid Q, P(Z), D = 1))\\ &\quad +(1 - D)\mathrm{E}(U_0 \mid Q, P(Z), D = 0). \qquad (7.15)\end{aligned}$$

The conditional means of the error terms $\mathrm{E}(U_1 \mid Q, P(Z), D = 1)$ and $\mathrm{E}(U_0 \mid Q, P(Z), D = 0)$ are called *control functions* (Heckman and Robb, 1985a, 1986a). Under the assumptions that U_0, U_1 are statistically independent of Q and Z, these functions may be written as

$$\mathrm{E}(U_0 \mid Q, P(Z), D = 0) = K_0(P(Z)),$$

$$\mathrm{E}(U_1 \mid Q, P(Z), D = 1) = K_1(P(Z)).$$

Specific distributional assumptions about (U_0, V) and (U_1, V) produce specific functional forms for K_0 and K_1. Heckman and MaCurdy (1986) present a catalogue of parametric models including the normal sample selection model of Heckman (1976, 1979).

Under these conditions, Eq. (7.15) is really just two sample selection bias equations applied to non-participants and participants respectively:

$$\mathrm{E}(Y_0 \mid Q, P(Z), D = 0) = g_0(Q) + K_0(P(Z)), \qquad (7.16a)$$

$$\mathrm{E}(Y_1 \mid Q, P(Z), D = 1) = g_1(Q) + K_1(P(Z)). \qquad (7.16b)$$

The most common form of the model writes $g_0(Q)=Q\beta_0$ and $g_1(Q)=Q\beta_1$, but this is not strictly required. We can use the $D = 1$ and $D = 0$ samples to recover the parameters of the model.

Assuming that there is at least one exclusion restriction (a variable in Z not in Q), and that $K_0(P(Z))$ and $K_1(P(Z))$ are not perfectly collinear with Q, we can identify $g_0(Q)$ and $g_1(Q)$ up to intercepts for any K_0 and K_1 functions. The intercepts are not determined. Any intercept in $g_0(Q)$ can be allocated to K_0 and vice versa; the same remark applies to the allocation of intercepts between $g_1(Q)$ and K_1. To identify the intercepts, it is necessary to have some Z values, say Z_0, such that $K_0(P(Z_0)) = 0$ and some Z values, say Z_1, such that $K_1(P(Z_1)) = 0$. Using such values, one can identify the unique intercepts for g_0 and g_1, respectively (Heckman, 1990).[55] Another way to determine the intercepts is to assume specific functional forms for K_0 and K_1 that exclude intercept terms as in the conventional normal selection bias model.

Many non-parametric and semiparametric selection bias strategies have been proposed that do not impose functional form assumptions on K_0 and K_1. All of these strategies require that we identify the intercepts on sets Z_0 and Z_1, respectively. See the comprehensive surveys by Heckman (1990), Powell (1994), and Honoré and Kyriazidou (1997). Andrews and Schafgans (1998) extend a method proposed in Heckman (1990) to identify the intercepts.

With g_0 and g_1 in hand, we can estimate

$$\mathrm{E}(Y_1 - Y_0 \mid Q) = g_1(Q) - g_0(Q).$$

To form $\mathrm{E}(Y_1 - Y_0 \mid Q, P(Z), D = 1)$ observe that from the preceding analysis we know $g_0(Q)$, $g_1(Q)$ and

$$\mathrm{E}(U_1 \mid Q, P(Z), D = 1) = K_1(P(Z)).$$

We do not directly estimate $\mathrm{E}(U_0 \mid Q, P(Z), D = 1)$. However, under our assumptions about the (mean) independence of U_0 and (Q,Z), we can write

$$0 = \mathrm{E}(U_0 \mid Q, P(Z), D = 1)P(Z) + \mathrm{E}(U_0 \mid Q, P(Z), D = 0)(1 - P(Z)).$$

Because we know both the second term in this expression and $P(Z)$, we can form

$$\mathrm{E}(U_0 \mid Q, P(Z), D = 1) = -K_0(P(Z))\frac{1 - P(Z)}{P(Z)}.$$

Thus we can construct[56]

$$\mathrm{E}(Y_1 - Y_0 \mid Q, P(Z), D = 1) = g_1(Q) - g_0(Q) + K_1(P(Z)) + K_0(P(Z))\frac{1 - P(Z)}{P(Z)}.$$

[55] This type of identification on limit sets is sometimes called "identification at infinity" because for some models the values of Z_0 and Z_1 that set K_0 and K_1 to zero are $\pm\infty$.

[56] Björklund and Moffitt (1987) construct $\mathrm{E}(Y_1 - Y_0 \mid X, D = 1)$ in exactly this way for a normal selection model.

To estimate $E(Y_1 - Y_0 \mid Q, D = 1)$ we simply integrate out (average out) $P(Z)$ against the density of $P(Z)$ conditional on $D = 1$ and Q, which can be estimated. Thus, by making separability, exclusion and intercept identification assumptions, we can identify the parameter of interest (see Heckman et al. (1998b) for details.)

The control function method parameterizes the bias function $B(P(Z))$ in terms of $K_1(P)$ and $K_0(P)$ and estimates these functions along with the other parameters of the model. The dependence induced between U_0 and D operating through the V is called "selection on unobservables." The dependence induced between U_0 and D operating through dependence between Z and U_0 is termed "selection on observables" (Heckman and Robb, 1985a, 1986a). In this context, the method of matching assumes selection on observables, because conditioning on Z controls the dependence between D and U_0, producing a counterpart to (7.6a) for the residuals: $E(U_0 \mid Z, D = 1) = E(U_0 \mid Z, D = 0)$. When selection is on unobservables, it is impossible to condition on Z and eliminate the selection bias. We next turn to the method of instrumental variables which, like matching, assumes that selection only occurs on the observables.

7.4.3. The method of instrumental variables

The method of instrumental variables (IV) applied to estimate $E(Y_1 - Y_0 \mid X, D = 1)$ is a variant of the method of matching. It augments the X variables in matching with instruments Z so that

$$E(U_1 - U_0 \mid X, Z, D = 1) = E(U_1 - U_0 \mid X, D = 1), \tag{7.17a}$$

$$E(U_0 \mid X, Z) = E(U_0 \mid X) \tag{7.17b}$$

and that

$$\Pr(D = 1 \mid X, Z) \tag{7.17c}$$

depends in a non-trivial way on both X and Z. In particular, there must be at least two values of Z, say Z' and Z'', such that for any X where we seek to identify the parameter of interest, $\Pr(D = 1 \mid X, Z') \neq \Pr(D = 1 \mid X, Z'')$. We assume that (X,Z) satisfies the non-causality condition (7.A.1) replacing X in that condition with (X,Z).

Condition (7.17a) rules out any dependence between $U_1 - U_0$ and Z given X and D. It is implied by the condition

$$\Pr(D = 1 \mid X, Z, U_1 - U_0) = \Pr(D = 1 \mid X, Z).$$

The second condition (7.17b) says that U_0 may depend on X but not on Z. This is not a standard IV condition but it is analogous to the balance of bias condition in matching. Applying these conditions, we use the law of iterated expectations to write

$$E(Y \mid X, Z') = g_0(X)$$

$$+[g_1(X) - g_0(X) + E(U_1 \mid X, D = 1) - E(U_0 \mid X, D = 1)]\Pr(D = 1 \mid X, Z') + E(U_0 \mid X).$$

We can express $E(Y \mid X, Z'')$ similarly for the same X, but a different $Z = Z''$. By subtracting the $E(Y \mid X, Z'')$ from $E(Y \mid X, Z')$, we can form the following expression:

$$\frac{E(Y \mid X, Z') - E(Y \mid X, Z'')}{\Pr(D = 1 \mid X, Z') - \Pr(D = 1 \mid X, Z'')} = g_1(X) - g_0(X) + E(U_1 - U_0 \mid X, D = 1)$$

$$= E(Y_1 - Y_0 \mid X, D = 1). \tag{7.18}$$

Condition (7.17a) ensures us that when we further condition on Z, it does not affect the conditioning of $U_1 - U_0$ on $D = 1$ and X. Condition (7.17c) assures us that the denominator of the expression is not zero.

Observe that if we assume that $E(U_0 \mid X) = 0$ and $E(U_1 \mid X) = 0$ (so $g_0(X) = \mu_0(X)$ and $g_1(x) = \mu_1(X)$),[57] and if we assume that

$$(U_0, U_1) \perp\!\!\!\perp D \mid X, Z', \tag{7.19}$$

then IV also identifies

$$E(Y_1 - Y_0 \mid X) = g_1(X) - g_0(X) = \mu_0(X) - \mu_1(X),$$

the effect of treatment on a randomly chosen person with characteristics X. Under these assumptions, matching and IV are now indistinguishable except that IV augments the original X variables by Z.[58]

If individuals select into the program on the basis of the gain in unobservables, $U_1 - U_0$, or on the basis of variables that are (stochastically) dependent on the gain in unobservables, the conditions required for IV estimators to consistently estimate $E(Y_1 - Y_0 \mid X, D = 1)$ are not satisfied (Heckman and Robb, 1985a, 1986a,b; Heckman, 1997) unless $U_1 = U_0$ or $U_1 - U_0$ is unknown or not acted on at the time program participation decisions are made. If the instrument Z is correlated with the gain in unobservables, and if individuals base their participation decisions at least in part on that gain, then the instrument is correlated with the error in the outcome equation. For the parameter of interest, treatment on the treated, failure of (7.17a) produces:

$$E(Y_1 - Y_0 \mid X, Z, D = 1) = (g_1(X) - g_0(X)) + E(U_1 - U_0 \mid X, Z, D = 1).$$

Because the instrument enters the second term on the right hand side, it is not a valid instrument. The outcome equation may be written as

$$Y = g_0(X) + DE(Y_1 - Y_0 \mid X, Z, D = 1)$$

$$+ \{U_0 + D[(U_1 - U_0) - E(U_1 - U_0 \mid X, Z, D = 1)]\}.$$

[57] If $E(U_0 \mid X) = 0$, then (7.17b) is the more familiar IV condition $E(U_0 \mid X, Z) = E(U_0 \mid X) = 0$.

[58] Observe that even if $E(U_0 \mid X) \neq 0$ and $E(U_1 \mid X) \neq 0$, under conditions (7.17a) to (7.17c), IV identifies $E(Y_1 - Y_0 \mid X) = g_1(X) - g_0(X) + E(U_1 - U_0 \mid X)$.

The term in braces is the unobservable when the parameter of interest is the impact of training on the trained. For Z to be a valid instrument, it must be mean independent of this error term. But if the gain in unobservables determines participation, then Z conditional on $D = 1$ is related to the gain and the expectation of the error term conditional on Z is certainly not equal to zero. The implication of this result is that when the response to training varies among individuals, and the parameter of interest is the impact of treatment on the treated, the method of instrumental variables requires a strong behavioral assumption about how persons make their decisions about program participation.

To make this point more concretely, consider an example in which program evaluators use the distance between a person's residence and the training center as an instrument. They assume that the distance to the training center affects outcomes only through the participation indicator in the earnings equation. The problem that arises in the heterogeneous response framework is that we would expect persons who live far away from the training center to participate in training only when their expected gain from training is relatively large – large enough to offset their higher cost of participation. By contrast, persons closer to the training center, who therefore face a lower cost of participation, will have smaller average expected gains from training. As a result, if an individual participates in training, their post-training earnings also depend on how far away they live from a training center. Therefore the instrument, distance, is correlated with the unobserved component of the gain from training for those who take training ($D = 1$) even if it is not for a random sample of persons in the population. Put differently, knowing how far trainees live from a training center tells us something about their expected earnings even conditional on their training status, which means that distance from the training center is not a valid instrument in this case.[59]

7.4.4. The instrumental variable estimator as a matching estimator

Heckman (1998c) shows how most evaluation estimators, including IV estimators, can be interpreted as matching estimators using the weighting framework of Eqs. (7.8) and (7.10). To see the basic idea, consider the simple random coefficient model

$$Y = \beta(X) + \alpha D + U.$$

We define β and α as functions of X where $\mathrm{E}(U \mid X, D) \neq 0$. Assume a valid instrument Z that satisfies conditions (7.17a)–(7.17c). Then

$$\mathrm{E}(Y \mid X, Z) = \beta(X) + \mathrm{E}(\alpha \mid X, D = 1)\mathrm{E}(D \mid X, Z) + \mathrm{E}(U \mid X, Z).$$

Now we can express the outcome equation as follows:

[59] Notice that this is an alternative interpretation that explains the "discount rate bias" recently discussed by Card (1995). Instrumenting by distance to a school or a training center may *raise* the estimated return to schooling or training if responses to schooling or training are heterogeneous and persons act on this heterogeneity in enrolling in schooling or training programs.

$$Y = \beta(X) + \mathrm{E}(\alpha \mid X, D = 1)\mathrm{E}(D \mid X, Z) + U$$

$$+[\alpha - \mathrm{E}(\alpha \mid X, D = 1)][\mathrm{E}(D \mid X, Z) + W] + \mathrm{E}(\alpha \mid X, D = 1)W,$$

where $D = \mathrm{E}(D \mid X, Z) + W$ and where, under our assumptions, the error terms have mean zero conditional on X and Z.[60] If we have a valid instrument, then $\mathrm{E}(U \mid X, Z) = \mathrm{E}(U \mid X)$ and $\mathrm{E}(\alpha \mid X, Z, D = 1) = \mathrm{E}(\alpha \mid X, D = 1)$. To identify $\mathrm{E}(\alpha \mid X, D = 1)$ we may form pairwise comparisons between person i and *anyone* else, provided that the matched partner for i, say i', has the same X but a different $Z = Z'$, where

$$\mathrm{E}(D \mid X, Z) \neq \mathrm{E}(D \mid X, Z').$$

If this condition is satisfied, we may match a suitable i' to form the pairwise estimate of the gains as follows:

$$\frac{Y_i - Y_{i'}}{\mathrm{E}(D_i \mid X, Z_i) - \mathrm{E}(D_{i'} \mid X, Z_{i'})}.$$

Therefore,

$$\mathrm{E}\left[\frac{Y_i - Y_{i'}}{\mathrm{E}(D_i \mid X, Z_i) - \mathrm{E}(D_{i'} \mid X, Z_{i'})}\right] = \mathrm{E}(\alpha \mid X, D = 1).$$

Accordingly, we can write our estimate of $\mathrm{E}(\alpha \mid X, D = 1)$ as a weighted average of contrasts:

$$\hat{\alpha} = \sum_{i,i'}\left[\frac{(Y_i - Y_{i'})}{\mathrm{E}(D_i \mid X, Z_i) - \mathrm{E}(D_{i'} \mid X, Z_{i'})}\right]W(i, i') \tag{7.20}$$

for i,i' such that $\mathrm{E}(D_i \mid X, Z_i) \neq \mathrm{E}(D_i' \mid X, Z_i')$, and where the weights are given by

$$W(i, i') = \frac{(\mathrm{E}(D_i \mid X, Z_i) - \mathrm{E}(D_{i'} \mid X, Z_{i'}))^2}{\sum_{i,i'}(\mathrm{E}(D_i \mid X, Z_i) - \mathrm{E}(D_i \mid X, Z_{i'}))^2}.$$

Formally, we set

$$\frac{Y_i - Y_{i'}}{\mathrm{E}(D_i \mid X, Z_i) - \mathrm{E}(D_{i'} \mid X, Z_{i'})} = 0$$

for i,i', where $\mathrm{E}(D_i|X, Z_i) = \mathrm{E}(D_{i'}|X, Z_{i'})$ and we get the same result summed over all i,i' since for these cases $W(i, i') = 0$.

Eq. (7.20) reveals that propensity score matching with Z as the propensity score estimates $\mathrm{E}(\alpha \mid X, D = 1)$ by taking a weighted average of all i,i' contrasts for values of (X,Z)

[60] As we have stressed, all we need is that the error terms depend only on X in order to recover $\mathrm{E}(\alpha(X) \mid X, D = 1)$.

with distinct probability values. Instrumental variable estimation is just a weighted average of contrasts of conditional means constructed in terms of propensity scores. Observe that this method only requires (7.17b) and not that $E(U \mid X, Z) = 0$. Thus, like matching and randomized trials, the IV method does not eliminate conventional econometric exogeneity bias – it just balances the bias.

7.4.5. IV estimators and the local average treatment effect

Imbens and Angrist (1994) reinterpret the output of IV Eq. (7.18) as the effect of treatment on those who change state in response to a change in Z. It is a discrete approximation to the marginal treatment effect (3.14) previously discussed in Section 3.4 and defined as the effect of a marginal change of a policy on those induced to change state as a consequence of the policy. Keeping the conditioning on X implicit, their parameter is $E(Y_1 - Y_0 \mid D(z) = 1, D(z') = 0)$ where $D(z)$ is the conditional random variable D given $Z = z$, and where z' is distinct from z, so $z \neq z'$. This conditions on people who switch from "0" to "1" as a consequence of the change in Z. This parameter is termed "LATE" for Local Average Treatment Effect.

The LATE parameter has several non-standard features. It is *defined* by variation in an instrumental variable that is external to the outcome equation. Unlike the instrumental variables discussed in the preceding section, in LATE, different instruments *define* different parameters. In the traditional IV literature, Z is used to identify the effect of X on outcomes. In LATE, variation in Z defines the parameter and no distinction between X and Z is made. When the instruments are indicator variables that denote different policy regimes, or when the instruments are different levels of intensity of a policy within a given regime (i.e., the level of φ in terms of the analysis of Section 3.4), LATE identifies the response to policy changes for those who change their program participation status in response to the policy change. When the instruments refer to personal or neighborhood characteristics used to predict an endogenous variable, say schooling in an earnings equation, LATE has a less clear cut interpretation and its relevance for policy analysis is questionable.

The measured variation in Z among people could be due to their choices of Z. If distance to the nearest school or training center is the instrument, LATE estimates the effect of variation in the distance to school on the earnings gain of persons who are induced to change their schooling or training status as a consequence of the different commuting costs they face. If a personal characteristic is used as an instrument, for example, family income, the parameter defines the marginal change in the outcome with respect to the variation in family income among those who would have changed their state in response to the sample variation in family income.

To define the LATE parameter more precisely, let $D(z)$ be the conditional random variable D given $Z = z$. (Recall that conditioning on X is kept implicit in this section). Since $D(z)$ is defined conditional on a particular realization of $Z = z$, it is independent of Z.[61] Imbens and Angrist (1994) assume that:

$(Y_0, Y_1, D(z))$ are independent Z and Pr(D=1|Z=z) is a non-trivial function of Z where these random variables are understood to be defined conditional on X. (7.IA.1)

As a consequence of this assumption, for a given person (with fixed Y_1,Y_0), and recalling that for $Z = z$, $Y = Y_0(1 - D(z)) + Y_1 D(z)$, we may write

$$\mathrm{E}(Y \mid Z = z) - \mathrm{E}(Y \mid Z = z') = \mathrm{E}[D(z)Y_1 + (1 - D(z))Y_0 \mid Z = z]$$

$$-\mathrm{E}[D(z')Y_1 + (1 - D(z'))Y_0 \mid Z = z']$$

$$= \mathrm{E}((D(z) - D(z'))(Y_1 - Y_0)). \tag{7.21}$$

The final step follows from assumption (7.IA.1) and depends crucially on the conditional independence of Y_1,Y_0 and $D(z)$ from Z.

In the Imbens–Angrist thought experiment, all of the random variables in the expression are defined for the same person. Thus for different values of $Z = z$, Y_1 and Y_0 do not change and $\{D(Z)\}$ for z in the support of Z is a collection of not necessarily independent random variables produced by changing Z and either not changing any other random variable or changing them only in the way specified in assumption (7.IA.2) below. In terms of the index model of discrete choice theory with index function $H(Z,V)$, which may be a net profit or net utility function, we have

$$D = 1(H(z, V) \geq 0) \tag{7.22}$$

and V is a random variable. In the Imbens and Angrist (1994) thought experiment, V stays fixed while z is varied.

From Eq. (7.21) it follows that

$$\mathrm{E}(Y \mid Z = z) - \mathrm{E}(Y \mid Z = z') = \mathrm{E}(Y_1 - Y_0 \mid D(z) - D(z') = 1)\Pr(D(z) - D(z') = 1)$$

$$+\mathrm{E}(Y_1 - Y_0 \mid D(z) - D(z') = -1)\Pr(D(z) - D(z') = -1). \tag{7.23}$$

This is the total effect on the outcome measure of a change in Z, including the effect on those induced to enter the program and the effect on those induced to leave the program. In terms of our discussion in Section 3.4, if Z is a policy variable, this produces the net effect of a change in Z on the aggregate measure of Y. This is one of the necessary ingredients for a cost benefit analysis of the effect of a marginal change in a policy variable on outcomes.

Imbens and Angrist (1994) break up the total effect into two terms: $\mathrm{E}(Y_1 - Y_0 \mid D(z) - D(z') = 1)$ and $\mathrm{E}(Y_1 - Y_0 \mid D(z) - D(z') = -1)$, defined for those induced into the program and induced out of it, respectively, and they present conditions that make it

[61] For two random variables (J,K) let f be the density (or frequency). Then $f(J,K) = f(J \mid K)f(K)$ so J given K is statistically independent of K although $f(J \mid K)$ may be functionally dependent on K.

possible to identify one of these. To identify the Imbens and Angrist (1994) "causal" parameters, a second assumption about the hypothetical random variables is required:

For all z, z' in the support of z, either $D(z) \geq D(z')$ for all persons or
$D(z) \leq D(z')$ for all persons. (7.IA.2)

Assuming that the denominator is not zero, this monotonicity assumption zeros out one of the two terms in (7.23). The assumption regarding the denominator is a technical condition. Even if the denominator is zero the program may have an effect on the aggregate through a shift in the composition of participants and non-participants. The variation across z and z' is made holding the error term constant. Condition (7.IA.2) makes either $\Pr(D(z) - D(z') = 1)$ or $\Pr(D(z) - D(z') = -1)$ zero for everyone. Thus, under their conditions the effect of a change in Z is to shift people into one sector or the other but not both. Suppose $D(z) \geq D(z')$, then $\Pr(D(z) - D(z') = -1) = 0$ and using (7.23) we obtain

$$\mathrm{E}(Y_1 - Y_0 \mid D(z) - D(z') = 1) = \frac{\mathrm{E}(Y \mid Z = z) - \mathrm{E}(Y \mid Z = z')}{\Pr(D = 1 \mid Z = z) - \Pr(D = 1 \mid Z = z')}. \tag{7.24}$$

If the monotonicity assumption is violated, IV estimates a weighted average of the LATE arising from people flowing from 0 to 1 and a reverse LATE arising from people flowing from 1 to 0, with the weights being

$$\frac{\Pr(D(z) - D(z') = 1)}{\Pr(D = 1 \mid Z = z) - \Pr(D = 1 \mid Z = z')} \quad \text{and} \quad \frac{\Pr(D(z) - D(z') = -1)}{\Pr(D = 1 \mid Z = z) - \Pr(D = 1 \mid Z = z')},$$

respectively. Because LATE is *defined* in terms of population moments, it can be consistently estimated by instrumental variables methods replacing population moments by sample moments.

Comparing (7.18) with (7.24) reveals that "LATE" looks like what the standard IV converges to except for one important difference: the LATE parameter is z dependent. Both the LATE and $\mathrm{E}(Y_1 - Y_0 \mid X, D = 1)$ are identified by taking the ratio of the change in the outcome induced by Z and dividing by the change in the probability of being in sector 1 induced by $Z = z$. The parameter $\mathrm{E}(Y_1 - Y_0 \mid X, D = 1)$ does not depend on Z while the LATE parameter does. Observe further that if conditions (7.17a) through (7.17c) are satisfied, the LATE estimator also identifies $\mathrm{E}(Y_1 - Y_0 \mid X, D = 1)$. Thus, in the case of a common coefficient model, or in the case where responses to training are heterogeneous, but not acted on by agents, LATE identifies $\mathrm{E}(Y_1 - Y_0 \mid X, D = 1) = \mathrm{E}(Y_1 - Y_0 \mid X)$.

Condition (7.IA.2) is satisfied if (7.22) characterizes choices. It is also satisfied by any index $IN = H(z, V_z)$ where

$$D(z) = 1(IN > 0 \mid Z = z)$$

characterizes participation in the program being evaluated, provided that H is increasing in

z, V_z is increasing in z and H is increasing in V_z. This would be satisfied in the case of a scalar z if

$$V_z = V_{z'} + \sigma(z),$$

for $z > z'$, where $\sigma(z)$ is a random variable with $\sigma(z) > 0$ when $z > z'$. If, however, $\sigma(z)$ is permitted to be both positive and negative, condition (7.IA.2) would not be satisfied.

The Roy model estimated by Heckman and Sedlacek (1985) has a decision rule of the form (7.22) or (6.5):

$$IN = Y_1 - Y_0 + k(z)$$

and

$$D = 1(IN > 0 \mid Z = z).$$

If $k(z)$ is monotonic in z, this decision rule produces a model consistent with (7.IA.2). To see this, assume that Y_1 and Y_0 are continuous random variables and that Z is independent of $(Y_1 - Y_0)$ so that the conditions of (7.IA.1) are satisfied. In the Imbens and Angrist (1994) thought experiment that defines their estimator, $Y_1 - Y_0 = V$ is fixed and different realizations of Z are considered. In this set up, the event $D(z) - D(z') = 1$ is described by the inequalities

$$Y_1 - Y_0 + k(z) > 0 \qquad \text{and} \qquad Y_1 - Y_0 + k(z') < 0$$

so that

$$-k(z') > Y_1 - Y_0 > -k(z)$$

and the condition $D(z) - D(z') = 1$ induces a partition of $Y_1 - Y_0$. Now the LATE "causal parameter" is

$$\mathrm{E}(Y_1 - Y_0 \mid D(z) - D(z') = 1) = \mathrm{E}(Y_1 - Y_0 \mid -k(z) < Y_1 - Y_0 < -k(z')),$$

which clearly depends on the choice of z and z'.[62] This example is a clear illustration of how under its assumptions LATE sidesteps the problem that Z is not a valid instrument for the treatment on the treated parameter in the Roy model. It estimates a different parameter that under its assumption approximates part of the marginal effect of a policy change derived in Section 3.4.

Consider once more our example of distance from the training center as an instrument for estimating the impact of training on earnings. If the LATE assumptions apply, but assumption (7.17a) does not, then LATE identifies the impact of commuting distance variation on training outcomes for those induced to participate by the change in the commuting distance. The LATE estimator with distance to the training center used as Z

[62] This example illustrates the point that statistical independence of two random variables does not imply their functional independence.

does not identify the impact of training for other samples or the LATE associated with different instruments.[63]

In general, the LATE parameter depends on the particular choice of the z and z' as well as X. Factors external to the outcome equation define the LATE parameter and a different parameter is produced for each choice of z and z'. If there are multiple instruments, there are multiple parameters. Additional instruments do not improve the efficiency with which a fixed parameter is estimated as they would in standard "policy invariant" structural models. Instead different instruments define different parameters. However, we have presented cases where the instruments are policy changes and LATE identifies a policy relevant parameter.

Heckman and Vytlacil (1999a,b) introduce a new parameter – the Local Instrumental Variable (LIV) parameter – which is a limit form of LATE when the instruments are continuous. A variety of evaluation parameters including LATE can be generated from it by using different weighting schemes. LIV is the fundamental building block of evaluation analysis. Heckman and Vytlacil (1999a) use LIV to bound parameters when treatment effects are not identified.

Imbens and Angrist (1994) claim that their identifying assumptions are much weaker than the more familiar identifying assumptions used in econometrics based on index models or latent variables crossing thresholds. In fact, their assumptions are *equivalent* to assuming a latent variable so there is no added generality in their approach (see Vytlacil, 1999).

7.4.6. Regression discontinuity estimators

Regression discontinuity estimators constitute a special case of "selection on observables." Originally introduced by Campbell and Stanley (1963), evaluations based on them have been presented by Goldberger (1972), Cain (1975), Barnow et al. (1980), Trochim (1984) and, more recently, by van der Klaauw (1997) and Hahn et al. (1998). In this model, treatment depends on some observed variable, S, according to a known, deterministic rule, such as $D = 1$ if $S < \bar{S}$, $D = 0$ otherwise. If Y_0 depends on S, and if $\alpha \neq 0$, then this assignment rule will induce a discontinuity in the relationship between $Y = Y_0 + D\alpha$ and S at the point $S = \bar{S}$.[64] Two features distinguish this case from the standard selection on observables case discussed in Section 7.4.1. First, there is no common support for participants and non-participants. For all values of S, $\Pr(D = 1 \mid S) \in \{0,1\}$. Thus, matching is impossible. Recall our example in Section 6.3 where the analyst knows (c, Y_{0k}). Conditioning on both of these variables violates the assumption (7.5). Thus the regression discontinuity estimator takes over when there is selection on observables but the overlapping support condition required for matching breaks down. Alternatively, the regression

[63] If the same parameter is estimated for all choices of commuting distances, then (7.17a) holds and the LATE estimator, which is formally equivalent to the IV estimator, recovers the impact of treatment on the treated. This is the basis for a test of whether the LATE is equivalent to the impact of treatment on the treated over the range of distance values for which it is estimated.

[64] We consider the assignment rule $S < \bar{S}$ for simplicity. The case with $S > \bar{S}$ is symmetric.

discontinuity design estimator is a limit form of matching at one point. Second, the selection rule is assumed to be deterministic and known.

Barnow et al. (1980), present a simple example of this estimator. They consider a hypothetical enrichment program for disadvantaged children based loosely on the US Head Start program. Children with family incomes below a cutoff level receive the program, all other children do not. The outcome variable of interest is the children's test scores. As shown in Fig. 9, the underlying relationship between test scores and family income is assumed to be linear. The line segment above the cutoff level reflects this relationship, which would continue (as shown by the broken line) to lower levels of family income in the absence of the program. The discontinuity in the regression line at the cutoff point represents the effect of the program, which is assumed to be a constant α. Under the assumptions of a common effect model and of linearity in the relationship between children's test scores and pre-program family income, α can be estimated without bias by OLS estimation of:

$$Y = \beta_0 + \alpha D + \beta_1 S + U. \tag{7.25}$$

Now consider the random coefficient case where α varies. We let α_i be the value of α for person i. The deterministic selection rule assumed in the regression discontinuity design precludes individuals choosing to participate in the program based on α_i. However, if α_i varies with S, then the mean impact of the treatment on the treated may differ from the mean impact of the treatment on a randomly selected person in the population. Due to the lack of a common support, no information about the impact of treatment on the untreated is available except at the point of discontinuity other than through extrapolation of the impact estimated for participants via functional form assumptions. Such an extrapolation

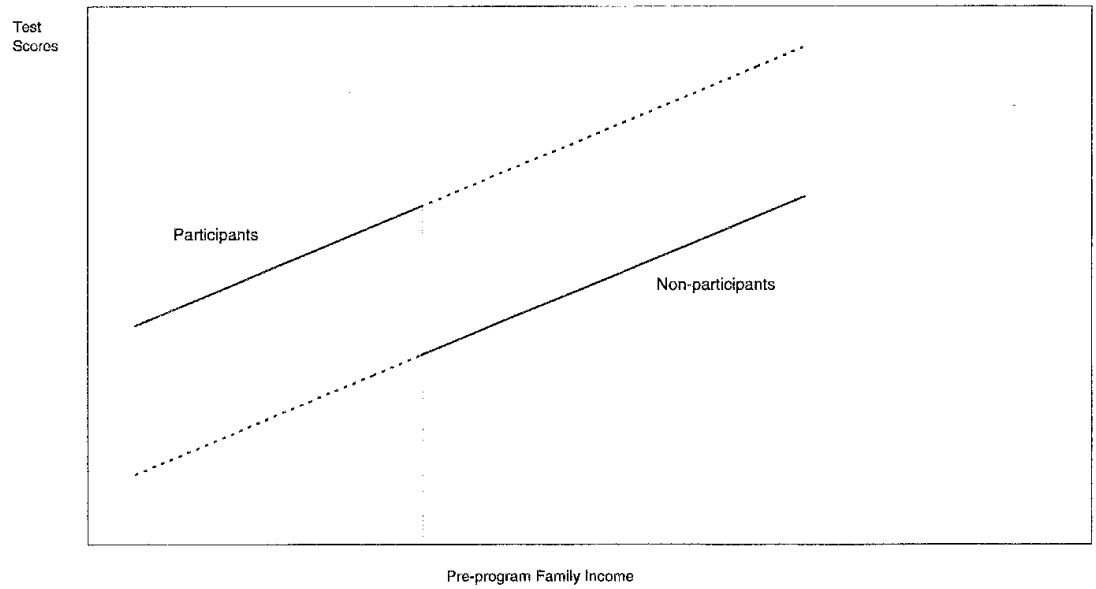

Fig. 9. Barnow et al. (1980): Head Start impact extrapolation example.

is illustrated in Fig. 9 in the upper broken line. This is a limitation of the estimator because one policy change of interest would be to increase the cutoff level to allow persons presently excluded at the margin to be included in the program. Some continuity has to be assumed to use the estimator in this situation. If it is assumed that the functional form of the relationship $Y_0(S)$ is known or can be determined using the available data, then we can estimate the impact of the treatment as a function of S, $\alpha(S)$, for persons in the program as the difference between the extrapolated $Y_0(S)$ and the observed outcomes of participants at each value of S. In the simplest case, if α_i is a linear function of S, then OLS estimation of

$$Y = \beta_0 + \alpha_0 D + \beta_1 S + \alpha_1 DS + U \tag{7.26}$$

will yield unbiased estimates of the linear relationship $\alpha_i = \alpha_0 + \alpha_1 S$ for participants. With knowledge of this relationship, we can readily determine the effects of a policy of cutting back the program by changing the cutoff point to exclude more people.

The most important issue in the application of the regression discontinuity design estimator is the extent to which the functional form of $Y_0(S)$ is known (possibly from samples not subject to treatment) or can be estimated. The older literature (e.g., Cook and Campbell, 1979) considers various methods such as selecting among polynomials in S through a combination of formal testing and visual inspection of the data. The more recent literature (e.g., Hahn et al., 1998) avoids this problem by estimating the impact of the treatment locally at the cutoff point using non-parametric methods. The former approach has the advantage of putting all of the data to use in identifying $Y_0(S)$; conditional on choosing the correct functional form, the parametric approach yields more precise estimates. The non-parametric approach has the advantage of avoiding extrapolation bias. In the random coefficient case, the non-parametric approach obtains only a local average treatment effect – the effect of treatment on participants with values of S close to $\bar{S}$. With multiple points of discontinuity, however, this problem becomes less severe. The parametric approach can still identify $\alpha(S)$ among participants as the difference between the observed relationship between Y and S conditional on $D = 1$ and the relationship $Y_0(S)$ estimated using the non-participants.

Another issue that arises in the analysis of regression discontinuity designs are so-called "fuzzy" assignment rules where assignment to the treatment is not a completely deterministic function of S. Except in the case of random variation in assignment conditional on S, fuzziness in the assignment rule changes this problem from one of selection on observables to one of selection on unobservables, conditional on S. The general methods for dealing with selection on unobservables discussed in this chapter can be applied in this situation, but much of the simplicity of the regression discontinuity design is lost (see the discussion in van der Klaauw, 1997). Still, the discontinuity assumed for S can aid in identifying parameters. For a random effect model, and under local monotonicity, Hahn et al. (1998) identify a LATE effect.

A final issue that arises in regression discontinuity designs concerns non-participation in the program by persons whose values of S make them eligible for it. Unless the common effect model is assumed, or the random effect model is assumed but participation does not

depend on the person-specific component of the impact, the simple estimation schemes described above no longer identify the mean impact of the treatment on persons satisfying the cutoff condition, even if the functional form of $Y_0(S)$ is known. They do, however, provide unbiased estimates of the impact of the treatment on the treated. If we seek to identify the impact of treatment on all persons below the cutoff point, $S < \bar{S}$, which would be of interest in regard to proposals to increase participation among persons already eligible (i.e., raising the "take-up rate"), we must apply modified versions of the non-experimental methods discussed in this section.

7.5. Using aggregate time series data on cohorts of participants to evaluate programs

For the model of Section 6.3 with one possible program enrollment period over the life-cycle, (e.g., schooling, or army service), and for many other models, it is sometimes possible to identify the effect of treatment on the treated using only data on cohort means, without knowing the treatment status of any individual in the cohort. As noted in Section 3.4, in principle one can evaluate a program using aggregate time series data and thereby avoid the selection problem. Initially, assume a time homogeneous environment. Estimates of the aggregate cohort mean outcomes formed in two or more cross-sections of unrelated persons measured before and after the age where participation in the program is possible can be used to obtain estimates of the effect of treatment on the treated free of selection bias even if the training status of each person is unknown so long as the cohort proportion of trainees is known or can be consistently estimated. With more data, the time homogeneity assumption can be partially relaxed, as we will demonstrate.

Assuming a time homogeneous environment and access to repeated cross-section data governed by random sampling, it is possible to identify $\alpha = \mathrm{E}(Y_1 - Y_0 \mid D = 1)$ (a) without any instrumental variables, (b) without need to specify the joint distribution of U_0, U_1 and V, and (c) without any need to know which individuals in the sample enrolled in training. However, the proportion of training must be known or consistently estimable (Heckman and Robb, 1985a, 1986b.)

To show how this is possible, suppose that no regressors appear in the earnings function.[65] Assuming that random sampling generates the data, the expectation of the cohort means (denoted by "-") is

$$\mathrm{E}(\bar{Y}_t) = \mathrm{E}(\beta_t + \alpha D + U_t) = \beta_t + \mathrm{E}(\alpha \mid D = 1)P, \qquad \text{for } t > k,$$

and

$$\mathrm{E}(\bar{Y}_{t'}) = \mathrm{E}(\beta_{t'} + U_{t'}) = \beta_{t'}, \qquad \text{for } t' < k,$$

[65] If regressors appear in the earnings functions, condition on X. See Heckman and Robb (1985a, 1986a) for the general case.

where $P = \Pr(D = 1)$. In a time homogeneous environment, $\beta_t = \beta_{t'}$ and

$$\frac{\mathrm{E}(\bar{Y}_t) - \mathrm{E}(\bar{Y}_{t'})}{P} = \mathrm{E}(\alpha \mid D = 1).$$

Replacing sample means with population means defines the estimator. The estimator can be formed within X strata. This is a grouping estimator that averages out the error term. Nowhere does it exploit any covariance term to identify the parameter. Hence, it is possible to identify the parameter when U is correlated with D and there is no conventional instrumental variable.

With more than two years of repeated cross-section data, one can apply the same principles to identify $\mathrm{E}(\alpha \mid D = 1)$ while relaxing the time homogeneity assumption. For instance, suppose that the time trend for cohort mean earnings lies on a polynomial of order $L - 2$:

$$\beta_t = \pi_0 + \pi_1 t + \cdots + \pi_{L-2} t^{L-2}.$$

From L temporally distinct cross-sections, it is possible to consistently estimate the $L - 1$ π-parameters and $\mathrm{E}(\alpha \mid D = 1)$ provided that the number of observations in each cross-section becomes large and there is at least one pre-program and one post-program cross-section.

If the effect of training differs across periods, it is still possible to identify $\mathrm{E}(\alpha_t \mid D = 1)$, provided that the environment changes in a "sufficiently regular" way. For example, suppose

$$\beta_t = \pi_0 + \pi_1 t,$$

$$\mathrm{E}(\alpha_t \mid D = 1) = \phi_0(\phi_1)^{t-k}, \qquad t > k.$$

In this case, π_0, π_1, ϕ_0, and ϕ_1 are identified from the means of four cross-sections, as long as at least two of these means come from a pre-program period and two come from successive post-program periods.[66]

Heckman and Robb (1985a, 1986b) state the conditions required to consistently estimate $\mathrm{E}(\alpha_t \mid D = 1)$ using repeated cross-section data on cohort aggregates which do not record the training identity of individuals under general conditions about cohort and time effects. Section 7.7 studies the sensitivity of this class of estimators to violations of the random sampling assumption.

7.6. Panel data estimators

Access to repeated observations on the same persons followed over time enables analysts to exploit the time series properties of the outcome equations and their relationship with

[66] Heckman and Robb (1985a) show how to solve the four equations for means in terms of the four unknown parameters.

program participation equations. Like the classical econometric selection bias estimators, panel data estimators exploit additive separability between model and error.

This subsection consists of four parts. In the first part, we consider panel data estimators for the common coefficient model (3.10). We allow α to depend on X but we assume only one error term $U_{1t} = U_{0t} = U_t$. A model with two errors in (3.10), $U_t = DU_{1t} + (1 - D)U_{0t}$, complicates the analysis and alters the conclusions reached for the simpler case of a single error term. This case requires a separate analysis because many longitudinal estimators are not robust to the introduction of two regime-specific error terms into the model.

In the second part, we extend the panel data models to apply to repeated cross-section data. We demonstrate how many conventional panel data evaluation estimators can be applied to repeated cross-sections of the same populations sampled over time. This is fortunate since repeated cross-section data are much more commonly available around the world than are panel data. In the third part, we extend these results to allow for a two component model, so that there is heterogeneity in responses to program participation on unmeasured outcomes ($U_{0t} \neq U_{1t}$). Finally, in the fourth part we show how the panel data estimators can be placed within the matching framework of Section 7.4.1.

7.6.1. Analysis of the common coefficient model

We start the analysis by assuming model (3.10) with $U_{1t} = U_{0t} = U_t$ so that $Y_{1t} - Y_{0t} = \alpha$ but α may depend on X, $\alpha(X)$.We consider more general cases below. The cases considered in this section are the familiar models used in conventional panel data analysis.

7.6.2. The fixed effects method

We begin our analysis with the conventional fixed effect model. Eq. (6.7) presents the key identifying assumption of the fixed effect method. If we allow Eq. (3.10) to include observed characteristics, the identifying assumption is:

$$\mathrm{E}(U_{0t} \mid X, D = 1) = \mathrm{E}(U_{0t'} \mid X, D = 1), \qquad \text{for some } t > k > t'. \tag{6.8'}$$

Recall that k is the period of program participation. Suppose that this condition holds and the analyst has access to 1 year of pre-program and 1 year of post-program outcome data. Regressing the difference between the outcomes in the years t and t' on a dummy variable for training status produces a consistent estimator of α. (This method is well exposited in Hsiao, 1986.) A variety of efficient estimators have been developed that exploit the multiplicity of contrasts that are sometimes available.

Some program participation rules and error processes for earnings justify condition (6.8'). For example, consider a certainty environment in which the earnings residual has a permanent-transitory structure:

$$U_t = \phi + \varepsilon_t,$$

where ε_t is a mean zero random variable independent of all other values of $\varepsilon_{t'}$ for $t \neq t'$, and is distributed independently of ϕ, a mean zero person-specific time-invariant random

variable. Assuming that the V ($=c + Y_{0k}$) in participation rule (6.4) are distributed independently of all ε_t, except possibly for ε_k, condition (6.8′) will be satisfied provided that decision rule (6.3) generates participation. However, this condition is violated if there are imperfect credit markets as in Section 6.3.4. With two periods of data (in t and t', $t > k > t'$), α is identified. With more periods of panel data, the model is overidentified and hence we can test condition (6.8′). See the discussion in Hsiao (1986).

The permanent-transitory error structure is very special. As already discussed in Section 6, much evidence speaks against this error specification as a description of earnings residuals. (See also the discussion of the evidence in Section 8.4.) This method is crucially dependent on additivity of the errors, strong assumptions about program participation rules and special assumptions about the time series properties of the errors. Thus it is not surprising that LaLonde (1986) finds the method to be one of the least reliable non-experimental estimators for evaluating training programs.

7.6.3. U_t follows a first-order autoregressive process

We consider a more general model and assume that U_t follows the first-order autoregression given by Eqs. (7.1a)–(7.1c). Substitution into (3.10) yields

$$Y_t = [X_t - (X_{t'}\rho^{t-t'})]\beta + (1 - \rho^{t-t'})D\alpha + \rho^{t-t'}Y_{t'} + \left\{\sum_{j=0}^{t-(t'+1)} \rho^j \varepsilon_j\right\}, \qquad \text{for } t > t' > k. \tag{7.27}$$

This expression is an alternative form of (7.2) that includes regressors. Assume further that either (i) the perfect foresight rule of Eq. (6.3) determines enrollment and the ε_j are distributed independently of X or (ii) that the post-k ε_t are not known at k, and are forecast to have zero means. (Heckman and Wolpin (1976) invoke similar assumptions in their analysis of affirmative action programs.) If the X are independent of $\varepsilon_{j'}$ for all j,j',[67] then least squares applied to Eq. (7.27) consistently estimates α in large samples.[68] Unlike the fixed effects model, the autoregressive model does not require preprogram earnings and hence can be used to evaluate schooling or training programs for youth. As is the case with the fixed effect estimator, the model becomes overidentified (and hence testable) for panels with more than two time periods. If we assume imperfect credit markets of the form presented in Section 6.3.4, the estimator is inconsistent because participation depends on all lagged and future ε_t and D is correlated with the error in (7.27).

7.6.4. U_t is covariance stationary

The next procedure invokes an assumption about the time series properties of the error that is implicitly used in many papers on training (Ashenfelter, 1978; Bassi, 1983), and exploits

[67] This condition can be weakened to mean independence: $E(\varepsilon_j \mid X_1, \ldots, X_T) = 0$ for all j.

[68] A non-linear regression that imposes restrictions across coefficients increases efficiency.

the assumption in a novel way (Heckman and Robb 1985a, 1986a). We assume the following:

(a) U_t is covariance stationary so $E(U_t U_{t-j}) = \sigma_j$ for $j \geq 0$;

(b) there is access to at least two observations on pre-program earnings in t' and $t' - j$ as well as one observation on post-program earnings in t where $t - t' = j$; and

(c) $E(U_{t'} \mid D = 1)P \neq 0$, where $P = \Pr(D = 1)$.

Unlike the two previous models, here we make no assumptions about the appropriate participation rule or about the stochastic relationship between U_t and the cost of enrollment in (3.10) or (6.3). We can define this model conditional on X values.

We write the model as

$$Y_t = \beta_t + D\alpha + U_t, \qquad \text{for } t > k,$$

$$Y_{t'} = \beta_{t'} + U_{t'}, \qquad \text{for } t' < k,$$

where β_t and $\beta_{t'}$ are period-specific shifters and the conditioning on X is kept implicit.

Using a random sample of pre-program earnings from periods t' and $t' - j$, we can consistently estimate $\sigma_j = \mathrm{Cov}(Y_{t'}, Y_{t'-j})$ using the least squares residuals. If $t > k$ and $t - t' = j$, so that the post-program earnings data are as far removed in time from t' as t' is removed from $t' - j$, the covariance $\mathrm{Cov}(Y_t, Y_{t'})$ satisfies

$$\mathrm{Cov}(Y_t, Y_{t'}) = \sigma_j + \alpha P E(U_{t'} \mid D = 1), \qquad \text{for } t > k > t'.$$

The covariance between D and $Y_{t'}$ is

$$\mathrm{Cov}(Y_{t'}, D) = PE(U_{t'} \mid D = 1), \qquad \text{for } ' < k.$$

Assuming $E(U_{t'} \mid D = 1)P \neq 0$ for $t' < k$, we obtain

$$\alpha = \frac{\mathrm{Cov}(Y_t, Y_{t'}) - \mathrm{Cov}(Y_{t'}, Y_{t'-j})}{\mathrm{Cov}(Y_{t'}, D)}.$$

Using sample moments in place of population moments defines the estimator. For panels of sufficient length (e.g., more than two pre-program observations or more than two post-program observations), the stationarity assumption can be tested. Increasing the length of the panel converts a just-identified model to an overidentified one. Heckman and Robb (1985a) consider a variety of other assumptions that exploit the time series properties of the panel data including factor structure models for error processes.

7.6.5. Repeated cross-section analogs of longitudinal procedures

We can apply most longitudinal procedures to repeated cross-section data. Such data are cheaper to collect, and they do not suffer from the problems of non-random attrition which often plagues panel data.[69] The previous section presented longitudinal estimators of α

[69] These points were first made in Heckman and Robb (1985a, 1986a).

that are based on identifying moment conditions. In all cases but one, however, we can identify α with repeated cross-section data. Heckman and Robb (1985a, 1986b) give many additional examples of longitudinal estimators which can be implemented on repeated cross-section data.

7.6.6. The fixed effect model

As in Section (7.6.2), assume that condition (6.8′) holds so that

$$\mathrm{E}(U_t \mid D = 1) = \mathrm{E}(U_{t'} \mid D = 1),$$

$$\mathrm{E}(U_t \mid D = 0) = \mathrm{E}(U_{t'} \mid D = 0),$$

for all t,t' such that $t > k > t'$. As before, we can condition on X. $\mathrm{E}(Y_t \mid D = 1)$ is the mean outcome of participants in year t and $\mathrm{E}(Y_t \mid D = 0)$ is the mean outcome of non-participants in year t, with sample counterparts $\bar{Y}_t^1$ and $\bar{Y}_t^0$ respectively. The parameter can be written in terms of population moments as

$$\alpha = [\mathrm{E}(Y_t \mid D = 1) - \mathrm{E}(Y_t \mid D = 0)] - [\mathrm{E}(Y_{t'} \mid D = 1) - \mathrm{E}(Y_{t'} \mid D = 0)]$$

with sample counterpart

$$\hat{\alpha} = (\bar{Y}_t^{(1)} - \bar{Y}_t^{(0)}) - (\bar{Y}_{t'}^{(1)} - \bar{Y}_{t'}^{(0)}).$$

Assuming random sampling, consistency of $\hat{\alpha}$ follows immediately. As in the case of the longitudinal version of this estimator, with more than two cross-sections, condition (6.8′) can be tested.

In one respect this example is contrived. It assumes that in the pre-program cross-sections we know the identity of future trainees. Such data might exist (e.g., individual person records can be matched to subsequent training records). One advantage of longitudinal data for identifying and estimating α is that we know the training status of all persons without resort to further sampling or matching of records across different data sources.

7.6.7. The error process follows a first-order autoregression

Suppose, instead, that U_t follows a first-order autoregressive process given by Eq. (7.1c) and that

$$\mathrm{E}(\varepsilon_t \mid D) = 0, \qquad \text{for } t > k.$$

It is possible to identify α with three successive post-program cross-sections in which the identity of trainees is known.

To establish this result, let the three post-program periods be t, $t + 1$ and $t + 2$. Assuming, as before, that no regressor appears in Eq. (7.2), or, alternatively, conditioning on X, we obtain:

$$\mathrm{E}(Y_j \mid D = 1) = \beta_j + \alpha + \mathrm{E}(U_j \mid D = 1),$$

$$\mathrm{E}(Y_j \mid D = 0) = \beta_j + \mathrm{E}(U_j \mid D = 0),$$

From condition (7.1c),

$$\mathrm{E}(U_{t+1} \mid D = 1) = \rho \mathrm{E}(U_t \mid D = 1),$$

$$\mathrm{E}(U_{t+1} \mid D = 0) = \rho \mathrm{E}(U_t \mid D = 0),$$

$$\mathrm{E}(U_{t+2} \mid D = 1) = \rho^2 \mathrm{E}(U_t \mid D = 1),$$

$$\mathrm{E}(U_{t+2} \mid D = 0) = \rho^2 \mathrm{E}(U_t \mid D = 0),$$

Then

$$\rho = \frac{[\mathrm{E}(Y_{t+2} \mid D = 1) - \mathrm{E}(Y_{t+2} \mid D = 0)] - [\mathrm{E}(Y_{t+1} \mid D = 1) - \mathrm{E}(Y_{t+1} \mid D = 0)]}{[\mathrm{E}(Y_{t+1} \mid D = 1) - \mathrm{E}(Y_t \mid D = 0)] - [\mathrm{E}(Y_t \mid D = 1) - \mathrm{E}(Y_t \mid D = 0)]} \tag{7.28a}$$

and

$$\alpha = \frac{[\mathrm{E}(Y_{t+2} \mid D = 1) - \mathrm{E}(Y_{t+2} \mid D = 0)] - \rho[\mathrm{E}(Y_{t+1} \mid D = 1) - \mathrm{E}(Y_{t+1} \mid D = 0)]}{1 - \rho}. \tag{7.28b}$$

Replacing population moments by sample moments defines the estimator.[70]

For this model, the advantage of longitudinal data is clear. Only two time periods of longitudinal data are required to identify α, but three periods of repeated cross-section data are required to estimate the same parameter. However, if Y_t is subject to measurement error, the apparent advantages of longitudinal data become less clear. Repeated cross-section estimators are robust to mean zero measurement error in the variables. The longitudinal regression estimator discussed in the preceding section does not identify α unless the analyst observes earnings without error or has access to instruments to adjust for the measurement error. Given 3 years of longitudinal data and assuming that measurement error is serially uncorrelated, one could instrument Y_t, in Eq. (7.2), using earnings in the earliest year as an instrument. This requires one more year of data. Thus one advantage of the longitudinal estimator disappears in the presence of measurement error. With four or more repeated cross-sections, the model is obviously overidentified and hence subject to test.

7.6.8. Covariance stationary errors

For simplicity, we implicitly condition on X (see Heckman and Robb, 1985a, 1986a, for the case in which regressors are present.) For any model with stationary errors

[70] Notice that a test that the numerator is zero is a test that $\rho = 1$. Thus one can test the identifying condition that $\rho \neq 1$.

$$\text{Var}(Y_t) = \alpha^2(1 - P)P + 2\alpha\text{E}(U_t \mid D = 1)P + \sigma_U^2, \qquad \text{for } t > k,$$

$$\text{Var}(Y_{t'}) = \sigma_U^2, \qquad \text{for } t' < k,$$

$$\text{Cov}(Y_t, D) = \alpha\, P(1 - P) + \text{E}(U_t \mid D = 1)P.$$

Note that $\text{E}(U_t^2) = \text{E}(U_{t'}^2)$ by virtue of our assumption of stationarity. Then[71]

$$\alpha = [P(1 - P)]^{-1}[\text{Cov}(Y_t, D) - \{[\text{Cov}(Y_t, D)]^2 - P(1 - P)[\text{Var}(Y_t) - \text{Var}(Y_{t'})]\}^{1/2}].$$

Replacing sample moments with population moments defines the estimator. Different features of the covariance stationarity assumptions are being exploited. The longitudinal procedure only requires that $\text{E}(U_tU_{t-j})=\text{E}(U_{t'}U_{t'-j})$ for $j > 0$; variances need not be equal across periods. The repeated cross-section analog above requires only that variances be stationary; covariances could differ among equispaced pairs of the U_t. With more than two cross-sections, the covariance stationarity assumption is overidentifying and hence subject to test.

7.6.9. The anomalous properties of first difference or fixed effect models

Almost all of the estimators considered in this chapter require a comparison group (i.e., a sample of non-trainees). The only exception is the fixed effect estimator in a time homogeneous environment where $\beta_t = \beta_{t'}$. In this case, if condition (6.8′) holds, and if we let $X_{it}\beta = \beta_t$ to simplify the exposition, then

$$\alpha = \text{E}(Y_t \mid D = 1) - \text{E}(Y_{t'} \mid D = 1).$$

No information on non-participants is needed, although information on participation or non-participation by the same persons is required.[72] This is not a general feature of the other estimators that we have considered. Even in stationary environments, other estimators require both participants and non-participants. Even the fixed effect estimator requires information on non-participants in a non-stationary environment.

Many of the procedures considered here can be implemented using only post-program data. It is not necessary to have pre-program background data. The covariance stationarity estimators of Section 7.6.4, certain repeated cross-section estimators, and first difference methods constitute exceptions to this rule. In this sense, those estimators are anomalous.

Fixed effect estimators also are robust to departures from the random sampling assumption. For instance, suppose condition (6.8′) is satisfied, but that the available data oversample or undersample trainees (i.e., the proportion of trainees in the sample does not

[71] The negative root of the quadratic equation for α derived from the three moments presented in the text does not identify the parameter. For details, see Heckman and Robb (1985a).

[72] Strictly speaking, we can implement the estimator by sampling participants in the same cohorts without sampling the same persons in t and t'. Recall our discussion of the repeated cross-section estimators.

converge to $P = \mathrm{E}(D)$). Suppose further that the analyst does not know the true value of P. Nevertheless, a first difference regression continues to identify α. As noted in Section 7.7, many other procedures do not share this property.

7.6.10. Robustness of panel data methods in the presence of heterogeneous responses to treatment

It is not surprising that estimators that exploit properties of covariances and variances of model residuals are affected by changes in the properties of the residuals. We have already noted in Section 3 that when responses to treatment are heterogeneous, we acquire a non-standard error term (see (3.7) and (3.9)). As we move from the common coefficient case to the heterogeneous-response case, we encounter new phenomena. Some of the estimators we have considered are robust to the introduction of heterogeneity. Others are not.

In this chapter, we focus on estimating the impact of treatment on the treated so Eq. (3.9) and its error term are the appropriate objects of attention. The induced heteroscedasticity clearly makes the repeated cross-section estimator based on stationarity invalid whether or not $U_{1t} - U_{0t}$ is anticipated in making program participation decisions. Without modification the longitudinal estimator based on covariance stationarity also is invalid.

In contrast, the fixed effect estimator (applied to panels or repeated cross-sections) is robust to heterogeneity in responses provided that the object of an evaluation is to identify $\mathrm{E}(Y_{1t} - Y_{0t} \mid X, D = 1)$. To see this point notice that the fixed effect estimators (for panels or repeated cross-sections) only use conditional mean properties of the errors. From the definition of the parameter (3.8) and Eq. (3.9), the error component induced by heterogeneous responses has mean zero $(\mathrm{E}[(U_1 - U_0) - \mathrm{E}(U_1 - U_0 \mid X, D = 1) \mid X, D = 1] = 0)$. Thus the properties of the estimator are not affected by heterogeneity in response when treatment on the treated is the parameter of interest. The selection problem arises solely from dependence between U_0 and D.

The autoregressive estimators provide an interesting example where the introduction of response heterogeneity affects the panel data version the estimator for the effect of treatment on the treated but not the repeated cross-section version. We develop our analysis of the autoregressive estimator in this context in two stages. First assume that the difference in the outcomes in any two periods is time invariant:

$$Y_{1t} - Y_{0t} = \alpha \qquad t > k.$$

Then letting

$$U_{0t} = \rho U_{0,t-1} + \varepsilon_t,$$

where ε_t is independent and identically distributed (i.i.d), we may now write the outcome equation as

$$Y_t = [X_t - X_{t'}\rho^{t-t'}]\beta + (1 - \rho^{t-t'})D\mathrm{E}(\alpha \mid X, D = 1)$$

$$+\rho^{t-t'}Y_{t'} + \sum_{j=0}^{t-(t'+1)} \rho^j \varepsilon_{t-j} + (1-\rho^{t-t'})D[\alpha - \mathrm{E}(\alpha \mid X, D=1)]. \tag{7.29}$$

Observe that even if the ε_t do not determine program participation for periods $t > k$,[73]

$\mathrm{Cov}(Y_{t'}, D(\alpha - \mathrm{E}(\alpha \mid X, D=1))) \neq 0.$

Consequently $Y_{t'}$ is correlated with the error term in the model and additional identifying information is required. But recall that the repeated cross-section estimator only uses group means. Just as in the case of the fixed effect estimator, the final error component of Eq. (7.29) averages out to zero when means are constructed. Thus the repeated cross-section estimator is robust to the introduction of response heterogeneity in the model, while the panel data version of the estimator is not.

This point is more general. All repeated cross-section estimators based on means that identify the parameter in the case of a common effect are consistent for $\mathrm{E}(Y_{1t} - Y_{0t} \mid X, D=1)$ in the random coefficient case. The new error component introduced when responses to treatment are heterogeneous averages out over persons. This is a property of the additive separability that underlies the entire class of estimators examined in this section of the chapter and clearly demonstrates the dependence of the properties of these estimators on functional form assumptions.

For the more general autoregressive processes given by

$$U_{1t} = \rho_1 U_{1,t-1} + \varepsilon_{1t},$$

$$U_{0t} = \rho_0 U_{0,t-1} + \varepsilon_{0t},$$

where $\mathrm{E}(\varepsilon_{1t}) = \mathrm{E}(\varepsilon_{0t}) = 0$, and $(\varepsilon_{1t}, \varepsilon_{0t})$ is i.i.d across persons but $\mathrm{E}(\varepsilon_{1t}\varepsilon_{0t}) \neq 0$, the autoregressive estimator is no longer clearly defined. The parameter "treatment on the treated" is now, in general, period dependent, even if $\beta_{1t} = \beta_{0t} = \beta$, because $\mathrm{E}(U_{1t} - U_{0t} \mid X, D=1)$ depends on period t. In addition, unless $\rho_1 = \rho_0$, it is no longer true that we can exploit the trick used to obtain Y_t in terms of lagged $Y_{t'}$ and eliminate $U_{t'}$ in (7.1a) to (7.1c).

When $\rho_1 = \rho_0 = \rho$, U_t still can be written in autoregressive form:

$$U_t = DU_{1t} + (1-D)U_{0t} = \rho^{t-t'}[DU_{1t'} + (1-D)U_{0t'}] + \sum_{j=0}^{t-(t'+1)} \rho^j [D\varepsilon_{1,t-j} + (1-D)\varepsilon_{0,t-j}].$$

Assuming a common β in both regimes except for a time-invariant difference in intercepts

[73] Observe that the error term for $Y_{t'}$ includes $D(\alpha - \mathrm{E}(\alpha \mid X, D=1))$. Then for the variable coefficient model the final term in (7.29) is correlated with this component of $Y_{t'}$. The covariance is $(1-\rho^{t-t'})\mathrm{E}(D(\alpha - \mathrm{E}(\alpha \mid X, D=1))^2) = (1-\rho^{t-t'})\mathrm{Var}(\alpha \mid X, D=1)\mathrm{P} \neq 0$ where $P = \Pr(D=1)$. The phenomenon here is similar to the fixed effect bias analyzed by Balestra and Nerlove (1966) except that the fixed effect in our model is state contingent: D times the fixed effect $\alpha - \mathrm{E}(\alpha \mid X, D=1)$.

α, the parameter treatment on the treated for period t is

$$[\alpha + \mathrm{E}(U_{1t} - U_{0t} \mid X, D = 1)] = \alpha + \rho^{t-t'}\mathrm{E}(U_{1t'} - U_{0t'} \mid X, D = 1)$$

$$+ \sum_{j=0}^{t-t'} \rho^j \mathrm{E}(\varepsilon_{1,t-j} - \varepsilon_{0,t-j} \mid X, D = 1). \tag{7.30}$$

Applying the autoregressive transform, we obtain

$$Y_t = (X_t - X_{t'}\rho^{t-t'})\beta + [\alpha(1 - \rho^{t-t'}) + \sum_{j=0}^{t-(t'+1)} \rho^j \mathrm{E}(\varepsilon_{1,t-j} - \varepsilon_{0,t-j} \mid X, D = 1)D + \rho^{t-t'}Y_{t'}$$

$$+ \sum_{j=0}^{t-t'} \varepsilon_{0,t-j} + D\left[\sum_{j=0}^{t-(t'+1)} \rho^j [\varepsilon_{1,t-j} - \varepsilon_{0,t-j} - \mathrm{E}(\varepsilon_{1,t-j} - \varepsilon_{0,t-j} \mid X, D = 1)]\right].$$

If the $\varepsilon_{l,t}$, $t > k$, $l = 0$ or 1, are not forecastable at $t = k$, then the parameter treatment on the treated in period $t > k$ is

$$\alpha + \mathrm{E}(U_{1t} - U_{0t} \mid X, D = 1) = \alpha + \rho^{t-k}\mathrm{E}(U_{1k} - U_{0k} \mid X, D = 1).$$

All of the innovations after $t > k$ are independent of D and hence α and ρ can be identified, as before, by least squares.

If the $\varepsilon_{j,t}$ are evenly partly forecastable by the agents being analyzed, then the final component of the error is correlated with $Y_{t'}$:

$$\mathrm{Cov}\left(Y_{t'}, D\sum_{j=0}^{t-t'} \rho^j(\varepsilon_{1,t-j} - \varepsilon_{0,t-j} - \mathrm{E}(\varepsilon_{1,t-j} - \varepsilon_{0,t-j} \mid X, D = 1))\right) \neq 0.$$

Since two different errors appear in the earnings stream for the $D = 1$ and $D = 0$ choices, they do not difference out as they do in the common coefficient case. In this case, the panel data form of the estimator is inconsistent for the parameter: it is necessary to instrument $Y_{t'}$.

In the general case, with $\rho_1 \neq \rho_0$, the autoregressive estimator breaks down. Different components of the error term decay at different rates, and it is no longer possible to simultaneously eliminate $U_{0t'}$ and $D(U_{1t'} - U_{0t'})$ by one substitution. Thus the method is not in general robust to heterogeneous responses. This lack of robustness to heterogeneous responses is a general feature of many of the panel data estimators discussed in Heckman and Robb (1986a).

7.6.11. Panel data estimators as matching estimators

The simple before–after estimator can be written as a matching estimator using the weighting scheme introduced in Section 7.4.1. To begin, accept assumption (4.A.1) as

valid. For person i at time $t > k$ (k is the program participation period in the notation of Section 4) who has participated in the program, the match is with himself/herself in period $t' < k$. Assume a stationary environment. Letting the match partner be the same individual at time $t' < k$, we match $Y_{0,i,t'}$, $t' < k$ to obtain the following:

$$Y_{1,i,t} - W(i,t')Y_{0,i,t'}, \qquad \text{for } t' < k,$$

where the weight $W(i,t') = 1$. More generally if we have access to more than one pre-program observation per person, one can weight the various terms by functions of the variances determined using the optimal weighting schemes in minimum distance estimation (see Heckman, 1998c, for details.) Thus the comparison group for person i at time t is a weighted average of the available observations for that person over the pre-program observation period:

$$Y^c_{0,i,t} = \sum_{j=0}^{k-1} W(i,j)Y_{0,i,j}, \qquad \text{for } j < k, \tag{7.31}$$

where

$$\sum_{j=0}^{k-1} W(i,j) = 1.$$

Each post-program period can be matched in this way with the pre-program observations. The weights can be chosen to minimize the variance in the sum of the contrasts. (Heckman, 1998c). Assuming that the same treatment effect characterizes all post-program periods, and summing over all post-program observations, we can estimate the treatment on the treated parameter by the sample analog of

$$\sum_{t=k+1}^{T} (Y_{1,i,t} - Y^c_{0,i,t})\varphi(i,t),$$

where

$$\sum_{t=k+1}^{T} \varphi(i,t) = 1$$

and $\varphi(i,t)$ are weights chosen to minimize the variance of this expression. If the treatment effects are different for each post-program period, there is no point in summing across post-program periods.

There is no necessary reason why the weights should be the same on the components. Thus we may write

$$\sum_{t=k+1}^{T} (\alpha\,(i,t)Y_{1,i,t} - \beta\,(i,t)Y^c_{i,t}),$$

provided that

$$\sum_{t=k+1}^{T} \alpha\,(i,t) = 1 \qquad \text{and} \qquad \sum_{t=k+1}^{T} \alpha\,(i,t) = \sum_{t=k+1}^{T} \beta\,(i,t),$$

for all i. These conditions enable us to difference out common components and retain identification of $\mathrm{E}(\alpha \mid X,\ D = 1)$.

If there are trends operating on participants, it is necessary to eliminate them to estimate the parameter of interest. If the trends are common across participants, we are led to using the differences-in-differences method as long as assumption (4.A.2) is valid. In this setting, it is necessary to use a group of persons who do not receive treatment. Accordingly, we can think of creating a comparison person i' for treatment person i:

$$Y_{0,i',t} - \sum_{j=1}^{k-1} W(i',j) Y_{0,i',j}, \qquad \text{for } t > k > j,$$

where

$$\sum_{j=1}^{k-1} W(i',j) = 1 \qquad \text{and} \qquad W(i,j) = W(i',j),$$

for all i,i' and j. This transforms the comparison group to be conformable with the treatment group. We thus create a pairing $i \rightarrow i'$, such that persons i and i' have the same weights, i is in the treatment group and i' is in the comparison group, and we can form the difference-in-differences estimator for person i paired with person i' as follows:

$$\left[Y_{1,i,t} - \sum_{j=1}^{k-1} W(i,j) Y_{0,i,j} \right] - \left[Y_{0,i',t} - \sum_{j=0}^{k-1} W(i',j) Y_{0,i',j} \right] \tag{7.32}$$

and $W(i,j) = W(i',j)$ for any (i,i') and all j and where

$$\sum_{i} W(i,j) = 1 \qquad \text{and} \qquad \sum_{i'} W(i',j) = 1.$$

This procedure eliminates common trends and weights the comparison group and treatment group symmetrically. Different weights are required for models with different serial correlation properties (Heckman, 1998c).

More generally, we can form other pairings in the comparison group and compare i to an entire collection of non-treated persons who are operated on by a common trend. For example, we can form an alternative difference-in-differences estimator as follows:

$$\left[Y_{1,i,t} - \sum_{j=0}^{k-1} W(i,j) Y_{0,i,j} \right] - \frac{1}{N_c} \sum_{i'=1}^{N_c} \left[Y_{0,i',t} - \sum_{j=0}^{k-1} W(i',j) Y_{0,i',j} \right] \varphi(i'), \tag{7.33}$$

where N_c is the number of persons in the comparison sample, $\varphi(i')$ is a weight and where

$$\frac{1}{N_c}\sum_{i'=1}^{N_c}\varphi(i') = 1 \qquad \text{and} \qquad \frac{1}{N_c}\sum_{i'=1}^{N_c} W(i',j)\varphi(i') = W(i,j).$$

Difference (7.33) eliminates age- or period-specific common trends or year effects. We can form variance weighted versions of (7.33) to pool information across i to estimate $\mathrm{E}(Y_1 - Y_0 \mid X, D = 1)$ efficiently if the effect is constant (see Heckman, 1998c).

The same scheme can be used to estimate models with person-specific, time-varying variables. Time-invariant variables are eliminated by subtraction. Consider the before–after estimator. Let $A_{it}(Y_{it})$ be an "adjustment" to Y_{it}, where

$$A_{it}(Y_{i,t}) = Y_{i,t} - g(X_{i,t}).$$

Then the comparison group for person i based on his preprogram adjusted outcomes can be written as

$$A_{it}^{c}(Y_{i,t}) = \sum_{j=0}^{k-1} W(i,j)A_{jt}(Y_{0,i,j})$$

and the before–after estimator can now be written in terms of adjusted outcomes as follows:

$$A_{it}(Y_{1,i,t}) - A_{i,t}^{c}(Y_{i,t}).$$

We can make a similar modification to the difference-in-differences scheme:

$$\left[A_{it}(Y_{1,i,t}) - \sum_{j=0}^{k-1} W(i,j)A_{jt}(Y_{0,j,t})\right] - \left[A_{i',t}(Y_{1,i',t}) - \sum_{j=0}^{k-1} W(i',j)A_{i',j}(Y_{0,i',t})\right],$$

where $W(i,j) = W(i',j)$ for all i,i', and

$$\sum_{j=0}^{k-1} W(i',j) = 1 \qquad \text{and} \qquad \sum_{j=0}^{k-1} W(i,j) = 1.$$

This modification eliminates non-invariant components. This enables us to generalize the simple before-after estimator to a case where person-specific and period-specific shocks operate on agents. This produces a large class of longitudinal estimators as special cases of the weighting scheme introduced in our discussion and is the basis for a unified treatment of a variety of evaluation estimators. Heckman (1998a) presents a comprehensive analysis and many examples of weights for different traditional econometric estimators.

7.7. Robustness to biased sampling plans

Virtually all estimation methods can be readily adjusted to account for choice-based sampling (i.e., oversampling of trainees relative to comparison group members) or

measurement error in training status among the comparison group (some of the comparison group members have taken training). Some methods require no modification at all.

The data available for analyzing the impact of training on earnings are often nonrandom samples. Frequently they consist of pooled data from two sources: (a) a sample of trainees selected from program records and (b) a sample of non-trainees selected from some national sample. Typically, such samples overrepresent trainees relative to their proportion in the population. This creates the problem of choice-based sampling first analyzed in a more general form by Rao (1965, 1986) and applied by Manski and Lerman (1977) and Manski and McFadden (1981).

A second problem, contamination bias, arises when the training status of certain individuals is recorded with error. Many control samples such as the US Current Population Survey or the US Social Security Work History data do not reveal whether or not persons have received training. These sampling situations combine the following types of data:

(A) outcomes, observable characteristics and participation status for a sample of trainees ($D = 1$);

(B) outcomes, observable characteristics and participation status for a sample of non-trainees ($D = 0$);

(C) outcomes and observable characteristics for a national comparison sample of the population (e.g., CPS or Social Security records) where the training status of persons is not known.If type (A) and (B) data are combined and the sample proportion of trainees does not converge to the population proportion of trainees, the combined sample is a choice-based sample. If type (A) and (C) data are combined with or without type (B) data, there is contamination bias because the training status of some persons is not known.

We can modify most procedures developed in the context of random sampling to consistently estimate $\mathrm{E}(\alpha \mid X, D = 1)$ using choice-based samples or contaminated comparison groups. In some cases, a consistent estimator of the population proportion of trainees is required. We illustrate these claims by showing how to modify the instrumental variables estimator to address both sampling schemes. We briefly consider several other methods as well. Heckman and Robb (1985a, 1986b) give explicit case-by-case treatment of these issues for a variety of estimators including all of the panel data estimators considered in this paper.

7.7.1. The IV estimator and choice-based sampling

If condition (7.17b) is strengthened to read

$$\mathrm{E}(U_0 \mid X, Z, D) = \mathrm{E}(U_0 \mid X), \qquad \text{for } D = 0, 1, \tag{7.17b$'$}$$

the IV estimator is consistent for $\mathrm{E}(\alpha \mid X, D = 1)$ in choice-based samples. The important point to notice is that identification condition (7.17b) is written for the population. By contrast, (7.17b$'$) is written for a subset of the population conditional on $D = 1$ or $D = 0$. If we reformulate the IV condition to apply to the $D = 0$ and $D = 1$ subpopulations, it does not matter how we reweight the subpopulations to form samples – the orthogonality conditions apply to any combinations of them.

To see how to form consistent estimators under the assumptions of Section 7.4.3, let D^* be the event that "a trainee is observed in a choice-based sample." In a sample generated by choice-based sampling, the probability of participation $\Pr(D^* = 1) = P^* \neq P = \Pr(D = 1)$, where P is the probability of participation in the case of random sampling.

Now in the choice-based sample, let U_0^* be the random variable U_0 generated from choice-based sampling, so that

$$\mathrm{E}(U_0^* \mid X, Z) = \mathrm{E}(U_0 \mid X, Z, D^* = 1)P^* + \mathrm{E}(U_0 \mid X, Z, D^* = 0)(1 - P^*).$$

If (7.17b′) applies, then we can write

$$\mathrm{E}(U_0^* \mid X, Z) = \mathrm{E}(U_0 \mid X, Z)P^* + \mathrm{E}(U_0 \mid X, Z)(1 - P^*) = \mathrm{E}(U_0 \mid X, Z).$$

Provided P is known, it is possible to reweight the data to secure consistent IV estimators for $\mathrm{E}(\alpha \mid X, D = 1)$ under the assumptions of Section 7.4.3. Simply multiply both dependent and independent observations by the weight

$$\omega = D\frac{P}{P^*} + (1 - D)\left(\frac{1 - P}{1 - P^*}\right)$$

and apply IV to the transformed data. This weighting ensures that (7.17b) applies to the reweighted data. The IV method applied to the reweighted samples consistently estimates the parameters of interest provided that other identifying assumptions are maintained (see Heckman and Robb, 1985a, 1986a).

7.7.2. The IV estimator and contamination bias

For data of type (C), D is not observed. Applying the IV estimator to pooled samples of type (A) and (C) data assuming that all observations in the type (C) data have $D = 0$ produces an inconsistent estimator if the type (C) data includes some trainees. However, with a minimal amount of additional information, it is possible to identify the estimator in this case.

In terms of the IV Eqs. (7.18) or (7.20), it is possible to generate $\mathrm{E}(Y \mid X, Z)$ from the type (C) sample. The type (A) data yield the sample joint distribution of (Y,X,Z) given $D = 1$ and in particular the joint distribution $f(X, Z \mid D = 1)$. Since we know

$$f(X, Z) = f(X, Z \mid D = 1)P + f(X, Z \mid D = 0)(1 - P),$$

we can solve for $f(X, Z \mid D = 0)$ if we know P. From Bayes' rule, we can write (denoting "f" as the density)

$$\Pr(D = 1 \mid X, Z) = \frac{f(X, Z, D = 1)}{f(X, Z)}.$$

The two densities can be constructed from the information in the type (C) and type (A) samples. Thus with knowledge of P, it is possible to estimate $\Pr(D = 1 \mid X, Z)$ for each person and hence to construct the IV estimator for contaminated samples. One can think of this procedure as a data imputation exercise. See Heckman and Robb (1985a,

1986a), Imbens and Lancaster (1996) and Heckman (1998a) for the econometric details.

7.7.3. Repeated cross-section methods with unknown training status and choice-based sampling

The repeated cross-section estimators discussed in Section 7.6.5 are inconsistent when applied to choice-based samples unless additional conditions are assumed.[74] For most of the repeated cross-section estimators, it is necessary to know the identity of the trainees to weight the sample back to the proportion of trainees that would be produced by a random sample to obtain consistent estimators. Hence, the class of estimators that does not require knowledge of individual training status is not robust to choice-based sampling.

Some of the estimators that we have examined are robust to choice-based sampling. Any estimator that is constructed conditional on D has the property of being robust to choice-based sampling. (Recall our discussion of instrumental variables estimators where the condition (7.17b) was modified to hold conditionally on D.) A control function estimator constructs

$$\mathrm{E}(U_{1t} \mid X, Z, D), \tag{7.34a}$$

$$\mathrm{E}(U_{0t} \mid X, Z, D), \tag{7.34b}$$

and works with the purged residuals

$$U_{1t} - \mathrm{E}(U_{1t} \mid X, Z, D)$$

and

$$U_{0t} - \mathrm{E}(U_{0t} \mid X, Z, D)$$

from the original model. Then the parameters of (7.34a) and (7.34b) are estimated along with the remaining parameters of the model. Identification conditions for control function models are given in Heckman and Robb (1985a).[75] The selection bias terms $K_0(P(Z))$ and $K_1(P(Z))$ in Eqs. (7.16a) and (7.16b) are examples of control functions with the inverse Mills' ratio as the leading example used in empirical work. Likewise, the autoregressive estimator of Heckman and Wolpin (1976) discussed in Section 7.6.3 is a control function estimator where

$$K_t = \rho^{t-t'} U_{t'}, \qquad \text{for } t > t' > k$$

and where $Y_{t'} - \beta_{t'} - \alpha D = U_{t'}$. The higher-order autoregression schemes discussed in

[74] This is not always true. For example, when the environment is time homogeneous, $(\bar{Y}_t - \bar{Y}_{t'})/P$ remains a consistent estimator of $\mathrm{E}(\alpha \mid X, D = 1)$ in choice-based samples as long as the same proportion of trainees are sampled in periods t' and t.

[75] They present conditions under which it is possible to identify the control functions apart from the parameters of the model. See also Heckman (1990).

Heckman and Robb (1985a, p. 223) are also control functions. They discuss additional control functions based on factor models and optimal forecasting schemes.

The basic principle of the control function is that of constructing conditional means of the errors in each regime ($D = 0, 1$) and estimating these conditional means and the other parameters of the model. As long as the control function is defined to be conditional on D, the control estimator is robust to choice-based sampling.

7.8. Bounding and sensitivity analysis

Since the problem of "causal analysis" is intrinsically a missing data problem, methods from the missing data literature can be used to solve the problem of causal inference, and to provide bounds on the missing data. Various bounding schemes proposed in the recent literature can be regarded as applications of the 1970s and 1980s literature on missing data.

The prototype for this approach is presented in a paper by Smith and Welch (1986) who consider both a sensitivity analysis and a bounding analysis in examining the effect of selection bias on the measured wage of blacks. Commenting on a paper by Butler and Heckman (1977), who attribute some part of the growth in black real wages observed in the US in the 1960s to selective withdrawal of the least skilled workers from the labor force, Smith and Welch (1986) apply the law of iterated expectations to write the true wage of all blacks $\mathrm{E}(W_{\mathrm{B}})$ as

$$\mathrm{E}(W_{\mathrm{B}}) = \mathrm{E}(W_{\mathrm{B}} \mid L_{\mathrm{B}} = 1)P(L_{\mathrm{B}} = 1) + \mathrm{E}(W_{\mathrm{B}} \mid L_{\mathrm{B}} = 0)P(L_{\mathrm{B}} = 0), \tag{7.35}$$

where $\mathrm{E}(W_{\mathrm{B}} \mid L_{\mathrm{B}} = 1)$ is the wage of black workers, $\mathrm{E}(W_{\mathrm{B}} \mid L_{\mathrm{B}} = 0)$ is the wage of black labor force dropouts would have received if they would have worked and $\mathrm{E}(W_{\mathrm{B}})$ is the mean wage of all blacks.[76] $P(L_{\mathrm{B}} = 1)$ is the proportion of the black population that is working. Observed (consistently estimable) are $\mathrm{E}(W_{\mathrm{B}} \mid L_{\mathrm{B}} = 1)$ and $P(L_{\mathrm{B}} = 1)$ (and hence $1 - P(L_{\mathrm{B}} = 1)$). Missing data on the wages of non-participants make $\mathrm{E}(W_{\mathrm{B}} \mid L_{\mathrm{B}} = 0)$ non-identified and hence $\mathrm{E}(W_{\mathrm{B}})$ is non-identified.

Smith and Welch (1986) adopt several solutions to this identification problem which have been widely applied in the evaluation literature. The first is to use panel data to follow non-workers over time and find the wage that is observed most recently to replace the missing wage. The second is to bound the missing parameter $\mathrm{E}(W_{\mathrm{B}} \mid L_{\mathrm{B}} = 0)$ assuming that $[\mathrm{E}(W_{\mathrm{B}} \mid \mathrm{L}_{\mathrm{B}} = 0) = \gamma \mathrm{E}(W_{\mathrm{B}} \mid L_{\mathrm{B}} = 1)]$ for $0.5 \leq \gamma \leq 1$. By varying γ over a range of values, they perform a sensitivity analysis or bounding analysis that has recently become fashionable in applied social science. Their methods apply directly to the selection problem. Suppose we know $\mathrm{E}(Y_0 \mid D = 0)$. We seek to know $\mathrm{E}(Y_0 \mid D = 1)$ to construct the counterfactual $\mathrm{E}(Y_1 - Y_0 \mid D = 1)$. By using bounds connecting $\mathrm{E}(Y_0 \mid D = 0)$ to $\mathrm{E}(Y_0 \mid D = 1)$, it is possible to bound $\mathrm{E}(Y_1 - Y_0 \mid D = 1)$. (Recall that $\mathrm{E}(Y_1 \mid D = 1)$ is known).

Glynn and Rubin (1986) present a similar analysis of what they call "mixture models."

[76] We use a simplified notation to convey the main idea in their work.

Like Smith and Welch (1986), they analyze two cases: (a) one where the missing data in one period can be obtained in another period and (b) one where they perform a "sensitivity" analysis by varying the unidentified parameters of the model. Rosenbaum (1995) summarizes a series of papers going back to the late 1950s that bound estimated causal effects by bounding the range of the unobserved parameters of the model.

In this section of the chapter, we draw on the comprehensive analysis of Balke and Pearl (1993, 1997), Balke (1995), and Chickering and Pearl (1996) on bounding causal parameters. Using linear programming methods, they extend the work of Robins (1989) and Manski (1995) to present the tightest possible *non-parametric* bounds for causal parameters. These methods exploit certain classical inequalities of probability theory. Instead of analyzing a model with a high level of generality, consider a specific model of missing data that links recent analyses of bounds for causal parameters to the classical problem of missing data in contingency analysis. The Holland (1986, 1988) and Rubin (1974, 1978) model is essentially one for a contingency table with missing data. Results in the literature on missing data in the contingency tables apply directly to the model of causal effects.

Fig. 7 considers a model of potential outcomes for each person i when there are two possible values for each potential outcome $Y_0 \in \{0,1\}$, $Y_1 \in \{0,1\}$, $D \in \{0,1\}$. This produces a $2 \times 2 \times 2$ table. In the case of a randomized experiment where randomization is done after persons have attempted to enroll in the program, the row and column margins of the left ($D = 1$) table are known but not the individual cells. One piece of identifying information is missing. A monotonicity assumption (e.g., $P_{101} = 0$) fully identifies the table. This assumption says that among the persons who enter the program, there are no persons who would switch from 1 to 0 status. One can use the Frechet bounds to obtain ranges of possible values, using the column and row marginal distributions for the table (see Heckman and Smith, 1993; Heckman et al., 1997c, for discussions and the first applications of these bounds to the evaluation problem).[77] These bounds produce the tightest possible bounds on the elements of a contingency table given the marginal distributions. In practice, these bounds are usually very wide as those authors, and the vast literature in statistics that precedes them, have shown.

The more general case with observational data is one where the column totals are known for the $D = 1$ table and the row totals are known for the $D = 0$ table. The remaining elements are not known.

[77] For any joint distribution for discrete or continuous random variables, $F(a,b)$, with marginal distributions $F(a)$ and $F(b)$, $\text{Max}[F(a) + F(b) - 1, 0] \leq F(a,b) \leq \text{Min}[F(a), F(b)]$. The upper bound is a trivial consequence of the fact that $\Pr(A \leq \text{a} \cap B \leq b) \leq \text{Min}(\Pr(A \leq a); \Pr(B \leq b))$. The lower bound is equally straightforward to derive. Partition the space (A,B) into four mutually exclusive regions: $R_1 = (A \leq a, B \leq b)$, $R_2 = (A \leq a, B > b)$, $R_3 = (A > a, B \leq b)$, $R_4 = (A > a, B > b)$, where (*) is defined as $\Pr(R_1) + \Pr(R_2) + \Pr(R_3) + \Pr(R_4) = 1$. Observe that $R_1 \cup R_2 = (A \leq a)$ while $R_1 \cup R_3 = (B \leq b)$. $\Pr((R_1 \cup R_2) \cup (R_1 \cup R_3)) = \Pr(R_1 \cup R_2 \cup R_3) = 1 - \Pr(R_4)$. (**) $\Pr(R_1 \cup R_2) + \Pr(R_1 \cup R_3) = \Pr(A \leq a) + \Pr(B \leq b)$. Subtracting (**) from (*) and rearranging, we obtain $\Pr(R_1) = \Pr(A \leq a) + \Pr(B \leq b) - 1 + \Pr(R_4)$. Since $\Pr(R_4) \geq 0$, $\Pr(R_1) \geq \Pr(A \leq a) + \Pr(B \leq b) - 1$ so $F(a,b) \geq F(a) + F(b) - 1$ but since probabilities cannot go negative $F(a,b) \geq \max(0, F(a) + F(b) - 1)$.

Consider bounds for the treatment on the treated parameter TT: $\mathrm{E}(Y_1 - Y_0 \mid D = 1)$. In terms of cell proportions:

$$TT = \mathrm{E}(Y_1 - Y_0 \mid D = 1) = \frac{P_{.11} - P_{1.1}}{P_{..1}} = \frac{P_{011} - P_{101}}{P_{..1}}.$$

For the case of observational data the solution is straightforward. The linear program to bound the parameter is

Max TT subject to

$$\hat{P}_{.01} = P_{001} + P_{101} \qquad \text{(columns determined)} \tag{7.36a}$$

$$\hat{P}_{.11} = P_{011} + P_{111} \tag{7.36b}$$

and

$$\hat{P}_{0.0} = P_{000} + P_{010} \qquad \text{(rows determined)} \tag{7.36c}$$

$$\hat{P}_{1.0} = P_{100} + P_{110}. \tag{7.36d}$$

We are free to make P_{011} maximal by setting $P_{111} = 0$ (so $\hat{P}_{.11} = P_{011}$) and to make P_{101} minimal by setting $P_{001} = \hat{P}_{.01}$. No constraints are violated because we have freedom to pick the row totals in the $D = 1$ table. By the same token we can make P_{011} minimal by setting $P_{111} = \hat{P}_{.11}$ so $P_{011} = 0$ and make P_{101} maximal by setting $P_{001} = 0$ and $\hat{P}_{0.1} = P_{1.1}$.

$$\Pr(Y_1 = 1 \mid D = 1) = \frac{\hat{P}_{.11}}{P_{..1}} \geq TT \geq -\frac{\hat{P}_{.0}}{P_{..1}} = -\Pr(Y_1 = 0 \mid D = 1).$$

Access to experimental data sharpens these bounds to a point. In this case, we know both the row totals and the column totals of the $D = 1$ table. We supplement linear inequalities (7.36a) and (7.36b) by

$$\hat{P}_{0.1} = P_{001} + P_{011}, \tag{7.36e}$$

$$\hat{P}_{1.1} = P_{101} + P_{111}. \tag{7.36f}$$

Now the formal optimization problem is apparently harder; Max TT subject to (7.36a), (7.36b) and (7.36e), (7.36f). Using (7.36a) and (7.36b) we obtain

$$\hat{P}_{.11} - \hat{P}_{1.0} = P_{011} - P_{101}$$

so the parameter is exactly identified. Using the Balke–Pearl methods, we can bound any parameter, or any empty cell in a contingency table analysis, using linear programming methods.

It is important to recognize that these are *non-parametric* bounds. They *do not* capture

the full potential variability in the estimated parameter values when parametric structure is imposed on the P, as is commonly done in applied work. Nor do they capture uncertainty about the X. To get the full range of variability in the parameter requires solving the non-linear program across models M and possible regressors $\bar{\bar{X}}$ used to generate the $P_{ijk}(x,m)$, where x is a choice of regressors and m is the particular model. A full characterization of model variability in this framework is given by choosing that m and x that maximize

$$\frac{P_{011}(x,m)}{P_{..1}(x,m)} - \frac{P_{101}(x,m)}{P_{..1}(x,m)},$$

that is

$$\underset{m\in M, x\in\bar{\bar{X}}}{\text{Max}} \left[\frac{P_{011}(x,m)}{P_{..1}(x,m)} - \frac{P_{101}(x,m)}{P_{..1}(x,m)} \right]$$

subject to appropriate (i.e., modified for X and m) constraints. These bounds account for model uncertainty and regressor misspecification. A full characterization of this problem remains to be developed.

8. Econometric practice

One of the most important lessons from the literature on evaluating social programs is that choices made by evaluators regarding their data sources, the composition of their comparison groups, and the specification of their econometric models have important impacts on the estimated effects of training. As noted in Section 7, the choice of a comparison sample can affect the statistical properties of an estimator applied to that sample. Under the conditions specified there, for certain comparison groups, simple mean comparisons between treatments and controls identify the parameters of interest.

The purpose of this section is to draw from the empirical literature to show why and how these choices matter. To begin our discussion, we first discuss the types of data used in most evaluations of active labor market policies and show how the source of data affects the impact estimates. Next, we draw on the work of Heckman et al. (1996b, 1998b), who collect unusually rich data compared to what is usually available to program analysts, to analyze the sources of measured selection bias. Their findings provide an informative guide to the construction of datasets for future evaluations.

In the third section, we present a small scale simulation study of alternative evaluation estimators which make different assumptions about program participation decision rules, outcome equations and their interrelationship. This simulation study summarizes the lessons of Section 7 and reveals that no universally valid estimator exists or is ever likely to be found. In the fourth and concluding section, we consider the logic that underlies the use of widely-applied "specification tests" to check the validity of an evaluation model by determining if it "aligns" the earnings (or other measures) of participants and non-participants prior to their enrollment in the program. The method is not guaranteed to pick a

correct evaluation model. We demonstrate the practical importance of this point and show how two different alignments used in the literature produced two very different and controversial impact estimates for the same program.

8.1. Data sources

To evaluate active labor market policies requires choosing data sources from which to construct comparison groups and treatment groups. In this subsection, we discuss these issues and describe the advantages and disadvantages of the various types of data typically used to evaluate employment and training programs. The decision about what data source or data sources to use has important implications for several aspects of an evaluation. In both experimental and non-experimental evaluations, the decision affects how much the evaluation will cost, how large the analysis sample will be (which affects the size of the training effect that can be statistically distinguished), what outcome variables can be studied, the time period over which the outcome variable can be measured and the amount and type of measurement error in the outcome variable. In non-experimental evaluations, the decision also affects which of the non-experimental evaluation methods discussed in this chapter can be used and whether or not the comparison group can be located in the same local labor markets as the participants. By affecting these aspects of an evaluation, the choice of a data source affects the final impact estimates.

A comparison between the studies of Fraker and Maynard (1987) and LaLonde (1986) illustrates that the choice of a data source can vitally affect the impact estimates obtained in a social experiment. Both of these studies examined the National Supported Work Demonstration. The demonstration included one baseline and up to four followup surveys of the treatments and controls. LaLonde (1986) used this survey data for his analysis, while Fraker and Maynard (1987) used administrative data on annual earnings from the US Social Security Administration (SSA). There exist striking differences between the experimental impact estimates reported in the two studies. Using the survey data, the annual impact of Supported Work on the earnings of AFDC (welfare) women was $1641 in 1978 and $851 in 1979. By contrast, when using the SSA earnings data on the same participants and controls, the annual impact was $505 in 1978 and $351 in 1979. The different data sources produce a difference in the estimated experimental impacts of $1135 in 1978 and of $500 in 1979. The sensitivity of the impact estimates to the data used in the analysis is similar in magnitude to their sensitivity to different econometric modelling assumptions and is large enough to affect the conclusions of a cost-benefit analysis.[78]

[78] Similar sensitivity to the choice of data source was found in the National JTPA Study. For male youth, estimates using survey data showed a negative and statistically significant impact from the program, while estimates using administrative data from state Unemployment Insurance (UI) records showed essentially a zero impact. See Bloom et al. (1993). Some of the difference between the estimates shown in Table 4 based on the official 18 and 30 month NJS impact reports results from the fact that the 18 month estimates rely only on survey data while the 30 month estimates rely on a combination of survey data and earnings data from state UI records.

8.1.1. Using existing general survey data sets

In non-experimental evaluations, existing survey datasets constitute one potential source from which comparison groups can be drawn. Examples of such datasets in the US include the Current Population Survey, a large cross-sectional survey which was the source of comparison groups for some of the CETA evaluations, or the National Longitudinal Survey of Youth (NLSY), a widely used panel dataset. Such datasets are not generally used to collect information on participants because they usually collect little, if any, information on receipt of public sector training.

The key advantages of using existing datasets as a source for non-experimental comparison groups are cost and sample size. Using an existing dataset avoids the costs of designing, testing and fielding a survey as well as the costs of locating potential comparison group members. General purpose datasets typically have large samples and are available for a modest fee. Depending on the dataset, they may provide either repeated cross-sectional samples, as with the US CPS, or a long panel, as with the NLSY. In general, a large list of regressors is available for subgroup analysis.

Existing survey data have four key disadvantages for evaluation research. First, it is often difficult to construct comparison groups of persons in the same local labor markets as participants from existing datasets due to sample size limitations and constraints imposed by privacy concerns on the level of detailed locational information made available to researchers. As we show in the next section, this is a severe limitation because variation across local labor markets plays a large role in explaining the earnings and employment variation of unskilled workers who are the targets for active labor market policies (Heckman et al., 1998b). Second, in contrast to what is possible when fresh survey data are collected, it is impossible to obtain specific variables of interest for the program being evaluated not already present in the existing data. Such variables might include the detailed information on recent labor force status histories noted as important determinants of program participation in Section 6. Third, because receipt of public training is often not measured or is not measured well in these data, contamination bias becomes an issue (Heckman and Robb, 1985a) as some members of the comparison group are likely to have received the treatment being evaluated. (Recall our discussion in Section 7.7). Finally, using existing datasets to construct a comparison group often entails using different survey instruments with different definitions of the same outcome variable for participants and comparison group members in an evaluation. Comparing outcomes measured in two different ways adds an important potential source of bias to the impact estimates reported for a program (Smith, 1997b).

8.1.2. Using administrative data

Many evaluations of active labor market policies in the US and Scandinavia rely on administrative data. These pre-existing data generally consist of administrative earnings records collected for tax purposes and administrative records on social assistance receipts. They are often combined with administrative data on the receipt of training from program records. The key advantages of such data are the low cost of acquiring them and lack of

certain types of measurement error. The costs are low on several dimensions. The fixed costs of extracting administrative earnings records are typically modest compared to the costs of collecting comparable data from surveys.[79] Moreover, the marginal costs of increasing the sample size or the number of time periods of data obtained are often very small. For example, recent estimates of the marginal cost of obtaining 10 years of quarterly data on an individual's earnings and social assistance receipts are approximately $2.50. These low costs make such data particularly attractive for non-experimental evaluations in which longitudinal methods will be used. Because these earnings data are used for tax purposes, there are strong incentives for authorities to minimize reporting errors for earnings, so they are likely to be much more accurate, for the types of earnings they intend to measure, than earnings data obtained from surveys.

Administrative data also have important limitations in the context of evaluating employment and training programs. First, these data typically consist of quarterly or annual earnings and little else is reported. Monthly earnings, as well as other outcomes of interest such as wage rates, hours worked and employment spells, are nearly always unavailable. Consequently, in the US, where researchers have relied on such data, relatively little research has looked at the impact of training on wages. This outcome is of great theoretical interest, because higher wages for trainees indicate that training raised their productivity. An exclusive focus on earnings or employment rates does not determine what part of the training impact results from increased productivity of the workers as measured by their hourly wage rates and what part results from the displacement of non-trainees in the labor market (Johnson, 1979).

Second, because governments maintain administrative records for tax and benefits purposes, these earnings measures may not equal total earnings. For example, many recent US evaluations use earnings from state unemployment insurance (UI) records. These data include earnings from jobs "covered" by the UI system, but omit earnings from self-employment, from employers in other states, and from sources not covered by the UI system. As a result, administrative earnings measures tend to be lower than those reported by individuals in surveys (Kornfeld and Bloom, 1996; Smith, 1997b).

Finally, administrative data typically contain only very basic information on demographic characteristics. For example, Table 5 shows that Ashenfelter's (1978) study of MDTA, which uses detailed information on annual pre-program and post-program earnings histories from SSA records, includes only very limited demographic information - just age, sex and race. No information on labor force histories, educational levels, training history, family status or geographical location was available in the data. Lack of data on individual characteristics limits the subgroup analyses that evaluators can perform and makes it difficult to justify the application of non-experimental methods such as matching whose plausibility depends on access to a rich set of conditioning variables.

[79] There are exceptions to this rule. In the NJS, state personnel were unable to provide useable unemployment insurance earnings data in 4 of the 16 states containing training centers in the NJS despite repeated attempts.

8.1.3. Collecting new survey data

An alternative to using existing data sources is to collect fresh survey data on participants and on controls or comparison group members. This choice has both advantages and disadvantages. The first advantage relative to using either existing survey or administrative datasets is that the evaluator has complete control over the information collected on the survey, and so can design the survey in light of the variables of interest in the study and, in non-experimental evaluations, in light of the econometric methods to be used. The second advantage relative to using existing data is that the sampling plan for the survey can target comparison group members in the same local labor markets as participants. A third advantage of collecting new survey data is that relative to administrative data, the analyst can obtain additional outcome measures such as wage rates and employment transitions, and can conduct a wider variety of subgroup analyses.

The most important disadvantage of collecting fresh survey data relative to using either administrative data or existing survey datasets is the high cost of doing so. The total costs of collecting new survey data can vary widely depending on whether evaluators obtain these data through a survey sent through the mail, conducted over the telephone, or during a person-to-person interview. Surveys done through the mail are inexpensive, but typically are plagued by very low response rates; surveys conducted in person are expensive but have very high response rates. In some studies more than one type of survey is used to obtain the data. The fixed costs associated with surveys also can vary widely depending on whether evaluators use an existing survey instrument, whether the survey is automated, and whether or not the interviewers require training.

Most program evaluations based on new survey data use either a telephone or in-person survey. Phone surveys are attractive because the marginal cost of obtaining an additional response is relatively low. Such costs, which include the interview, editing, and coding of the data, are approximately $50 per observation. Longer interviews increase these costs modestly. Average costs are generally double this amount or more. Telephone surveys can be problematic especially when surveying low income populations, because response rates are often significantly lower than for in-person interviews. One practical problem in low income populations is that some respondents may not have a working telephone at the time of the survey. If the survey is done in person, the marginal cost of obtaining an additional observation more than doubles. Further, these marginal costs rise sharply with the response rate. Additional respondents become harder to find. Average costs for samples of modest size obtained from in-person surveys range as high as $500 per observation. If evaluators wish to return and resurvey the same sample at a later date, these costs may not fall appreciably in the second wave. Low income populations are often highly mobile and resources must be expended locating persons who have moved.

The costs of collecting new survey data are likely to be lower for program participants than for members of the comparison group. To obtain a sample of participants, evaluators can use the administrative records that contain information such as the individual's address, phone numbers, and sometimes the most recent employer. Such information is advantageous, because locating respondents is an important component of the cost of

surveys. In contrast, obtaining information on a comparison group requires evaluators to construct a list of comparable persons. One criterion for sample inclusion might be to include persons eligible for the program, who did not participate. An advantage of using non-participating program applicants is that they constitute a ready list from which evaluators can sample for their survey. Another method for selecting a comparison group is to first conduct a short "screening survey" to obtain a list of individuals who were eligible for the program, but did not participate. Even in low income neighborhoods, the fraction of respondents found to be eligible is typically low, so evaluators must conduct many short interviews to obtain a sufficient number of comparisons. Even when using a telephone survey, these procedures can double the marginal cost of obtaining an observation for a comparison group member.

Collecting new survey data can be expensive. As a result, there is no reason to expect that careful non-experimental evaluations that collect new survey data are appreciably less costly than experimental evaluations. The marginal cost per participant of administering an experiment is small. The cost of obtaining a high quality comparison group in a non-experimental evaluation can be very high. A dramatic example of the high cost of collecting new survey data in a non-experimental evaluation is the cost of obtaining the non-experimental comparison group used in the NJS. This sample cost $3.5 million (1990) to collect responses from 3000 persons, in two waves, from just four of the 16 sites included in the study (Smith, 1994). Most of these responses were obtained using a telephone survey. The average cost was a little more than 1000 per observation. Particularly large were the costs associated with locating eligible persons not participating in JTPA.[80]

Related to the general issue of cost is an important tradeoff that affects evaluators who collect their own survey data. Researchers often seek longterm followup data on outcomes, to determine whether shortterm program impacts persist. Non-experimental researchers planning to use many of the longitudinal methods considered in Section 7.6 require information on outcomes in periods prior to the decision to participate in training. The marginal cost of obtaining additional periods of outcome data either before or after participation is usually low for administrative data. With survey data, the evaluator must choose between constructing a panel by fielding costly additional surveys, or tolerating the degradation of data quality as the length of the survey recall period increases.[81]

8.1.4. Combining data sources

One solution to the limitations of any particular type of data is to construct a new dataset by combining more than one type of data. Evaluators often combine administrative data on outcomes with survey data on the characteristics of participants and of comparison or control group members. Analysts then have access to relatively rich data on individual

[80] We are grateful to personnel at Mathematica Policy Research, MDRC, NORC, Westat, and the W.E. Upjohn Institute for Employment Research for providing us with information on the cost of collecting survey data.

[81] Bound et al. (1994) provide evidence of recall effects in labor market survey data and Sudman and Bradburn (1982) discuss the general issue of recall in surveys.

regressors as well as a long panel of earnings data that allows implementation of longitudinal estimators of program impact. For example, many of the US CETA evaluations use a dataset that combines program records on trainees with comparison group data drawn from the CPS. The dataset includes matched administrative earnings data from the Social Security Administration for both groups. However, because the comparison group is drawn from the existing CPS dataset, it is not possible to match them to participants in the same local labor markets.

The NJS provides an example of a study which combined new survey data with administrative data. In this evaluation, treatment and control group members completed a baseline survey and one or two followup surveys. These data were combined with administrative earnings data from state UI systems, administrative income data from the US Internal Revenue Service and administrative data on social assistance from state welfare agencies. The NJS also collected both survey and administrative data on its non-experimental comparison group sample. Because the NJS researchers collected fresh survey data rather than using an existing dataset, they were able to locate comparison group members in the same local labor markets as participants.

8.2. *Characterizing selection bias*

We next draw on the work of Heckman et al. (1996b, 1997a, 1998b), and demonstrate the value of better data in conducting evaluations of active labor market policies. Placing people in the same local labor market and administering them the same survey instrument makes an enormous difference to the quality of an evaluation. So does comparing comparable people. We also summarize the best available evidence on the validity of the widely used practice of using "no shows" as a comparison group.

The mean selection bias in using non-participants to approximate participant outcomes conditional on X is given by

$$B(X) = \mathrm{E}(Y_0 \mid X, D = 1) - \mathrm{E}(Y_0 \mid X, D = 0). \tag{8.1}$$

Selective differences in uncontrolled variables (variables on which the analyst cannot condition) produce selection bias. Such differences may arise from self-selection decisions by the agents being studied or from uncontrolled differences between treatments and controls due to the inadequacy of the available data. We argue that much of the bias reported by LaLonde (1986) in his influential study of the effectiveness of econometric estimators arises from the second source – the inadequacy of the data. In ordinary non-experimental evaluations, B is unknown. This produces the evaluation problem. Using data from a social experiment conducted under the conditions specified in Section 5, it is possible to estimate the first term on the right hand side of (8.1). Using a non-experimental comparison group it is possible to estimate the second term.

The conventional measure of selection bias, B, used by LaLonde (1986), Ashenfelter (1978) and Heckman and Hotz (1989) is the mean difference between the earnings of controls and the earnings of comparison group members:

$$B = \mathrm{E}(Y_0 \mid D = 1) - \mathrm{E}(Y_0 \mid D = 0).$$

This is the coefficient on D of a regression of Y_0 on D in a pooled comparison group and control group sample, $Y_0 = \alpha_0 + BD + \tau$ when $\mathrm{E}(\tau \mid D) = 0$. It does not condition on X.

Heckman et al. (1996b, 1998b) estimate the bias term $B(X)$ using non-parametric methods. With their estimated bias, they test the identifying assumptions that justify matching, the classical econometric selection bias estimator and a non-parametric version of difference-in-differences. They show that it is possible to decompose the conventional measure of bias, B, which does not condition on X, into three components. The first component of B results from the fact that for certain values of X among participants there may be no comparison group members, and vice versa – in formal terms the supports (regions of X where the density function is not zero) of X in the participant and comparison groups may not completely overlap. The second component results from differences in the distribution of X between participants and comparison group members within the region of common support; i.e., for those values of X common to the two groups. The third component represents selection on unobservables as defined in Section 7. This decomposition is helpful for understanding the sources of selection bias as it is conventionally measured.

To reduce the set of conditioning variables, X, down to manageable size, Heckman et al. (1996b, 1998b) condition on the probability of program participation, $P(X)$, rather than directly on X. This is always possible, because we may write the outcome in the absence of training for the experimental controls as follows:

$$Y_0 = \mathrm{E}(Y_0 \mid P(X), D = 1) + V_1,$$

where $\mathrm{E}(V_1 \mid P(X), D = 1) = 0$. The corresponding expression for the comparison group members is given by

$$Y_0 = \mathrm{E}(Y_0 \mid P(X), D = 0) + V_0,$$

where $\mathrm{E}(V_0 \mid P(X), D = 0) = 0$. The residuals average out to zero within participant ($D = 1$) and non-participant ($D = 0$) samples.[82]

Using these methods, this bias B can be decomposed into three components:[83]

$$B = \mathrm{E}(Y_0 \mid D = 1) - \mathrm{E}(Y_0 \mid D = 0) = B_1 + B_2 + B_3. \tag{8.2}$$

To help define B_1, we first define S_P as the common support – the set of $P(X)$ values common to the $D = 1$ and $D = 0$ samples. In addition, let S_{1P} denote the set of $P(X)$ values found in the $D = 1$ sample and S_{0P} the set found in the $D = 0$ sample. The first bias term is given by

$$B_1 = \int_{S_{1P} \setminus S_P} \mathrm{E}(Y_0 \mid P(X), D = 1) \mathrm{d}F(P(X) \mid D = 1)$$

[82] This is a valid decomposition whether or not matching is a valid evaluation estimator.

[83] This decomposition was first published in Heckman et al. (1996b).

$$-\int_{S_{0P}\backslash S_P} \mathrm{E}(Y_0 \mid P(X), D = 0)\mathrm{d}F(P(X) \mid D = 0),$$

where $S_{1P}\backslash S_P$ is the subset of S_{1P} not in S_P, i.e., the set of $P(X)$ values present in the $D = 1$ sample but not in the $D = 0$ sample. The set $S_{0P}\backslash S_P$ is defined comparably for the $D = 0$ group. The second bias term arises from the different densities of $P(X)$ in the $D = 1$ and $D = 0$ samples:

$$B_2 = \int_{S_P} \mathrm{E}(Y_0 \mid P(X), D = 0)[\mathrm{d}F(P(X) \mid D = 1) - \mathrm{d}F(P(X) \mid D = 0)].$$

The third bias term is the contribution of selection bias rigorously defined:

$$B_3 = P_X \bar{B}_{S_P},$$

where

$$\bar{B}_{S_P} = \frac{\int_{S_P} B(P(X))dF(P(X) \mid D = 1)}{\int_{S_P} dF(P(X) \mid D = 1)},$$

is the average selection bias defined over the common support set, S_P, and $B(P(X)) = \mathrm{E}(U_0 \mid P(X), D = 1) - \mathrm{E}(U_0 \mid P(X), D = 0)$ is the selection bias at each point.

The first term on the right-hand side of (8.2) is the difference between the mean earnings of the controls and the comparison group members in the region outside the common support – that is, for those values of $P(X)$ that appear only among controls or only among comparison group members. This is the bias that arises from comparing non-comparable people – persons in $D = 1$ who have no counterpart in $D = 0$ and vice versa. The second term gives the bias due to the different densities of $P(X)$ in the control and comparison groups over the region in which the densities of $P(X)$ for the two groups overlap. This is the bias that arises from weighting comparable people incomparably.

Finally, the third term, or the "true" selection bias, is the weighted (by the distribution of $P(X)$ for controls) average difference between the earnings of controls and comparisons who have the same $P(X)$. If matching is an effective evaluation method, the third term, B_3, representing selection on unobservables, should be zero or close to it. Recall from the discussion in Section 7.4.1 that under the assumptions that justify matching, $B(P(X))=0$ for all $P(X)$. We can interpret estimates of this term as a measure of the extent to which matching does not balance the bias between treatment and comparison group members.

Heckman et al. (1996b, 1998b) estimate the components of selection bias using the experimental controls from the NJS and a sample of eligible non-participants (ENPs) from the same sites as well as using other, more traditional comparison groups of the sort discussed in Section 7.2.[84] Fig. 10A plots the densities of $P(X)$ for adult male controls

[84] Heckman et al. (1996b, 1998b) estimate the $\mathrm{E}(Y_0 \mid P(X), D = 0)$ terms using a local linear regression of the outcome Y_0 on $P(X)$. The estimates of $P(X)$ are obtained from logit models of participation in the JTPA program, but estimates using non-parametric P are very similar.

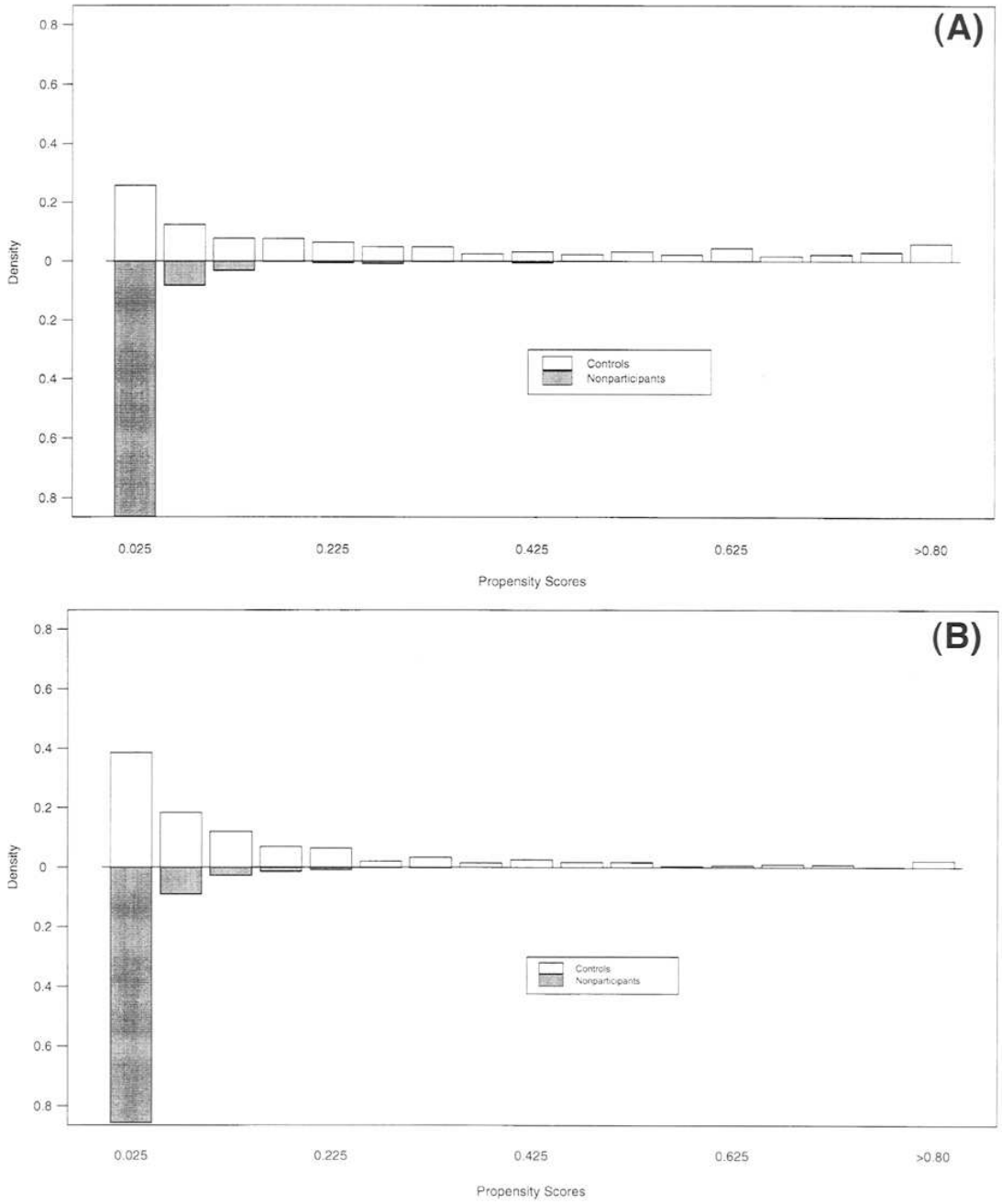

Fig. 10. Density of estimated probability of program particiaption for adult male (A) and female (B) controls and eligible non-participants in the National JTPA Study.

and ENPs. Fig. 10B plots the densities of the $P(X)$ for adult female controls and ENPs. In both groups, for a substantial range of $P(X)$ values in the control sample, there are few or no corresponding comparison group members. Among the adult males, nearly one half of the controls' $P(X)$ values are outside the region of overlapping support.

Table 8 presents estimates of the decomposition in (8.2) for adult males and females in the NJS. As shown by the second row of Table 8, differences in the support of $P(X)$ are an important source of bias. This source of bias is of at least the same order of magnitude as the conventional measure of selection bias presented in the first row of the table. The third row of Table 8 indicates that differences in the distributions of $P(X)$ between control and comparison group members in the region of common support are an important source of bias. Finally the fourth, fifth and sixth rows of the table show that for both groups the selection bias term, B_3, is relatively small compared to the other components of B, the bias as conventionally measured. However, B_3 is still quite large compared to the estimated program impact. This result indicates that matching on $P(X)$ mitigates but does not eliminate selection bias in the NJS data. Selection on unobservables is a substantial component of the experimentally estimated impact of treatment even using the rich data available in the NJS. It is likely to be even more important in cruder datasets, as we document below.

Eliminating selection bias in most non-experimental evaluations may be even more difficult than is suggested by Table 8. The NJS eligible non-participant comparison group was constructed specifically for the purpose of conducting a high quality non-experimental

Table 8
Decomposition of differences in mean earnings for adult participants in the US National JTPA Study (mean monthly earnings differences between experimental controls and comparison sample of eligible non-participants during the 18 months following the baseline in four sites)[a]

	Adult males	Adult females
Mean difference in earnings = B	−337	33
	(47)[b]	(26)
Non-overlapping support = B_1	298	106
	(35)	(13)
	[−88][c]	[318]
Different density weighting of propensity scores = B_2	−659	−118
	(42)	(20)
	[195]	[−355]
Selection bias = B_3	24	45
	(28)	(26)
	[−7]	[136]
Average selection bias when matching only in regions of common support	48	59
Selection bias as a percent of treatment impact	109	202
Control group sample size	508	696
Comparison group sample size	388	866

[a] Source: Heckman et al. (1996b, Table 1, p. 13418).

[b] The numbers in parentheses are the bootstraped standard errors. They are based on 50 replications with 100% sampling.

[c] The numbers in square brackets are the percentage of the mean difference in earnings (row 1) attributable to each component of the bias.

evaluation of JTPA. These data contain many more demographic and baseline characteristics than are commonly available to program evaluators. Further, the comparison group members reside in the same labor market as the trainees, are administered the same survey instruments, and are all eligible for JTPA. The encouraging news from the analyses of Heckman et al. (1997a, 1998b,c) is that less expensive comparison groups that contain limited labor force status histories but still place comparison group members in the same local labor markets as participants and administer the same surveys to both groups should do just as well as the richer data.

Table 9 presents the decomposition when no-shows are used as a comparison group. In the context of the NJS, no-shows are persons randomly assigned to the experimental treatment group who never enroll in JTPA and do not receive JTPA services (these are the dropouts of Section 5). In the absence of an experiment, no-shows are usually persons who enroll in a program but drop out prior to service receipt. Cooley et al. (1979) and Bell et al. (1995) advocate the use of no-shows as a comparison group. On a priori grounds, no-

Table 9
Decomposition of differences in mean earnings in the US National JTPA Study (mean monthly earnings differences during the 18 months following the baseline in four sites, no-shows)[a]

	Experimental controls and treatment group dropouts[b]		Experimental controls and SIPP eligibles[c]	
	Adult males	Adult females	Adult males	Adult females
Mean difference in earnings = B	29	9	−145	47
	(38)[d]	(23)	(56)	(23)
Non-overlapping support = B_1	−13	1	151	97
	(12)	(6)	(30)	(19)
	[−45][e]	[9]	[−104]	[206]
Different density weighting of propensity scores = B_2	3	−9	−417	−172
	(16)	(10)	(44)	(16)
	[11]	[−99]	[287]	[−367]
Selection bias = B_3	38	18	121	122
	(37)	(26)	(33)	(15)
	[135]	[190]	[−83]	[260]
Average selection bias when matching only in regions of common support	42	20	192	198
	(40)	(29)	(57)	(26)
Selection bias as a percent of treatment impact	97	68	440	676

[a] Source: Heckman et al. (1997a, Table 2).

[b] Treatment group dropouts (or "no-shows") are persons randomly assigned to the experimental treatment group who failed to enroll in JTPA.

[c] The SIPP eligibles are persons in the 1998 SIPP full panel who were eligible in month 12 of the 24 month panel using eligibility definition "B" from Devine and Heckman (1996).

[d] Bootstrap standard errors appear in parentheses. They are based on 50 replications with 100% sampling.

[e] The numbers in square brackets are the percentage of the mean difference in earnings (row 1) attributable to each component of the bias.

shows are not necessarily an attractive comparison group. Selective differences in unobservables between participants and no-shows will make the latter a poor comparison group if selection on unobservables (conditional on applying to and being accepted into the program) is an important component of bias. Yet, at the same time, no-shows are an attractive comparison group because they are located in the same labor market and administered the same questionnaire as participants.

The first two columns of Table 9 present the decomposition in (8.2) constructed using the experimental controls and the no-shows from the NJS. Fig. 11A,B presents the densities of $P(X)$ for the same groups. There is much more overlap in the supports of the no-show and control groups than there is in the comparison and control groups. Moreover, the shapes of the distributions of P are closer for no shows and control group members than they are for comparisons and controls (cf. Fig. 10A,B with Fig. 11A,B, respectively).

The evidence on no-shows is mixed. The raw measure of bias B is small for both males and females. In addition, the support and density weighting problems are much smaller than those reported in Table 8, although part of this difference results from the smaller set of X's available in the NJS data to construct $P(X)$ for the no-shows. However, as shown in the final row of Table 9, the selection bias for the no-shows remains sizeable when measured as a percentage of the treatment impact.

The biases obtained for the no-shows in the NJS or the comparison group are much smaller than the biases that result from comparing the NJS controls to a comparison group constructed from a general survey dataset. The last two columns of Table 9 present the bias decompositions based on a comparison group of persons eligible for JTPA drawn from the US Survey of Income and Program Participation (SIPP). The SIPP is a national survey dataset of the type widely used in evaluating active labor market policies. SIPP data are rich enough to determine program eligibility. The comparison group constructed from it is not drawn from the same local labor markets as the NJS control group due to sample size and confidentiality limitations. Moreover, the earnings measure in the SIPP differs substantially from that used for the NJS controls due to differences in the respective survey instruments (Smith, 1997a,b).

A comparison of the first rows of Tables 8 and 9 shows that for the SIPP eligible comparison group, the raw bias, B, is actually smaller for adult males than with the ENP comparison group. The raw bias is about the same magnitude for adult females using the two comparison groups, although of a different sign. However, B_3, selection bias rigorously defined, is much larger for the SIPP eligible comparison group than for the eligible non-participant comparison group in Table 8. This indicates that mismatch of labor markets and questionnaires between participants and comparison group members is a major source of selection bias.

Heckman et al. (1998b) examine these issues in greater depth. In particular, using the NJS data on controls and ENPs, they match controls at two sites with ENPs at the two remaining sites. This comparison shows the effect of putting comparison group members in different local labor markets while holding constant the survey instrument used to measure earnings in the two groups. They find that mismatching the local labor markets

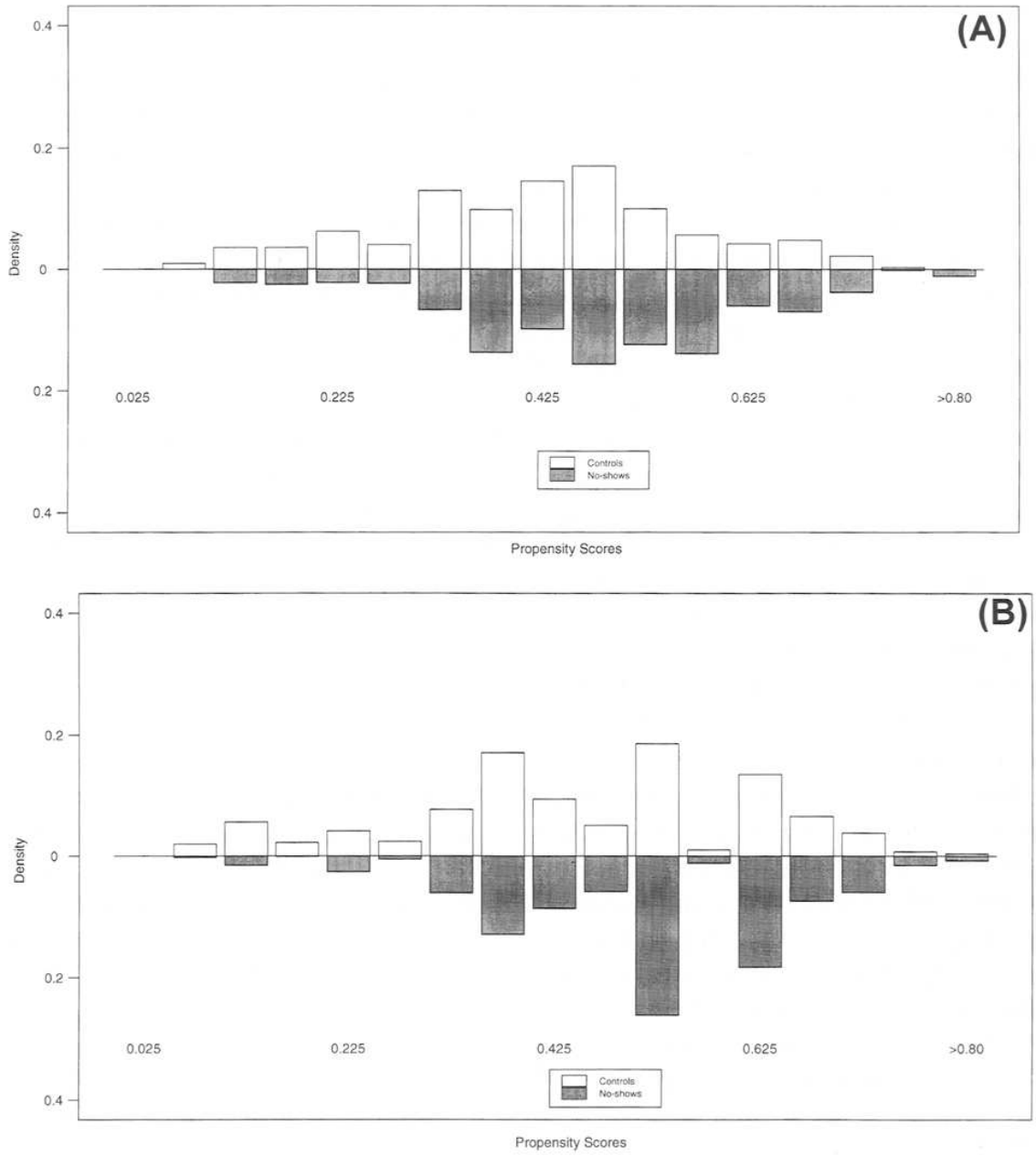

Fig. 11. Density of probability of program participation for adult male (A) and female (B) controls and no-shows in the National JTPA Study.

creates a substantial bias on the order of 30–40% of the estimated treatment effect.[85] Overall, comparing the fifth rows of Tables 8 and 9 suggests that putting participants and comparison group members in the same labor markets and giving them the same

[85] Friedlander and Robins (1995) report similar findings regarding the importance of drawing participants and non-participants from the same local labor markets.

questionnaire eliminates a substantial amount (around 50%) of selection bias, rigorously defined.

Those authors also report that a substantial bias results from using only those observations that fall into the common support of $P(X)$, S_P, for the control and comparison group samples to estimate the impact of treatment. Estimating the experimental treatment effect on the common support rather than on the full support of $P(X)$ among the controls increases the estimate by 50%. Put differently, the experimental impact estimate is higher for persons whose $P(X)$ lies in the common support.

The failure of the common support condition due to an absence of comparison group members comparable to participants in terms of X (or $P(X)$) is a major source of bias in conducting non-experimental evaluations. This motivates one of our major recommendations presented in Section 11 – that non-experimental comparison groups should be designed so that they have the same set of X or $P(X)$ values present among program participants.

An important advantage of an experimental control group in program evaluations is that randomization ensures that the support of treatment and control observed characteristics is the same, up to sampling variation. The results just discussed indicate that non-experimental methods may be able to mitigate major sources of selection bias that arise in the region of common support. Simple principles of using the same questionnaire, locating participants and comparison group members in the same labor markets, comparing comparable people and weighting comparison group members appropriately go a long way toward reducing the conventional measure of selection bias. However, because a significant source of the bias in non-experimental studies is the failure to find a comparison group for which the support of the observed characteristics largely overlaps that of the participants, such studies can only provide a partial description of the impact of treatment. Estimates obtained only over the region of common support may be a poor guide to the impact for all participants. We suspect that this source of bias is substantial for other programs besides the JTPA program where it has been studied.

Heckman et al. (1998b) use the estimated $B(P(X))$ functions to test among competing identifying assumptions for alternative evaluation estimators using the NJS data. Using a variety of X, they reach the following main conclusions:

(I) They reject the assumption: A_{M}: $B(P(X)) = 0$ for all X which justifies matching;

(II) They *do not* reject the assumption: A_{SS}: $B(X) = B(P(X))$ which says that the bias can be written as a function of $P(X)$ and which justifies the index sufficient classical sample selection model. However, since the support of $P(X)$ is limited, the method cannot recover $\mathrm{E}(Y_1 - Y_0 \mid X, D = 1)$ in their data because of the inability to identify the intercepts in the model. They decisively reject the normal sample selection model in their data.

(III) They do not reject the assumption: A_{DD}: $B_t(P(X)) - B_{t'}(P(X)) = 0$ for $t > k > t'$ which justifies the non-parametric difference-in-differences estimator introduced in Heckman et al. (1997a, 1998b). This estimator does not require the full support conditions required in the sample selection estimator although if they are not satisfied, the treatment effect defined only over a subset of the support of $P(X)$.

Finally, even though the assumptions justifying matching are rejected, matching, nonparametric difference-in-differences, and sample selection models do about equally well for the *average* of $E(Y_1 - Y_0 \mid X, D = 1)$ over the support where it can be defined although matching is somewhat inferior to the other two estimators. Their analysis demonstrates that over intervals where the bias balances out, fundamentally different estimators based on different identifying assumptions can identify the same parameter.

Heckman et al. (1998b) emphasize the importance of using semiparametric and nonparametric versions of all three estimators (matching, classical sample selection and difference-in-differences). When they use conventional parametric versions of these estimators, they estimate substantial biases.

The evidence presented in this subsection has major implications for the correct interpretation of LaLonde's (1986) influential examination of the effectiveness of non-experimental evaluation strategies for training programs. As noted in Table 6, LaLonde's non-experimental comparison groups were constructed from various non-comparable data sources. The comparison groups were located in different labor markets from program participants and had their earnings measured in different ways than the participants. His measure of selection bias, B, combines the three factors disentangled in the analyses of Heckman et al. (1996b, 1998b) just summarized.[86] In addition, like most of the studies summarized in Tables 5 and 6, he lacked information on recent preprogram labor force status dynamics which, as noted in Section 6.3, are an important predictor of participation in training. A major conclusion of the analysis of Heckman et al. (1998b) is that a substantial portion of the bias and sensitivity reported by LaLonde is due to his failure to compare comparable people and to weight them appropriately. Further, mismatch of labor markets and questionnaires are also likely important sources of the selection bias measured in LaLonde's study. Overall, the available evidence indicates that simple parametric econometric models applied to bad data do not eliminate selection bias. Instead, better data, including a rich array of X variables for use in constructing $P(X)$, and more appropriate comparison groups, go a long way toward eliminating the sensitivity problems raised in LaLonde's (1986) study.

8.3. A simulation study of the sensitivity of non-experimental methods

A theme of this chapter is that *every* estimator relies on identifying assumptions about the outcome and participation processes. When a particular estimator is applied to data where those assumptions fail to hold, bias results. This bias can be substantial. When different estimators are applied to the same data, the estimates they produce will vary because at most one set of underlying assumptions is consistent with the data. Only if there is no problem of selection bias would all estimators identify the same parameter.

[86] Some of LaLonde's (1986) measures of B are based on a linear regression model that "partials out" X in the sense that linear regression conditions on X. Heckman and Todd (1994) present the appropriate decomposition for this case. When estimated using the NJS controls and eligible non-participants, the same qualitative conclusions emerge about the importance of various components of bias.

To demonstrate these points, in this section we present a simulation study in which we examine the effects of alternative specifications of the processes that determine earnings and participation in training on the performance of various econometric estimators. Using earnings equations and participation rules that are consistent with the evidence from actual training programs, we apply a number of conventional econometric estimators to the simulated data. We vary aspects of the data generating process to see how the different components of the earnings and outcome equations affect the bias of the estimators discussed in Section 7.

8.3.1. A model of earnings and program participation

Building on the model of participation and earnings presented in Section 6.3, we specify a model to underlie our simulation study. Following the notation in Section 6.3, but augmenting it with "i" subscripts to distinguish individual variables from constants, we define the training period as period k, and let D_i be a dummy variable equal to 1 in periods $t > k$ if the individual receives training and 0 otherwise. Prior to the training period ($t < k$), D_i is identically equal to 0 for both future trainees and non-trainees. We further assume that individual i's earnings are determined by the following equation, where the error term combines an AR(1) (autoregressive of order one) process, as used, for example, in Eq. (7.1), with an individual-specific fixed effect, so that

$$Y_{it} = \beta + \alpha_i D_i + \theta_i + U_{it}, \tag{8.3}$$

where

$$U_{it} = \rho U_{i,t-1} + \varepsilon_{it}, \tag{8.4}$$

for all time periods t. $\mathrm{E}(\varepsilon_{it}) = 0$, where ε_{it} is independent and identically distributed over time and persons. The individual-specific fixed effect, θ_i, is drawn from a population distribution with mean zero. We assume that θ_i,ε_{it} and α_i are mutually independent. We assume random sampling so that all i-subscripted random variables are statistically independent of all i' subscripted variables, $i \neq i'$.

In this model,

$$Y_{it} = D_i Y_{1it} + (1 - D_i) Y_{0it}$$

and

$$Y_{1it} - Y_{0it} = \alpha_i.$$

This is a random coefficients model in which the effect of training, α_i, varies among individuals according to some population distribution. This specification of the outcome equation yields two parameters of interest: the mean effect of training in the population, $\mathrm{E}(\alpha_i)$, and the mean effect of training on those who actually receive training, $\mathrm{E}(\alpha_i \mid D_i = 1)$. The more standard common coefficient specification assumes that $\alpha_i = \alpha$ for all individuals, in which case the two parameters are equivalent.

Following the model of perfect certainty presented in Section 6.3, we assume that the

decision to participate in training depends on individuals' discounted lifetime gain from training, α_i/r, their opportunity costs or foregone earnings in period k, Y_{ik}, and their tuition costs or subsidy, c_i. More formally, we have

$$D_i = \begin{cases} 1 & \text{if } \alpha_i/r - Y_{ik} - c_i > 0 \text{ and } t > k, \\ 0 & \text{otherwise.} \end{cases} \tag{8.5}$$

As noted in Section 6.3, this model is consistent with Ashenfelter's dip in earnings among participants prior to participation. In some of the specifications analyzed below, we relax the perfect foresight assumption and consider the case where α is not known by the agent at the time program participation decisions are made. In Eq. (8.5), we introduce instruments as determinants of program costs and write $c_i = Z_{i'}\phi + V_i$, where Z_i is an observed characteristic that affects the cost of training and where V_i is a mean zero random disturbance. For simplicity, we assume that both Z_i and V_i are independent of all other variables and errors. We assume the trainees have zero earnings during the training period. Because D_i depends on foregone or "latent" earnings in period k, $E(D_i\theta_i)$ is non-zero and, in fact, is negative. Persons with higher values of θ_i have higher opportunity costs. As a result, OLS estimates of our parameters of interest are downward biased.

8.3.2. The data generating process

In our simulations, we set β=1000 and the treatment effect, α_i, is drawn from a normal distribution with a mean of 100 and standard deviation of σ_α. We explore the effects on the bias of different values of σ_α, including the common effect model where $\sigma_\alpha = 0$. The ε_{it} are randomly drawn from a normal distribution with mean zero and standard deviation σ_ε. We initialize the process by setting $U_{i,k-5} = \varepsilon_{i,k-5}$, where $k - 5$ is the initial period in the simulated data. We generate the θ_i from a normal distribution with mean zero and standard deviation σ_θ.

In the participation equation, the Z_i are randomly drawn from a $N(\mu_Z,\sigma_Z^2)$ distribution and the parameter ϕ is set equal to 1. The mean of the distribution of characteristics, μ_Z, is chosen so that, for each simulated sample, 10% of the population enters the program. Notice that because we draw the characteristics, Z_i, independently of both components of the outcome equation unobservable, θ_i and ε_{it}, Z_i is a valid instrument for the training variable D_i in the common coefficient model. When α_i varies among persons, and is acted on by agents, Z_i is not a valid instrument for the parameter $E(\alpha_i \mid D_i = 1)$ for the reasons given in Section 7.4.3. Only if the idiosyncratic component of α_i is not acted on in making participation decisions is IV a valid estimator of $E(\alpha_i \mid D_i = 1)$. We set the discount rate r to be 0.10. To complete the parameterization of the participation equation, we draw the disturbances, V_i, from a $N(0,\sigma_V^2)$ distribution.

Using this specification, in most of the runs we generate 100 samples each containing 1000 individuals. For each person in each sample, we generate 10 periods of earnings data. There are five pre-program periods, $k - 5$ to $k - 1$, one training period, k, and four post-program periods, $k + 1$ to $k + 4$ that we simulate. However, persons are assumed to live forever so the simple infinite horizon decision rule applies. Each sample consists of 100

participants and 900 non-participants. The "unmatched" comparison group used in the tables consist of all of the non-participants. Tables 10, 12 and 13 present estimates using unmatched comparison groups.

Matched samples are often formed prior to applying econometric estimators. As noted in Section 7.2, applying estimators to matched samples often invalidates the properties of an estimator that is appropriate in random or unmatched samples. In fact, matching is an estimator in its own right. The conventional practice of matching and then using an econometric estimator on the new samples created by matching is not in general justified. To illustrate the effects of this practice, our matched comparison group consists of non-participants matched to the participants using nearest neighbor matching with replacement. The sample sizes for the matched samples are much smaller. We have 100 treatment group members as before but at most 100 unique comparison group members in each matched sample – compared to the 900 members in the unmatched comparison group. Unless otherwise stated, the matching is on earnings two periods prior to participation, i.e., on $Y_{i,k-2}$. Similar matching or screening rules are widely used in the literature. Tables 11, 14 and 15 present estimates using the matched comparison groups, with the latter two tables examining the effects of alternative matching rules.

In the first column of Table 10, we present a set of "base case" estimates for a variety of models with a data generating process $\theta_i \sim N(0, 300)$, $\varepsilon_i \sim N(0, 450)$, $Z_i \sim N(0, 300)$, $\rho = 0.78$, and $\alpha_i = 100 + N(0, 300)$. These distributions are chosen to represent samples of the sort that appear in practice. The values for the standard deviations of θ_i and ε_i, as well as the value of ρ, are based on estimates reported in Ashenfelter and Card (1985). The value for the standard deviation of α_i is based on the estimate reported in Heckman et al. (1997c).[87] Column (1) considers the base case when $\mathrm{E}(\alpha_i \mid D_i = 1)$ is the parameter of interest while column (3) considers the base case when $\mathrm{E}(\alpha_i)$ is the parameter of interest. The expected value of the parameter of interest taken over all 100 simulated datasets appears in the column heading for each specification. In the base case, $\mathrm{E}(\alpha_i \mid D_i = 1) = 607.8$. Given that $\mathrm{E}(\alpha_i) = 100$, this indicates substantial selection into the program based on α_i. As previously discussed, the bias for $\mathrm{E}(\alpha_i)$ is the bias for $\mathrm{E}(\alpha_i \mid D_i = 1)$ plus $\mathrm{E}(\alpha_i - \mathrm{E}(\alpha_i) \mid D = 1) = \mathrm{E}(U_{1i} - U_{0i} \mid D = 1)$, the term incorporated into the definition of $\mathrm{E}(\alpha_i \mid D_i = 1)$.

In the remaining columns of Table 10, we vary one aspect of the data generating process at a time using the base case as a point of departure. Column (2) presents the common coefficient case, with $\alpha_i = \alpha = 100$ for all i. Column (4) presents the case of a random coefficient model where agents know $\mathrm{E}(\alpha_i)$ rather than α_i when making their program participation decisions. Thus there is ex ante homogeneity but ex post heterogeneity in realized outcomes so $\mathrm{E}(\alpha_i \mid D_i = 1) = \mathrm{E}(\alpha_i)$ and Z_i is a valid instrument for both parameters. Column (5) presents the base case with an increased variance of α_i. For each

[87] If α_i is a log concave random variable, then in a Roy model, the Heckman et al. (1997c) estimates of the variance of α_i are understated since they estimate $\mathrm{Var}(\alpha_i \mid D_i = 1)$ and not $\mathrm{Var}(\alpha_i)$.

specification in Table 10, Table 11 presents estimates using the matched comparison group sample.

In Section 7 we focused primarily on identification of the various parameters of interest under different assumptions about data generating processes. This focus follows much of the recent econometric literature on program evaluation, starting with Heckman and Robb (1985a, 1986a). In practice, securing identification is only a useful first step in determining a valid estimation strategy. The sampling variability of alternative estimators is an important consideration in picking an estimator. Different estimators converge to the true value at different rates. Table 12 presents some Monte Carlo evidence on the rates of convergence of the estimators we examine using different sample sizes.

Table 13 presents the results from simulations in which we reduce the standard deviations of the random variables determining outcomes and participation one at a time, holding the overall variances fixed, in order to explore the effect of the size of relative components of variance on the bias.

We stress that the Monte Carlo analysis reported in this chapter is illustrative rather than definitive. Heckman and Smith (1998e) present a much more comprehensive Monte Carlo study which examines the bias and small sample variability of the main non-experimental estimators presented in Section 7. Our work draws from their findings.

8.3.3. The estimators we examine

The assumptions required to justify each estimator are discussed in Sections 3, 4 and 7. Here we briefly discuss how each estimator was implemented in our simulation study. The estimators selected are those most commonly used in the literature. The entries in the tables indicate the mean and, in parentheses, the standard deviation of the estimates obtained from the 100 simulated samples. For the IV and Heckman (1979) estimators we present additional statistics of interest. Unless otherwise noted, the estimates presented in these tables reflect impacts on Y_{k+4}.

The first estimator in each table is the cross-section estimator applied to post-program earnings. Because we do not include any observables in the earnings equation, the cross-section estimator is the coefficient on D_i in a regression of $Y_{i,k+4}$ on D_i and is equivalent to the difference between the mean of participant and non-participant earnings. The cross-section estimator is biased downward when $\text{Var}(\theta_i) > 0$ or $\rho > 0$. When $\text{Var}(\theta_i) = \rho = 0$, the cross-section estimator identifies $\text{E}(\alpha_i \mid D_i = 1)$ when applied to post-program earnings.

The second, third and fourth rows in each table present three alternative versions of the difference-in-differences estimator based on the averages (over $D = 1$ and $D = 0$) of the comparisons:

$$Y_{it} - Y_{i,t'} = \alpha_i D_i + (U_{it} - U_{i,t'}), \tag{8.6}$$

where $t' < k < t$. In all three rows, $t = k + 3$ is the "after" period. The three rows differ based on the value chosen for the "before" period, to show the effect of differencing relative to different points in the sequence along Ashenfelter's dip and also to illustrate the symmetric differencing estimator. In the second row, the before period is $k - 1$, in the

Table 10
Bias in non-experimental estimates of the impact of training (unmatched comparison group samples)[a]

Estimator[b]	Base case[c]; parameter of of interest: $E(\alpha \mid D = 1) = 615.7$ (1)	Base case with common coefficient; parameter of interest: $E(\alpha \mid D = 1) = 100.0$ (2)	Base case; parameter of interest: $E(\alpha) = 100.0$ (3)	Base case where agent knows $E(\alpha)$, not α_i; parameter of interest $E(\alpha \mid D = 1) = 98.4$ (4)	Base case with increased variance of α; parameter of interest $E(\alpha \mid D = 1) = 971.4$ (5)
Cross-section					
Mean	−98.5	−494.7	417.2	−494.7	−60.5
SD	(61.9)	(47.9)	(66.6)	(47.9)	(59.2)
Diff-in-diff (−1,3)					
Mean	34.0	155.6	549.7	155.6	18.7
SD	(58.0)	(54.4)	(61.4)	(54.4)	(56.1)
Diff-in-diff (−3,3)					
Mean	−9.8	−55.2	505.9	−55.2	−7.9
SD	(60.7)	(52.3)	(64.5)	(52.3)	(58.5)
Diff-in-diff (−5,3)					
Mean	−40.2	−211.7	475.5	−211.7	−26.9
SD	(63.1)	(51.6)	(67.1)	(51.6)	(57.9)
AR(1) regression					
Mean	−15.7	−187.1	500.0	−174.0	5.0
SD	(203.5)	(203.3)	(202.9)	(210.3)	(197.4)
IV estimator					
Mean	342.7	−15.3	858.4	−15.1	−102.8
Median	−146.0	−5.8	369.1	−5.7	−247.7
SD	(4829.0)	(231.5)	(4829.1)	(227.7)	(5131.2)
Corr(Z,D)	0.0559	0.2755	0.0559	0.2755	0.0355

Ashenfelter (1979)					
Mean	75.9	166.4	591.6	166.1	86.5
SD	(59.3)	(53.3)	(62.7)	(53.7)	(58.8)
Heckman (1979)					
Mean	214.5	53.2	1139.8	54.8	350.2
Median	−90.9	53.5	204.8	61.2	−178.3
SD	(3595.3)	(221.4)	(12811.0)	(269.1)	(6618.0)
Kitchen sink					
Mean	−20.0	−160.7	495.7	−160.9	−11.9
SD	(54.9)	(54.8)	(58.6)	(55.0)	(52.4)

[a] Estimates are based on 100 simulated samples of 1000 observations each. The "mean" row presents the mean of the estimates from the 100 samples while the "SD" row presents the standard deviation of the estimates from the 100 samples. The "Corr(Z,D)" row for the IV estimates gives the average correlation between the participation indicator, D, and the instrument, Z.

[b] The cross-section estimator is the simple difference between participant and non-participant earnings in period $k+4$. The difference-in-differences estimates are based on the periods indicated, so that $(-1,3)$ is the difference between the change in participant earnings from period $k-1$ to period $k+3$ and the change in non-participant earnings over the same interval. The difference-in-differences $(-3,3)$ estimator is symmetric. The AR(1) estimates are based on a regression of Y_{k+4} on Y_{k+3} and D, with the estimate consisting of the coefficient estimate on D divided by $(1-\rho)$, where ρ is estimated by the coefficient on Y_{k+3}. The IV estimates use Z as an instrument for a regression of Y_{k+4} on D. The Ashenfelter (1979) estimator is described in Section 8.3.3. The dependent variable for this estimator is $Y_{i+4}-Y_k$. The Heckman (1979) estimator is a special case of the class of control function estimators presented in Section 7.4.2. In columns (1) and (5) the estimate is calculated as shown in Section 7.4.2. In columns (2), (3) and (4) the estimate is the coefficient on D when the estimated control functions are included. The dependent variable for the Heckman (1979) estimator is Y_{k+4}. The kitchen sink estimates are based on a regression of Y_{k+4} on Y_{k-1}, Y_{k-2} and Z.

[c] The base case has $\theta \sim N(0,300)$, $\varepsilon \sim N(0,280)$, $Z \sim N(0,300)$, $V \sim N(0,200)$, $\rho = 0.78$ and $\alpha = 100 + N(0,300)$. This case is based on estimates of the size of the permanent and transitory components of earnings from Ashenfelter and Card (1985) and of the variance in the impacts of training from Heckman et al. (1997c). In column (2), $\alpha = 100$. In column (5), $\alpha = 100 + N(0,500)$. In the base case in columns (1) and (3), the fractions of $\text{Var}(Y_{k+4} \mid D = 1)$ accounted for by α and θ are 0.0564 and 0.2670, respectively. In column (2), they are 0.0000 and 0.3132, respectively. In column (4), they are 0.2787 and 0.2246, respectively. In column (5), they are 0.1273 and 0.2467, respectively.

Table 11
Bias in non-experimental estimates of the impact of training (matched comparison group samples)[a]

Estimator[b]	Base case[c]; parameter of interest $E(\alpha\|D=1) = 615.7$ (1)	Base case with common coefficient; parameter of interest $E(\alpha\|D=1) = 100.0$ (2)	Base case; parameter of interest $E(\alpha) = 100.0$ (3)	Base case where agent knows $E(\alpha)$, not α_i; parameter of interest $E(\alpha\|D=1) = 98.4$ (4)	Base case with increased variance of α; parameter of interest $E(\alpha\|D=1) = 971.4$ (5)
Cross-section[c]					
Mean	−42.9	−233.0	472.8	−233.0	−27.4
SD	(80.3)	(70.4)	(81.8)	(70.4)	(74.4)
Diff-in-diff (−1,3)					
Mean	−5.8	−36.8	509.9	−36.8	−5.9
SD	(77.5)	(77.1)	(78.8)	(77.1)	(77.5)
Diff-in-diff (−3,3)					
Mean	−43.3	−243.2	472.4	−243.2	−28.1
SD	(82.5)	(70.9)	(84.4)	(70.9)	(82.2)
Diff-in-diff (−5,3)					
Mean	−33.9	−202.0	481.8	−202.0	−23.2
SD	(80.3)	(76.1)	(83.0)	(76.1)	(77.9)
AR(1) regression					
Mean	−6.9	−69.1	508.8	−13.7	5.4
SD	(333.7)	(479.3)	(333.7)	(588.9)	(323.3)
IV estimator					
Mean	−305.0	27.3	210.7	28.2	−427.5
Median	−41.9	18.1	474.1	10.1	36.2
SD	(4001.6)	(175.8)	(4000.8)	(191.0)	(3643.9)
Corr(Z,D)	0.0923	0.5035	0.0923	0.5035	0.0604
Ashenfelter (1979)					
Mean	81.5	209.3	597.2	208.6	90.0
SD	(83.0)	(85.0)	(84.7)	(84.7)	(79.3)

Heckman (1979)					
Mean	−22060.2	24.1	−6915.5	22.9	221.7
Median	−57.6	13.9	469.0	7.3	−41.5
SD	(222223.4)	(177.2)	(71920.5)	(198.5)	(3739.7)
Kitchen sink					
Mean	−22.8	−194.7	492.9	−194.5	−13.9
SD	(80.2)	(91.5)	(81.1)	(94.9)	(75.4)

[a] Estimates are based on 100 simulated samples of 1000 observations each. The "mean" row presents the mean of the estimates from the 100 samples while the "SD" row presents the standard deviation of the estimates from the 100 samples. The "Corr(Z,D)" row for the IV estimates gives the average correlation between the participation indicator, D, and the instrument, Z. Matching consists of nearest neighbor matching on Y_{k-2} with replacement. The average number of unique observations in a matched sample is 92.1 in columns (1) and (3), 79.8 in columns (2) and (4) and 92.3 in column (5).

[b] The cross-section estimator is the simple difference between participant and non-participant earnings in period $k+4$. The difference-in-differences estimates are based on the periods indicated, so that $(-1,3)$ is the difference between the change in participant earnings from period $k-1$ to period $k+3$ and the change in non-participant earnings over the same interval. The difference-in-differences $(-3,3)$ estimator is symmetric. The AR(1) estimates are based on a regression of Y_{k+4} on Y_{k+3} and D, with the estimate consisting of the coefficient estimate on D divided by $(1-\rho)$, where ρ is estimated by the coefficient on Y_{k+3}. The IV estimates use Z as an instrument for a regression of Y_{k+4} on D. The Ashenfelter (1979) estimator is described in Section 8.3.3. The dependent variable for this estimator is $Y_{k+4}-Y_k$. The Heckman (1979) estimator is a special case of the class of control function estimators presented in Section 7.4.2. In columns (1) and (5) the estimate is calculated as shown in Section 7.4.2. In columns (2), (3) and (4) the estimate is the coefficient on D when the estimated control functions are included. The dependent variable for the Heckman (1979) estimator is Y_{k+4}. The kitchen sink estimates are based on a regression of Y_{k+4} on Y_{k-1}, Y_{k-2} and Z.

[c] The base case has $\theta \sim N(0,300)$, $\varepsilon \sim N(0,280)$, $Z \sim N(0,300)$, $V \sim N(0,200)$, $\rho = 0.78$ and $\alpha = 100 + N(0,300)$. This case is based on estimates of the size of the permanent and transitory components of earnings from Ashenfelter and Card (1985) and of the variance in the impacts of training from Heckman et al. (1997c). In column (2), $\alpha = 100$. In column (5), $\alpha = 100 + N(0,500)$. In the base case in columns (1) and (3), the fractions of $\text{Var}(Y_{k+4} \mid D = 1)$ accounted for by α and θ are 0.0564 and 0.2670, respectively. In column (2), they are 0.0000 and 0.3132, respectively. In column (4), they are 0.2787 and 0.2246, respectively. In column (5), they are 0.1273 and 0.2467, respectively.

Table 12
Bias in non-experimental estimates of the impact of training (unmatched comparison group samples)[a]

Estimator[b]	Base case[c] with sample size = 2500; parameter of interest $E(\alpha \mid D = 1) = 615.1$ (1)	Base case with sample size = 5000; parameter of interest $E(\alpha \mid D = 1) = 614.6$ (2)	Base case with sample size = 10000; parameter of interest $E(\alpha \mid D = 1) = 614.6$ (3)
Cross-section			
Mean	−102.0	−106.0	−103.4
SD	(35.4)	(24.7)	(18.8)
Diff-in-diff (−1,3)			
Mean	34.5	34.5	34.6
SD	(35.1)	(25.3)	(17.0)
Diff-in-diff (−3,3)			
Mean	−13.0	−13.7	−10.5
SD	(38.5)	(26.7)	(18.4)
Diff-in-diff (−5,3)			
Mean	−48.4	−47.3	−44.4
SD	(33.8)	(22.0)	(16.8)
AR(1) regression			
Mean	−4.4	−20.9	−26.5
SD	(137.1)	(85.8)	(65.2)
IV estimator			
Mean	−191.2	−201.1	−173.4
Median	−118.7	−205.1	−170.8
SD	(837.4)	(470.9)	(322.2)
Corr(Z,D)	0.0536	0.0577	0.0577
Ashenfelter (1979)			
Mean	73.0	71.0	70.3
SD	(30.1)	(21.8)	(15.5)
Heckman (1979)			
Mean	2.0	−60.5	−38.8
Median	24.8	−27.6	−20.6
SD	(931.8)	(584.1)	(380.8)
Kitchen sink			
Mean	−20.3	−23.2	−22.6
SD	(30.2)	(20.9)	(15.9)

[a] Estimates are based on 100 simulated samples of the indicated size. The "mean" row presents the mean of the estimates from the 100 samples while the "SD" row presents the standard deviation of the estimates from the 100 samples. The "Corr(Z,D)" row for the IV estimates gives the average correlation between the participation indicator, D, and the instrument, Z.

[b] The "base case" has $\theta \sim N(0, 300)$, $\varepsilon \sim N(0, 280)$, $Z \sim N(0, 300)$, $V \sim N(0, 200)$, $\rho = 0.0$, $\alpha = 100 + N(0, 300)$. Estimates for the base case with samples of size 1000 appear in Table 10. This case is based on estimates of the size of the permanent and transitory components of earnings from Ashenfelter and Card (1985) and of the variance in the impacts of training from Heckman et al. (1997c). In column (1), the fractions of $\text{Var}(Y_{k+4} \mid D = 1)$ accounted for by α and θ are 0.0556 and 0.2678, respectively. In column (2), the fractions are 0.0561 and 0.2692, respectively. In column (3), the fractions are 0.0558 and 0.2695, respectively.

third row it is $k-3$, which is the symmetric case, and in the fourth row it is $k-5$. The general difference-in-differences estimator will only be consistent for $\mathrm{E}(\alpha_i \mid D_i = 1)$ when $\rho = 0$.

The fifth estimator is the simple autoregressive estimator discussed in Section 7.6:

$$\begin{aligned} Y_{it} &= \rho Y_{i,t-1} + (1-\rho)\beta + (1-\rho)\mathrm{E}(\alpha_i \mid D_i = 1)D_i + (1-\rho)\theta_i \\ &\quad + (1-\rho)D_i[\alpha_i - \mathrm{E}(\alpha_i \mid D_i = 1)] + \varepsilon_{it} \\ &= \rho Y_{i,t-1} + \beta^* + \alpha^* D_i + \theta^* + (1-\rho)D_i[\alpha_i - \mathrm{E}(\alpha_i \mid D_i = 1)] + \varepsilon_{it}, \end{aligned} \tag{8.7}$$

where $\beta^* = (1-\rho)\beta$ and $\alpha^* = (1-\rho)\mathrm{E}(\alpha_i \mid D_i = 1)$. We define $\hat{\alpha}_{\mathrm{AR}} = \hat{\alpha}^*/(1-\hat{\rho})$ where $\hat{\alpha}^*$and $\hat{\rho}$ are the OLS estimators of $(1-\rho)\mathrm{E}(\alpha_i \mid D_i = 1)$ and ρ, respectively. The autoregressive estimator identifies $\mathrm{E}(\alpha_i \mid D_i = 1)$ only when $\mathrm{Var}(\theta_i) = 0$ and $\sigma_\alpha = 0$, i.e., only when there are no fixed effects in the outcome equation and there is no heterogeneity in the impact of treatment.

The sixth estimator we consider is an instrumental variables (IV) estimator. We calculate the IV estimates using Z_i, the observable variable in the participation equation, as an instrument for the training indicator variable, D_i, in earnings Eq. (8.3). For post-program earnings, the IV estimator will consistently estimate $\mathrm{E}(\alpha_i \mid D_i = 1)$ if $\mathrm{E}(Z_i D_i) \neq 0$, $\mathrm{E}(Z_i \theta_i) = 0$, and $\mathrm{E}(Z_i \varepsilon_i) = 0$ for all t and if α_i is the same for everyone or, when it is heterogeneous, if agents do not choose to participate in the program based upon it. If agents select into the program based on α_i, then IV is inconsistent for $\mathrm{E}(\alpha_i \mid D_i = 1)$. However, in this case IV estimates the LATE associated with the instrument Z_i because our model satisfies the monotonicity and independence conditions (7.IA.1) and (7.IA.2) of Imbens and Angrist (1994). Accordingly, provided that the estimates converge adequately to large sample values, our Monte Carlo analysis reveals how much the LATE differs from treatment on the treated assuming that the estimator is consistent.

The seventh estimator we consider is Ashenfelter's (1979) difference-in-differences autoregressive estimator. His estimator may be written as

$$Y_{it} - Y_{ik} = (\rho^{t+1} - \rho)Y_{i,k-1} + \beta^{**} + \alpha^{**} D_{it} + \theta^{**} + U^{**}. \tag{8.8}$$

From knowledge of ρ, Ashenfelter proposes to estimate the parameter of interest, $\mathrm{E}(\alpha_i \mid D_i = 1)$, using

$$\hat{\alpha}_{ASH} = \left[\sum_{j=1}^{t-1} \rho^j (1-\rho)\right]^{-1} \hat{\alpha}^{**}, \tag{8.9}$$

where $\hat{\alpha}^{**}$ is the OLS estimate of α^{**} in Eq. (8.8). When $\rho \neq 0$, this estimator is biased and inconsistent for α in the common coefficient model and for both $\mathrm{E}(\alpha_i)$ and $\mathrm{E}(\alpha_i \mid D_i = 1)$ in the random coefficient model (Heckman, 1978).

The eighth estimator shown in each table is the Heckman (1979) two-step estimator based on the assumption that the unobservables in the outcome and participation equations

Table 13
Bias in non-experimental estimates of the impact of training (unmatched comparison group samples)[a]

Estimator[b]	Base case[c] with reduced variance of θ; parameter of interest $E(\alpha \mid D = 1) = 616.2$ (1)	Base case with reduced variance of ε; parameter of interest $E(\alpha \mid D = 1) = 615.1$ (2)	Base case with reduced variance of z; parameter of interest $E(\alpha \mid D = 1) = 615.8$ (3)	Base case with reduced variance of V; parameter of interest $E(\alpha \mid D = 1) = 616.1$ (4)	Base case with $\rho = 0$ and no fixed effect, θ; parameter of interest $E(\alpha \mid D = 1) = 615.3$ (5)
Cross-section					
Mean	−59.9	−185.4	−96.7	−97.9	7.5
SD	(58.0)	(50.7)	(55.6)	(61.0)	(57.8)
Diff-in-diff (−1,3)					
Mean	46.7	0.4	33.7	32.1	−0.4
SD	(69.2)	(6.2)	(53.3)	(58.7)	(89.2)
Diff-in-diff (−3,3)					
Mean	−15.8	−0.3	−11.9	−9.8	−3.1
SD	(71.7)	(6.2)	(59.9)	(59.9)	(78.0)
Diff-in-diff (−5,3)					
Mean	−61.6	−0.3	−41.9	−40.1	2.4
SD	(72.3)	(5.9)	(58.9)	(60.7)	(91.2)
AR(1) regression					
Mean	25.1	−445.6	−0.7	−20.2	7.4
SD	(157.1)	(4510.8)	(192.2)	(193.7)	(57.7)
IV estimator					
Mean	−605.0	−193.2	−2054.7	119.2	−207.6
Median	−239.1	−323.7	−230.0	−164.7	−147.2
SD	(3630.6)	(2020.2)	(17443.0)	(2678.8)	(1962.8)
Corr(Z,D)	0.0552	0.0576	0.0063	0.0677	0.0583

Ashenfelter (1979)					
Mean	96.8	43.4	76.1	73.5	247.7
SD	(64.7)	(6.6)	(55.5)	(56.6)	(83.6)
Heckman (1979)					
Mean	−887.5	50.0	−2515.9	290.5	−446.7
Median	−129.8	−56.2	−98.5	−21.7	−109.3
SD	(5429.7)	(3310.5)	(19052.7)	(3162.0)	(3546.4)
Kitchen sink					
Mean	−19.8	2.0	−18.9	−21.2	8.7
SD	(59.2)	(6.2)	(48.2)	(53.8)	(58.3)

[a] Estimates are based on 100 simulated samples of 1000 observations each. The "mean" row presents the mean of the estimates from the 100 samples while the "SD" row presents the standard deviation of the estimates from the 100 samples. The "Corr(Z,D)" row for the IV estimates gives the average correlation between the participation indicator, D, and the instrument, Z.

[b] The cross-section estimator is the simple difference between participant and non-participant earnings in period $k+4$. The difference-in-differences estimates are based on the periods indicated, so that $(-1,3)$ is the difference between the change in participant earnings from period $k-1$ to period $k+3$ and the change in non-participant earnings over the same interval. The difference-in-differences $(-3,3)$ estimator is symmetric. The AR(1) estimates are based on a regression of Y_{k+4} on Y_{k+3} and D, with the estimate consisting of the coefficient estimate on D divided by $(1-\rho)$, where ρ is estimated by the coefficient on Y_{k+3}. The IV estimates use Z as an instrument for a regression of Y_{k-4} on D. The Ashenfelter (1979) estimator is described in Section 8.3.3. The dependent variable for this estimator is $Y_{k-4}-Y_k$. The Heckman (1979) estimator is a special case of the class of control function estimators presented in Section 7.4.2. The estimates in all five columns are calculated as shown in Section 7.4.2. The dependent variable for the Heckman (1979) estimator is Y_{k-4}. The kitchen sink estimates are based on a regression of Y_{k+4} on Y_{k-1}, Y_{k-2} and Z.

[c] The base case has $\theta \sim N(0,300)$, $\varepsilon \sim N(0,280)$, $Z \sim N(0,300)$, $V \sim N(0,200)$, $\rho = 0.78$ and $\alpha = 100 + N(0,300)$. In column (1), $\theta \sim N(0,30)$ and $\varepsilon \sim N(0,337)$, in column (2), $\theta \sim N(0,538)$ and $\varepsilon \sim N(0,30)$, in column (3), $Z \sim N(0,30)$ and $V \sim N(0,359)$, in column (4), $Z \sim N(0,359)$ and $V \sim N(0,30)$ and in column (5), $\rho = 0$ and there is no fixed effect, θ. The base case is based on estimates of the size of the permanent and transitory components of earnings from Ashenfelter and Card (1985) and of the variance in the impacts of training from Heckman et al. (1997c). In column (1), the fractions of $\text{Var}(Y_{k+4} \mid D = 1)$ accounted for by α and θ are 0.0578 and 0.0028, respectively. In column (2), the fractions are 0.0540 and 0.8015, respectively. In column (3), the fractions are 0.0563 and 0.2683, respectively. In column (4), the fractions are 0.0560 and 0.2670, respectively. In column (5), the fractions are 0.0605 and 0.0000, respectively.

are jointly normally distributed. The general control function estimator, of which the Heckman (1979) estimator is a special case, is given by Eq. (7.15). Under its identifying assumptions, this estimator consistently estimates both $E(\alpha_i)$ and $E(\alpha_i \mid D_i = 1)$ using the procedures described in Section 7.4.2. Because we assume normal errors, our analysis is favorable to this estimator.

The final row in each table presents what we call the "kitchen sink" estimator. This estimator approximates the common practice of conditioning on whatever variables are available in an earnings equation that also includes an indicator for receipt of training. The Barnow et al. (1980) estimator is a version of the kitchen sink estimator (see the discussion in Section 7.4.1). We implement this estimator by regressing earnings in each post-program period on D_i, X_i, $Y_{i,k-2}$, and $Y_{i,k-1}$. This estimator is inconsistent for all of the specifications we consider except those with $\rho = 0$.

8.3.4. Results from the simulations

All of the specifications we consider depart from the base case presented in the first columns of Tables 10 (for the unmatched comparison group) and 11 (for the matched comparison group). In the base case, the cross-section estimator is biased downward in both the unmatched and matched samples because persons with low fixed effects, θ_i, are differentially more likely to participate in the program, which implies that participants have lower average earnings without training than do comparison group members. This bias is accentuated by selection into the program based on low values of ε_{it} in the enrollment period k, which persist over time due to the high value of ρ. Using a matched comparison group cuts the mean bias for the cross-section estimator roughly in half. The difference-in-differences estimator takes care of the selection on θ_i when that is the only source of bias, but not the selection bias due to the persistence in the transitory shocks. It has a lower mean bias than the cross-section estimator but is still inconsistent. Use of a matched comparison group has mixed effects on the bias in the difference-in-differences estimator.

The AR(1) estimator is consistent if $\sigma_\theta = 0$ and $\sigma_\alpha = 0$. In the base case, even though $\sigma_\theta > 0$ and $\sigma_\alpha > 0$, the estimator performs relatively well, with the lowest mean bias for the unmatched comparison group and one of the lowest with the matched comparison group. This is an artifact of the specific parameter values chosen for the base case model. For this model, several sources of bias just happen to cancel out, resulting in a lower overall bias (Heckman and Smith, 1998e). In particular, $Y_{i,t-1}$ is positively correlated with both θ_i and $D_i(\alpha_i - E(\alpha_i \mid D_i = 1))$ in the outcome equation error, and D_i is negatively correlated with θ_i. Heckman and Smith (1998e) present a comprehensive analysis of this case and demonstrate that perturbations in the base case specifications produce large biases in the AR(1) model.

The IV estimator is inconsistent for treatment on the treated in the base case because Z_i is correlated with the error term conditional on D_i as shown in Section 7.4.3. This inconsistency is reflected in large and highly variable biases with both the matched and unmatched comparison groups. However, IV consistently estimates the LATE associated

with Z_i. Using the median value, the LATE parameter is 25% lower than the treatment on the treated parameter. The Ashenfelter (1979) and kitchen sink estimators are inconsistent as well, but have relatively small estimated biases. In both cases, conditioning on lagged earnings appears to provide an imperfect but still helpful control for the effects of selection in both the matched and unmatched samples.

Column (2) of Tables 10 and 11 presents the bias for the common coefficient case in which $\sigma_\alpha = 0$ for the unmatched and matched comparison groups, respectively. Switching from the variable coefficient case to the common coefficient case has two important effects. First, in the common coefficient case selection into the program depends solely on θ_i and U_{it}. In contrast, in the random coefficient base case, persons with values of θ_i or U_{it} near zero, or even positive, will nonetheless select into training if they have a large enough value of α_i. Figs. 12 and 13 show that in the common coefficient case, the distribution of θ_i for trainees differs much more sharply from that for non-trainees than in the random coefficient base case. A further consequence of eliminating α_i as a determinant of program participation is that Ashenfelter's dip becomes much deeper in the common coefficient case, reflecting the stronger sorting on θ_i and U_i. Figs. 14 and 15 illustrate this difference.

In the random coefficient base case, selection into the program based on α_i acts like randomization for the parameter $\mathrm{E}(\alpha_i \mid D_i = 1)$ because α_i is uncorrelated with all of the components of post-program error. The more D_i is driven by variation in α_i, the more exogenous it is and the smaller the bias. To see this, compare the cross-section estimator in columns (1) and (2). Without the benefit of the pseudo-randomization induced by selection into the program based on α_i, the bias in the common coefficient case, which has the same variances of θ_i and U_{it} as the base case, is much greater. The stronger selection on θ_i and U_i in the common coefficient case and the deeper dip it induces substantially increase the mean bias in all cases except the IV and Heckman (1979) estimators for both the unmatched and matched comparison groups.

The second important effect of switching to the common coefficient model is to dramatically improve the performance of the IV and Heckman (1979) estimators. (This also shows up in their excellent performance in column (4) for the model in which $\mathrm{E}(\alpha_i \mid D_i = 1) = \mathrm{E}(\alpha_i)$ and there is no selection on α_i.) As discussed in Section 7.4, in the common coefficient case, Z_i is a valid instrument (or exclusion restriction) because it is no longer correlated with the error term conditional on D_i. As a result, both the mean bias and the variability in the estimates across samples fall.

Column (3) of Tables 10 and 11 shows the mean bias when $\mathrm{E}(\alpha_i)$ rather than $\mathrm{E}(\alpha_i \mid D_i = 1)$ is the parameter of interest. As indicated by the values in the column headings these parameters differ greatly in our base case model because there is strong selection into the program of persons with high values of α_i. As a result, estimators which estimate $\mathrm{E}(\alpha_i \mid D_i = 1)$ with low bias provide highly biased estimates of $\mathrm{E}(\alpha_i)$. For this parameter, the dependence of D_i on α_i is a source of bias rather than a solution to the bias problem as it is when the parameter of interest is $\mathrm{E}(\alpha_i \mid D_i = 1)$.

Column (4) of Tables 10 and 11 shows the bias for the case where α_i varies across

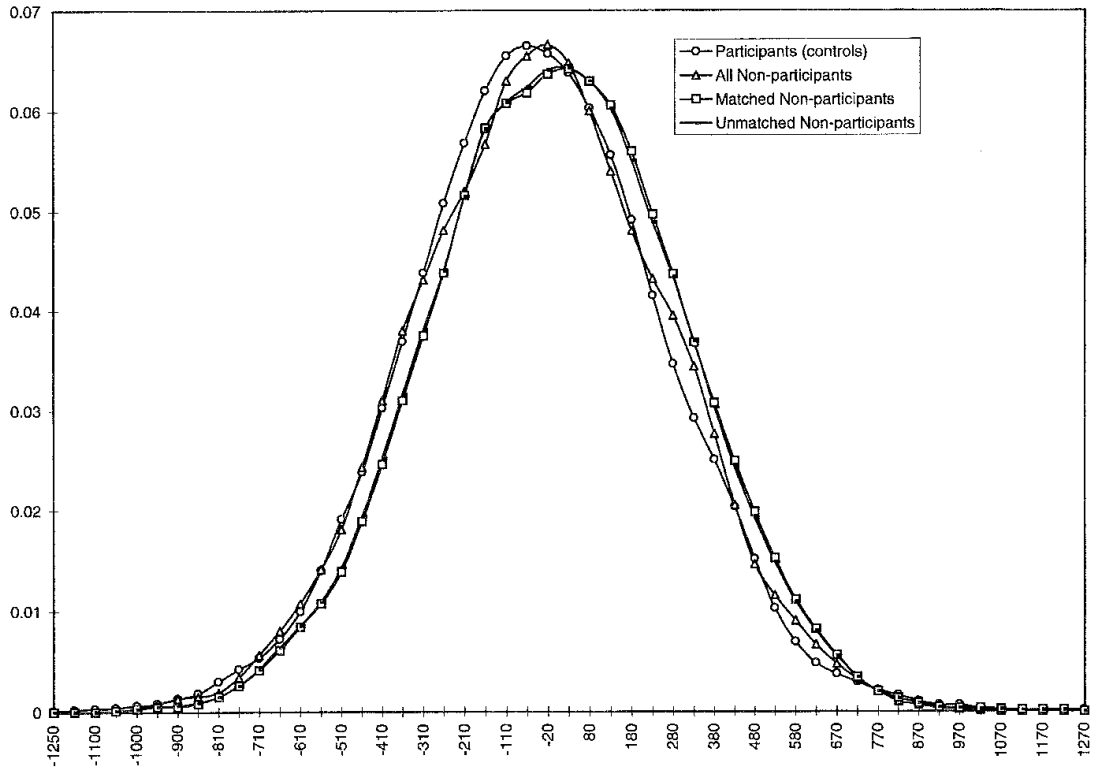

Fig. 12. Distribution of θ in base case with random coefficient.

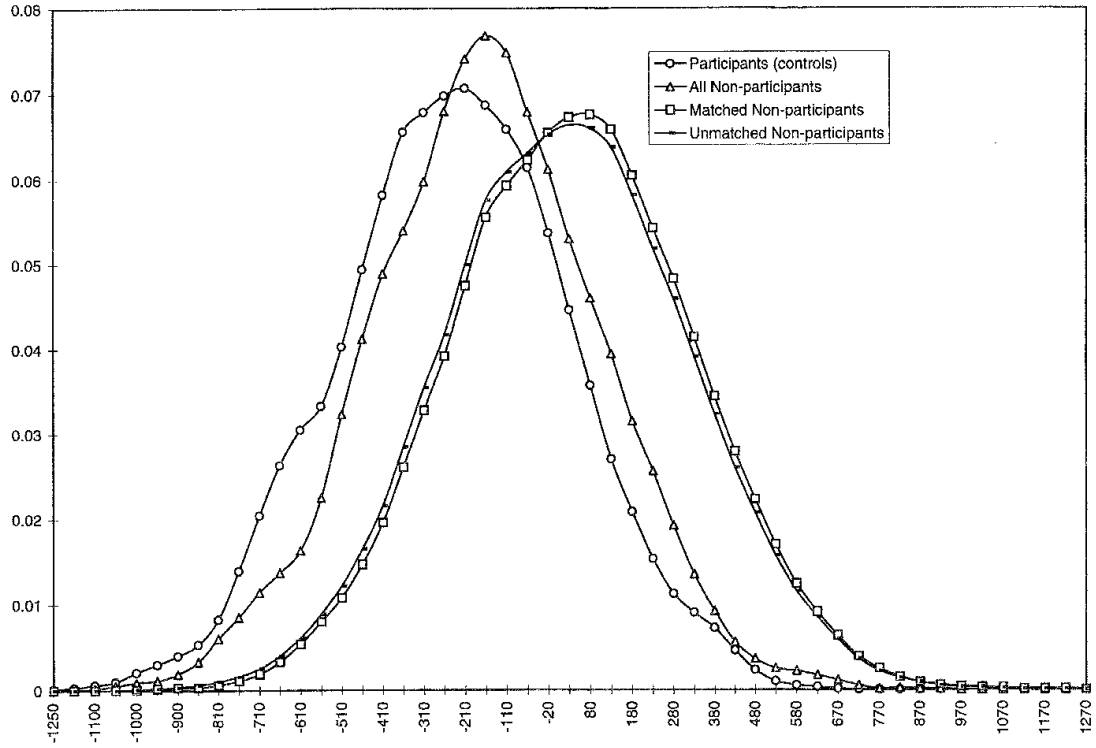

Fig. 13. Distribution of θ in base case with common coefficient.

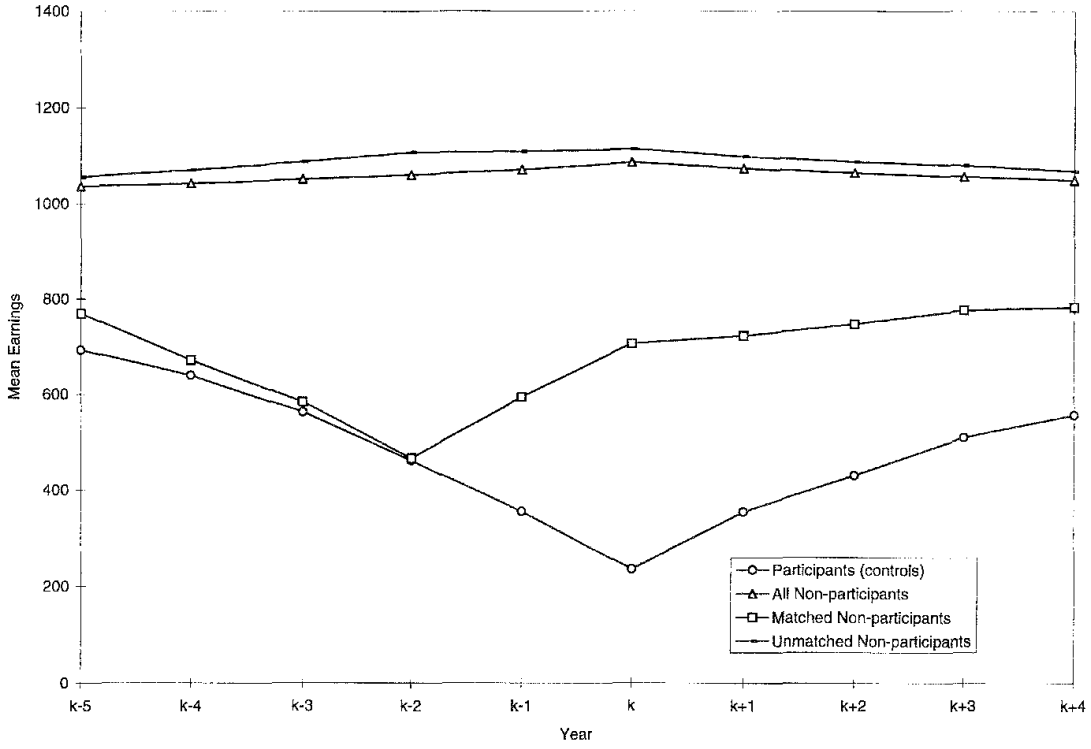

Fig. 14. Mean earnings in base case with common coefficient and matching on Y_{k-2}.

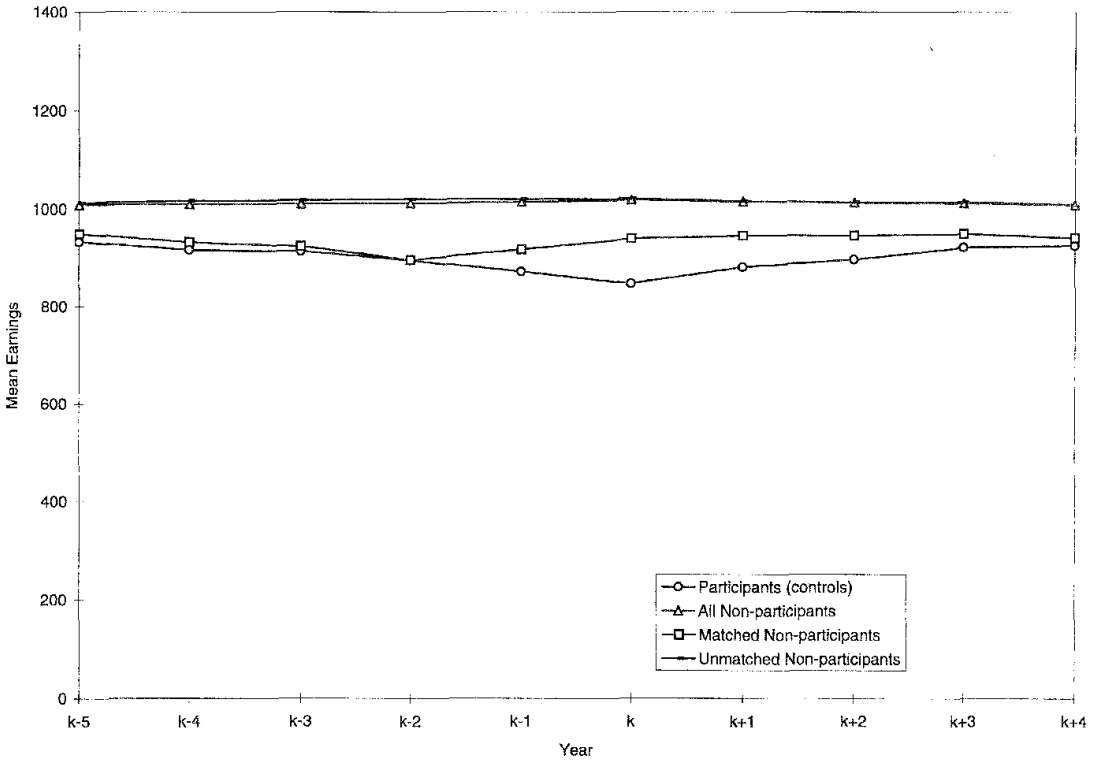

Fig. 15. Mean earnings in base case with random coefficient and matching on Y_{k-2}.

persons but selection decisions are based only on the expected value $E(\alpha_i)$. That is, in this case, potential trainees are assumed not to know the idiosyncratic component of their gain (or loss) from training. The estimated biases for the all of the estimators other than the AR(1) are essentially the same as in the common coefficient case, because in both cases variation in D_i is not driven by variation in α_i. For the AR(1) estimator, the additional error component in the earnings equation adversely affects the performance of the estimator.

The final column of Tables 10 and 11 presents the estimates when the base case is altered by increasing the variance of α_i while holding the variances of θ_i and U_{it} fixed. This essentially randomizes D_i against the error term $\theta_i + U_{it} + D_i(\alpha_i - E(\alpha_i \mid D_i = 1))$. Ceteris paribus, increasing the heterogeneity of the impact of treatment improves the performance of all of the estimators we examine except for the IV and Heckman (1979) estimators so long as $E(\alpha_i \mid D_i = 1)$ is the parameter of interest and agents act on α_i in making program participation decisions.

The evidence presented in Table 10 is based on Monte Carlo using simulated samples of 1000 observations. Though small, evaluators often use samples of this size in practice. In order to study how much of the bias reported in Table 10 results from failure to converge to the true bias values, and in order to gauge the reliability of large sample theory when applied to samples of the sizes used in practice, we present bias estimates from simulated samples of size 2500, 5000 and 10,000 in Table 12. The estimates of bias reported there correspond to those reported in column (1) of Table 10. We find that the estimates in Table 10 provide an accurate gauge of the bias present in all of the non-experimental estimators we examine other than the IV estimator. The IV estimator constitutes an important exception because it converges slowly and is unstable in small samples.

Table 13 shows the effects on the estimated bias for the base case model presented in Table 10 of changing the relative variances of the observed and unobserved variables affecting earnings and participation. The first two columns vary the contributions of U_{it} and θ_i to the outcome equation error variance, holding the overall variance, $\mathrm{Var}(U_{it}) + \mathrm{Var}(\theta_i)$, fixed. Columns (3) and (4) vary the relative contributions of Z_i and V_i in determining D_i. The final column presents the special case where there is no selection bias in post-program outcomes because $\mathrm{Var}(\theta_i) = 0$ and $\rho = 0$. In each case, the exact values for the variances appear in the table notes.

The results reported in the first two columns are pretty much as expected. The bias for the cross-section estimator increases when the contribution of θ_i to the outcome equation increases. The difference-in-differences estimators are designed for the case where θ_i is an important component of the bias. A comparison of columns (1) and (2) reveals that as the variance of θ_i increases, the bias from using this estimator decreases. The AR(1) estimator is designed to exploit the autoregressive properties of the error term in the outcome equation. Therefore it is not surprising that as the variance due to the autoregressive component declines and the variance due to the fixed effect increases, the performance of the AR(1) estimator deteriorates. The performance of the other estimators is not much affected by the relative variances of θ_i and ε_i. This is not surprising because they do not depend on the time series properties of the error terms in the outcome equation.

The second two columns present the bias in the base case model when the relative variances of the observables, Z, and unobservables, V, in the participation equation are changed, keeping the total variance fixed. These changes affect only the IV and Heckman (1979) estimators, which make explicit use of the participation equation. As the variance of Z declines from column (4) to column (3), the correlation of D with Z drops from 0.0677 to 0.0063, and the quality of both the IV and Heckman (1979) estimators declines, as evidenced by the increases in mean bias and in the variance of the bias across the simulated samples. Both estimators rely on an exclusion restriction and on variation in Z_i relative to the outcome equation error term, although they use this information in different ways. As a result, when the exogenous variation in Z_i is small, the performance of these estimators deteriorates.

The case of no selection bias shown in column (5) of Table 13 is an ideal case for all of the estimators other than the Ashenfelter (1979) estimator, which makes use of Y_{ik}. As expected, almost all of the estimators show a very low estimated mean bias. However, it is surprising that the IV estimator does so poorly in this case. This poor performance reflects the intrinsic variability in the IV estimator already noted in our discussion of Table 12.

Tables 14 and 15 indicate the sensitivity of the estimated biases to different matching rules when the matched comparison group samples are used. Table 14 reports estimates for the base case and Table 15 reports estimates for the common coefficient case. The first four columns present biases from matching on earnings at different lags, or on the sum of earnings over the five pre-program periods. The final column reports estimates based on matching on a propensity score obtained by estimating a probit model of participation including Z_i, $Y_{i,k-1}$, $Y_{i,k-2}$ and $Y_{i,k-3}$ as independent variables.

As noted in Section 7.2, using matching to construct a comparison sample alters the properties of the generated samples compared to random samples and thereby affects the properties of many estimators. The Heckman (1979) estimator is especially vulnerable because matching alters the joint distribution of the unobservables in the participation and outcome equations. The IV estimator is also sensitive to departures from random sampling for reasons analyzed in Section 7.7. In both cases, Table 15, as well as the estimates reported in Table 11 for matched samples, demonstrates that these effects are especially pronounced in the case of the common coefficient model where $\sigma_\alpha = 0$, as the variability in the bias from both estimators is substantially higher in the common coefficient case. As we have stressed throughout this section, and as is already evident in column (3) of Tables 10 and 11, increasing the variance of α_i when the parameter of interest is $\mathrm{E}(\alpha_i \mid D_i = 1)$, and persons select into the program on the basis of α_i (and other variables), reduces bias because more of the variation in D_i results from factors that do not contribute to selection bias.

8.4. Specification testing and the fallacy of alignment

The message of Sections 3–7 is that the choice of an estimator to evaluate a program requires making judgments about outcome equations, participation rules and the rela-

Table 14
Bias in non-experimental estimates of the impact of training (matched comparison group samples); parameter of interest: $E(\alpha \mid D = 1) = 615.7$[a]

Estimator[b]	Base case[c] with matching on Y_{k-3}	Base case with matching on Y_{k-2}	Base case with matching on Y_{k-1}	Base case with matching on sum of Y_{k-5} to Y_{k-1}	Base case with propensity score matching
	(1)	(2)	(3)	(4)	(5)
Cross-section					
Mean	−55.8	−42.9	−16.2	−38.4	−30.4
SD	(71.3)	(80.3)	(75.4)	(72.1)	(73.8)
Diff-in-diff (−1,3)					
Mean	2.6	−5.8	−28.1	−12.4	−40.0
SD	(68.7)	(77.5)	(62.9)	(75.3)	(70.8)
Diff-in-diff (−3,3)					
Mean	−64.5	−43.3	−22.3	−57.8	−40.7
SD	(65.9)	(82.5)	(78.3)	(82.7)	(74.6)
Diff-in-diff (−5,3)					
Mean	−52.6	−33.9	−14.6	−53.4	−31.6
SD	(81.5)	(80.3)	(69.4)	(79.0)	(81.9)
AR(1) regression					
Mean	14.7	−6.9	75.6	22.8	45.8
SD	(315.5)	(333.7)	(324.8)	(345.4)	(366.9)
IV estimator					
Mean	346.1	−305.0	399.2	−192.1	1745.8
Median	−242.5	−41.9	−329.4	−49.5	463.0
SD	(3106.5)	(4001.6)	(26174.9)	(5609.5)	(12488.3)
Corr(Z,D)	0.0937	0.0923	0.0926	0.0931	0.0045

Ashenfelter (1979)					
Mean	80.4	81.5	86.1	77.4	66.6
SD	(80.2)	(83.0)	(83.1)	(73.2)	(71.2)
Heckman (1979)					
Mean	123.5	−15382.5	−4625.8	−93.3	2443.1
Median	−77.4	−238.5	−228.8	81.7	718.4
SD	(3508.1)	(155427.6)	(33888.8)	(7386.2)	(11799.4)
Kitchen sink					
Mean	−19.3	−22.8	−12.4	−21.9	−30.4
SD	(70.3)	(80.2)	(77.0)	(70.1)	(72.6)

[a] Estimates are based on 100 simulated samples of 1000 observations each. The "mean" row presents the mean of the estimates from the 100 samples while the "SD" row presents the standard deviation of the estimates from the 100 samples. The "Corr(Z,D)" row for the IV estimates gives the average correlation between the participation indicator, D, and the instrument, Z. Matching consists of nearest neighbor matching with replacement in all cases. In Columns (1), (2) and (3), matching is on earnings in periods $k-3$, $k-2$ and $k-1$, respectively, where period k is the period of participation for those taking training. In Column (4), matching is on the sum of earnings in the five periods prior to period k. In the final column, matching is on a propensity score calculated by estimating a probit model with participation as the dependent variable and Z, Y_{k-3}, Y_{k-2} and Y_{k-1} as independent variables. The average number of unique observations in a matched sample is 92.1 in column (1), 92.1 in column (2), 92.2 in column (3), 91.7 in column (4) and 91.3 in column (5).

[b] The cross-section estimator is the simple difference between participant and non-participant earnings in period $k+4$. The difference-in-differences estimates are based on the periods indicated, so that $(-1,3)$ is the difference between the change in participant earnings from period $k-1$ to period $k+3$ and the change in non-participant earnings over the same interval. The difference-in-differences $(-3,3)$ estimator is symmetric. The AR(1) estimates are based on a regression of Y_{k+4} on Y_{k-3} and D, with the estimate consisting of the coefficient estimate on D divided by $(1-\rho)$, where ρ is estimated by the coefficient on Y_{k+3}. The IV estimates use Z as an instrument for a regression of Y_{k+4} on D. The Ashenfelter (1979) estimator is described in Section 8.3.3. The dependent variable for this estimator is $Y_{k+4}-Y_k$. The Heckman (1979) estimator is a special case of the class of control function estimators presented in Section 7.4.2. The estimates in all five columns are calculated as shown in Section 7.4.2. The dependent variable for the Heckman (1979) estimator is Y_{k+4}. The kitchen sink estimates are based on a regression of Y_{k+4} on Y_{k-1}, Y_{k-2} and Z.

[c] The "base case", has $\theta \sim N(0,300)$, $\varepsilon \sim N(0,280)$, $Z \sim N(0,300)$, $V \sim N(0,200)$, $\rho = 0.78$, and $\alpha = 100 + N(0,300)$. Estimates for the base case without matching appear in Table 10. This case is based on estimates of the size of the permanent and transitory components of earnings from Ashenfelter and Card (1985) and of the variance in the impacts of training from Heckman et al. (1997c). In the base case, the fractions of $\text{Var}(Y_{k+4} \mid D=1)$ accounted for by α and θ are 0.0564 and 0.2670, respectively.

Table 15
Bias in non-experimental estimates of the impact of training (matched comparison group samples); parameter of interest: $E(\alpha \mid D = 1) = 100.0$[a]

Estimator[b]	Base case[c] with common coefficient and matching on Y_{k-3} (1)	Base case with common coefficient and matching on Y_{k-2} (2)	Base case with common coefficient and matching on Y_{k-1} (3)	Base case with common coefficient and matching on sum of Y_{k-5} to Y_{k-1} (4)	Base case with common coefficient and propensity score matching (5)
Cross-section					
Mean	−287.4	−233.0	−165.3	−199.5	−194.6
SD	(72.3)	(70.4)	(80.0)	(73.2)	(103.6)
Diff-in-diff (−1,3)					
Mean	38.3	−36.8	−178.7	−33.0	−206.8
SD	(82.0)	(77.1)	(80.9)	(74.4)	(100.7)
Diff-in-diff (−3,3)					
Mean	−333.6	−243.2	−157.8	−296.5	−214.5
SD	(72.5)	(70.9)	(88.2)	(74.9)	(115.9)
Diff-in-diff (−5,3)					
Mean	−271.1	−202.0	−141.8	−302.6	−177.9
SD	(76.6)	(76.1)	(87.0)	(71.4)	(108.9)
AR(1) regression					
Mean	103.0	−69.1	−150.4	−144.6	37.6
SD	(1124.9)	(479.3)	(1010.3)	(2357.7)	(567.5)
IV estimator					
Mean	−35.8	27.3	89.4	42.4	9387.0
Median	−18.9	18.1	89.1	49.5	550.3
SD	(167.5)	(175.8)	(162.4)	(165.6)	(86337.9)
Corr(Z,D)	0.4781	0.5035	0.5398	0.5029	0.0109

Ashenfelter (1979)					
Mean	223.1	209.3	171.7	215.2	242.7
SD	(85.6)	(85.0)	(77.1)	(86.7)	(109.6)
Heckman (1979)					
Mean	− 32.3	24.1	81.2	42.5	9499.4
Median	− 23.4	13.9	80.1	52.1	559.7
SD	(167.2)	(177.2)	(171.0)	(167.0)	(87144.0)
Kitchen sink					
Mean	− 161.5	− 194.7	− 201.8	− 162.0	− 187.5
SD	(85.6)	(91.5)	(99.3)	(101.1)	(98.6)

[a] Estimates are based on 100 simulated samples of 1000 observations each. The "mean" row presents the mean of the estimates from the 100 samples while the "SD" row presents the standard deviation of the estimates from the 100 samples. The "Corr(Z,D)" row for the IV estimates gives the average correlation between the participation indicator, D, and the instrument, Z. Matching consists of nearest neighbor matching with replacement in all cases. In Columns (1), (2) and (3), matching is on earnings in periods $k-3$, $k-2$ and $k-1$, respectively, where period k is the period of participation for those taking training. In Column (4), matching is on the sum of earnings in the five periods prior to period k. In the final column, matching is on a propensity score calculated by estimating a probit model with participation as the dependent variable and Z, Y_{k-3}, Y_{k-2} and Y_{k-1} as independent variables. The average number of unique observations in a matched sample is 84.0 in column (1), 79.8 in column (2), 73.0 in column (3), 78.7 in column (4) and 56.3 in column (5).

[b] The cross-section estimator is the simple difference between participant and non-participant earnings in period $k+4$. The difference-in-differences estimates are based on the periods indicated, so that $(-1,3)$ is the difference between the change in participant earnings from period $k-1$ to period $k+3$ and the change in non-participant earnings over the same interval. The difference-in-differences $(-3,3)$ estimator is symmetric. The AR(1) estimates are based on a regression of Y_{k-4} on Y_{k+3} and D, with the estimate consisting of the coefficient estimate on D divided by $(1-\rho)$, where ρ is estimated by the coefficient on Y_{k-3}. The IV estimates use Z as an instrument for a regression of Y_{k+4} on D. The Ashenfelter (1979) estimator is described in Section 8.3.3. The dependent variable for this estimator is $Y_{k+4}-Y_k$. The Heckman (1979) estimator is a special case of the class of control function estimators presented in Section 7.4.2. The estimates in all five columns consist of the coefficient on D when the estimated control functions are included. The dependent variable for the Heckman (1979) estimator is Y_{k+4}. The kitchen sink estimates are based on a regression of Y_{k+4} on Y_{k-1}, Y_{k-2} and Z.

[c] The "base case", has $\theta \sim N(0,300)$, $\varepsilon \sim N(0,450)$, $Z \sim N(0,300)$, $V \sim N(0,200)$, $\rho = 0.78$, and $\alpha = 100$. Estimates for the base case without matching appear in Table 10. This case is based on estimates of the size of the permanent and transitory components of earnings from Ashenfelter and Card (1985). In the base case, with common coefficient, the fractions of $\mathrm{Var}(Y_{k+4} \mid D=1)$ accounted for by α and θ are 0.0000 and 0.3132, respectively.

tionship between the two. All estimators, including social experiments, are based on identifying assumptions which are often difficult if not impossible to test on the available data. For example, the validity of social experiments depends on assumption (5.A.1) or assumptions (5.A.2a) and (5.A.2b), which state that randomization does not disrupt the program being evaluated. Testing for disruption effects turns out to be a difficult task (see Heckman et al., 1996a). Testing whether a variable is a valid instrument is also difficult unless one has access to the true parameter via some other identifying assumption, such as another instrument, a valid social experiment or one of the other identifying restrictions discussed above or in Heckman and Robb (1985a, 1986a). The inability to test maintained identifying assumptions on the available data is a source of frustration to many.

One widely used practice in the evaluation literature apparently evades this problem by testing evaluation models on pre-program data and then using the models that pass the tests to evaluate the program. Papers by Ashenfelter (1978), Ashenfelter and Card (1985) and Heckman and Hotz (1989) exemplify this approach. The idea underlying this approach is that if a selection estimator correctly adjusts for differences in pre-program earnings levels (or some other outcome measure) between future participants and non-participants, it should also adjust correctly for post-program differences and therefore be a valid estimator for evaluating the program. This method could also be applied to the matching estimators defined in Section 7.4.1. According to this line of reasoning, a good match on pre-program outcome levels should produce a valid estimator for post-program levels.

The basic idea underlying this method is captured by the following testing framework. Write $A(Y_{1t'},X_{t'})$ for the adjusted pre-program earnings of program participants and $A(Y_{0t'},X_{t'})$ for the adjusted pre-program earnings of non-participants, where $t' < k$. Then, for a common $X_{t'}$, test the hypothesis

$$A(Y_{1t'}, X_{t'}) = A(Y_{0t'}, X_{t'}). \tag{8.10}$$

Most commonly such tests are based on the model of (3.10). In that context, the test for a valid comparison group is a test of the hypothesis H_0: $\alpha = 0$ in the equation

$$Y_{t'} = X_{t'}\beta + D\alpha + U_{t'}, \qquad t' < k,$$

estimated using pre-program data on participants and comparison group members. Here $D = 1$ denotes that a person will be a participant in period k. If H_0 is not rejected, the comparison group is deemed to be adequate.

This logic seems compelling, but is potentially misleading. The success of testing strategies based on the alignment of pre-program earnings depends on the serial correlation properties of the error term in the earnings equation. Suppose, for example, that program participants and non-participants have identical pre-program earnings histories but that participants experience a permanent loss in earnings at the time of enrollment in period k. In this case, finding that a particular estimator or comparison group correctly aligns earnings in periods prior to k tells little about the validity of a post-

program comparison. Even if the program had a strong positive impact on participant earnings compared to what they would have earned without the program, post-program comparisons between participants and non-participants based on estimators or comparison groups which correctly aligned pre-program earnings might still yield a negative impact estimate for the program because of the large negative shock experienced by participants.

Using tests based on the alignment of pre-program earnings or outcome levels to evaluate the validity of an estimator or comparison group or both is the alignment fallacy. The widely used Heckman and Hotz (1989) tests of the validity of non-experimental selection estimators using pre-program earnings are based on the alignment fallacy. Its practical importance can be illustrated by re-examining an old controversy in the evaluation literature. In the early 1980s, two major consulting firms – Westat, Inc. and SRI International – used matching to construct comparison groups to evaluate the US CETA training program. Both firms had access to the same large datasets and both hired expert statisticians who advocated matching as an evaluation estimator. They both chose their comparison groups to align the earnings of participants and comparison group members in the pre-program period.

As shown in Fig. 3, Ashenfelter's dip characterized the earnings of participants in the CETA program. SRI chose to match on earnings two periods prior to the enrollment period. It picked as comparison group members persons whose earnings were very similar to participants in period $k-2$. Westat aligned using earnings in period $k-1$. Using a simple matching estimator for post-program earnings, SRI reported a negative impact of CETA on participant earnings that was substantially lower than the impact reported by Westat. Figs. 15 and 16 demonstrate how this would happen. Those figures are based on our adaptation of the empirical model of Ashenfelter and Card (1985) used to generate the simulations in Section 8.3. That model is rich enough to generate Ashenfelter's dip. Figs. 15 and 16 show the earnings of participants, matched non-participants, unmatched non-participants and all non-participants for comparison groups based on matching in periods $k-2$ and $k-1$, respectively.

Comparing Figs. 15 and 16, when we match so that future participant and non-participant earnings are the same in period $k-2$, mean reversion causes the earnings after period k of persons aligned in $k-2$ to be higher than those of persons aligned based on earnings in period $k-1$.[88] This implies that the matching estimator used by SRI should produce a lower estimate of program impact than the matching estimator used by Westat, which is exactly what was found. Neither matching estimator may be correct, but the ordering of the estimates obtained from them is predicted from our knowledge of the earnings dynamics of program participants.

Alignment on pre-program earnings is not guaranteed to produce valid estimators of the impact of a program using post-program earnings. It is thus interesting, but not by any

[88] There were other matching variables used by both groups but the use of earnings at different lags to form matched samples plays the main role in explaining the discrepancy between the two studies.

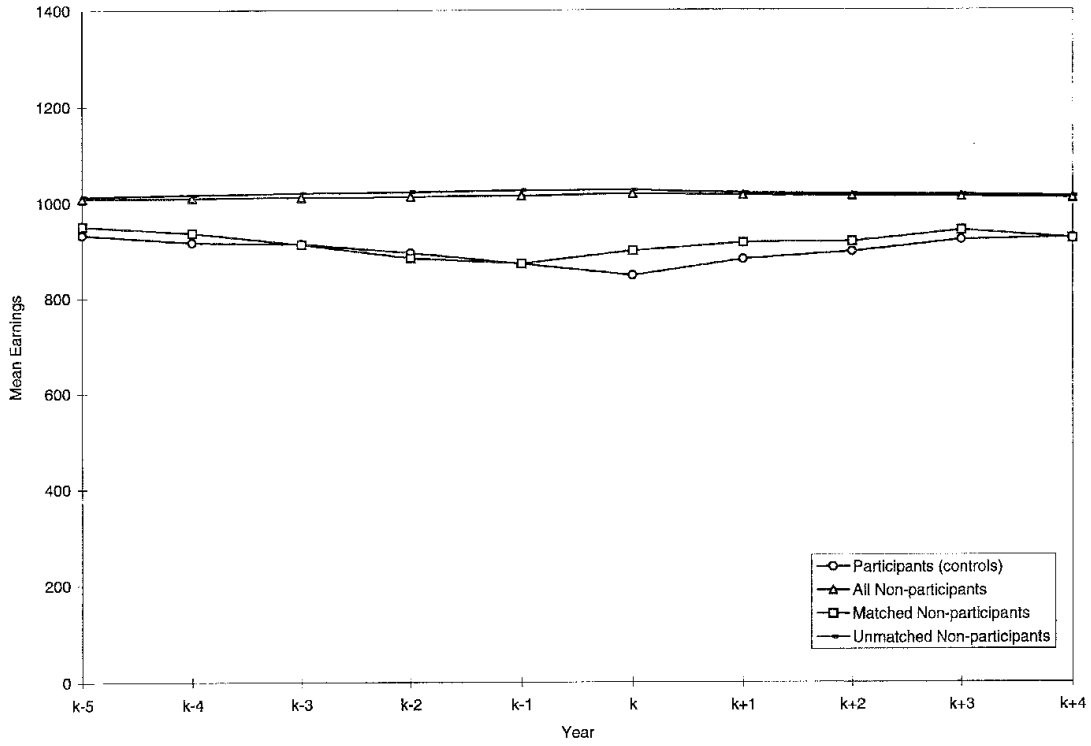

Fig. 16. Mean earnings in base case with random coefficient and matching on Y_{k-1}.

means conclusive, that specification tests based on alignment of pre-program earnings developed by Heckman and Hotz (1989) have been found by them and by others such as Friedlander and Robbins (1995) to eliminate from consideration the most biased estimators of training impact. Even in these studies, many estimators that survive the tests still exhibit substantial bias.

8.4.1. Testing identifying assumptions

As noted by Heckman and Robb (1985a, 1986a), most of the conventional econometric estimators make strong overidentifying restrictions which can be tested. The fixed effects and inverse Mills' ratio estimators are examples of evaluation models with strong over-identifying assumptions.[89] Heckman et al. (1997a) present tests of over-identifying assumptions for matching estimators for non-experimental data.

Nonetheless, Heckman and Robb (1985a, 1986a) also note that all econometric evaluation models can be weakened to a just-identified form, and they present many examples of how this can be done. Just-identified models offer one interpretation of the available data but other just-identified models are equally good descriptions of the same data. The only

[89] Tests of the fixed effect model for panels of length greater than $T = 2$ are presented in Chamberlain (1984), Hsiao (1986) and Baltagi (1995). Tests for the normal selection model based on the properties of censored normal residuals are discussed in Amemiya (1985) among other sources. See also Bera et al. (1984).

way to test the validity of just-identified models is to get better data to eliminate the effects of unobservables on selection.

9. Indirect effects, displacement and general equilibrium treatment effects

Except for our discussion of general equilibrium effects in Section 3.4, throughout this chapter we have followed most of the evaluation literature and used microeconomic partial equilibrium analysis as a framework for interpreting the estimates obtained from evaluation studies. As stated in Section 3, the key identifying assumption in this approach is that the no-treatment outcomes within a given policy regime closely approximate the outcomes in a no-program regime. In the language of Lewis (1963), this assumption allows analysts to ignore indirect effects. In the context of evaluating large scale employment and training programs at a national level, it is natural to ask whether this assumption is valid and the consequences for an evaluation if it is not. To answer these questions in a convincing fashion requires constructing a model of the labor market, a task that is rarely performed in conventional evaluation studies.

In this section, we summarize a line of previous research that attempts to unite the "treatment effect" literature with the general equilibrium policy evaluation literature. Calls for doing so originate in the work of Lewis (1963) and have also been made by Hamermesh (1971, 1993), Johnson and Layard (1986) and others. Within the framework of a Mortensen–Pissarides model, Davidson and Woodbury (1993, 1995) present a promising attempt to analyze the indirect effects of an unemployment bonus program. They assume that prices and wages are fixed and consider the effects of the bonus program on the search behavior of participants and non-participants. In a model with flexible skill prices, Heckman et al. (1998d) consider the effects of changes in tuition on schooling and earnings, accounting for general equilibrium effects on participants and non-participants. We consider both models in this section after briefly surveying the traditional approach to accounting for indirect effects.

Newly trained workers may displace previously trained workers if wages are inflexible, as they are in many European countries. For some training programs in Europe, substantial displacement effects have been estimated (OECD, 1993; Calmfors, 1994). If wages are flexible, the arrival of newly trained workers to the market tends to lower the wages of previously trained workers but does not displace any worker. In the framework of Section 3, even if the effect of treatment on the treated is positive, non-participants may be worse off as a result of the program compared to what they would have experienced in the no-program state. Non-participants who are good substitutes for the new trainees will be especially affected. Complementary factors will benefit. These spillover effects can have important consequences for the interpretation of traditional evaluation parameters. The benchmark "no-treatment" state is actually affected by the program.[90]

[90] Thus assumption (3.15) may be violated and instead $E(Y_0 \mid D = 0, \tilde{\varphi}) < E(Y_0 \mid D = 0, \varphi = 0)$.

To demonstrate these issues in a dramatic way, consider the effect of a wage subsidy for employment in a labor market for low-skill workers. Assume that firms act to minimize their costs of employment. Wage subsidies operate by taking non-employed persons and subsidizing their employment at firms.

As indicated in Table 2, many active labor market policies have a substantial wage-subsidy component. Suppose that the reason for non-employment of low-skill workers is that minimum wages are set too high. This case is a traditional justification for wage subsidies (see, e.g., Johnson, 1979; Johnson and Layard, 1986). If the number of subsidized workers is less than the number of workers employed at the minimum wage, a wage subsidy financed from lump sum taxes has no effect on total employment in the low wage sector because the price of labor for the marginal worker hired by firms is the minimum wage which is the same before and after the subsidy program is put in place. The marginal worker is unsubsidized both before and after the subsidy program is put in place.

The effects of the program are dramatic on the individuals who participate in it. Persons previously non-employed become employed as firms seek workers who carry a wage subsidy. Many previously-employed workers become non-employed as their employment is not subsidized. There are no effects of the wage subsidy program on GDP unless the taxes raised to finance the program have real effects on output. Yet there is substantial redistribution of employment. Focusing solely on the effects of the program on subsidized workers greatly overstates its impact on the economy at large.

In order to estimate the impact of the program on the overall economy, $A(\tilde{\varphi})$ in the notation of Section 3, it is necessary to look at outcomes for both participants and non-participants. Only if the benefits accruing to previously-non-employed participants are adopted as the appropriate criterion would the effect of treatment on the treated be a parameter of interest in this situation. Information on both direct participants and affected non-participants is required to estimate the net gain in earnings and employment resulting from the program.

In the case of a wage subsidy, comparing the earnings and employment of subsidized participants during their subsidized period to their earnings and employment in the pre-subsidized period can be a very misleading estimator of the total impact of the program. So is a cross-section comparison of participants and non-participants. In the example of a subsidy in the presence of a minimum wage, the before–after estimate of the gain exceeds the cross-section estimate unless the subsidy is extended to a group of non-employed workers as large as the number employed at the minimum wage. For subsidy coverage levels below this amount, some proportion of the unsubsidized employment is paid the minimum wage. Under these circumstances, commonly-used evaluation estimators produce seriously misleading estimates of program impacts.

The following example clarifies and extends these points to examine the effect of displacement on the trilogy of estimators discussed in Section 4. Let N be the number of participants in the low-wage labor market. Let N_E be the number of persons employed at the minimum wage M and let N_S be the number of persons subsidized. Subsidized persons receive the minimum wage. Subsidization operates solely on persons who would other-

wise have been non-employed and had no earnings. Assume $N_E > N_S$. Therefore, the subsidy has no effect on total employment in the market, because the marginal cost of labor to a firm is still the minimum wage. Workers with the subsidy are worth more to the firm by the amount of the subsidy S. Firms would be willing to pay up to $S + M$ per subsidized worker to attract them.

The estimated wage gain using a before–after comparison for subsidized participants is

$$\text{Before-after}: \quad S + M - 0 = S + M,$$

because all subsidized persons earn a zero wage prior to the subsidy. The estimated wage gain using cross-section comparisons of program participants and non-participants is

$$\text{Cross-section}: \quad (S + M) - M\frac{(N_E - N_S)}{(N - N_S)} = S + M\left(\frac{N - N_E}{N - N_S}\right) < S + M,$$

where $(S+M)$ is the average participant's wage and $M(N - N_E)/(N - N_S)$ is the average non-participant's wage. Since $N_E > N_S$, the before–after estimator is larger than the cross-section estimator. The widely used difference-in-differences estimator compares the before–after outcome measure for participants to the before–after outcome measure for non-participants.

Difference-in-differences :

$$(S + M - 0) - M\left(\frac{N_E - N_S}{N - N_S} - \frac{N_E}{N - N_S}\right) = S + M\left(\frac{N}{N - N_S}\right) > S + M.$$

The gain estimated from the difference-in-differences estimator exceeds the gain estimated from the before–after estimator which in turn exceeds the gain estimated from the cross-section estimator. The "no-treatment" benchmark in the difference-in-differences model is contaminated by treatment.

The estimate of employment creation obtained from the three estimators is obtained by setting $M = 1$ and $S = 0$ in the previous expressions. This converts those expressions into estimates of employment gains for the different groups used in their definition.

None of these estimators produces a correct assessment of wage or employment gain for the economy at large. Focusing only on direct participants causes analysts to lose sight of overall program impacts. Only an aggregate analysis of the economy as a whole, or random samples of the entire economy, would produce the correct assessment that no wage increase or job creation is produced by the program. The problem of indirect effects poses a major challenge to conventional micro methods used in evaluation research that focus on direct impacts instead of total impacts, and demonstrates the need for program evaluations to utilize market-wide data and general equilibrium methods.

9.1. Review of the traditional approaches to displacement and substitution

Calmfors (1994) presents a comprehensive review of the issues that arise in evaluating

active labor market programs in the context of a modern economy and an exhaustive list of references on theoretical and empirical work on this topic. He distinguishes a number of indirect effects including *displacement effects* (jobs created by one program are at the expense of other jobs), *deadweight effects* (subsidizing hiring that would have occurred in the absence of the program), *substitution effects* (jobs created for a certain category of workers replace jobs for other categories because relative wage costs have changed) and *tax effects* (the effects of taxation required to finance the programs on the behavior of everyone in society). A central conclusion of this literature is that the estimates of program impact from the microeconomic treatment effect literature provide incomplete information about the full impacts of active labor market programs. The effect of a program on participants may be a poor approximation to the total effect of the program, as our simple example has shown.

Forslund and Krueger (1997) illustrate both the traditional approach to estimating displacement and the problems with it. The standard reduced form approach pursued by Johnson and Tomola (1977), Gramlich and Ysander (1981) and others regresses employment in non-subsidized jobs in a geographical area on the number of subsidized jobs lagged one period and other control variables. Full displacement is said to occur if the estimated coefficient on lagged subsidized employment is minus one. For each subsidized job there is one fewer unsubsidized job. For Swedish construction workers, Forslund and Krueger estimate a coefficient of -0.69 so that for each public relief worker hired, there are 0.69 fewer private construction workers hired. For other groups, their estimates of displacement are unstable and they report only a broad range of values.

Forslund and Krueger discuss the problem of reverse causation. A negative shock to the economy may stimulate the use of relief workers. The estimated displacement effect may be a consequence of the feedback between macro shocks and the application of a public hiring policy. Although they present various ad hoc methods based on vector autoregressions to circumvent this problem, they sound a cautionary note about all of the reduced form methods used to estimate displacement and the evidence presented in the entire literature based on them.

9.2. *General equilibrium approaches*

A more clearly interpretable approach to the problem of measuring indirect effects of programs is to construct equilibrium models of the labor market in which both direct and indirect effects are modeled. One recent example is Davidson and Woodbury (1993). They consider these issues in the context of evaluating a bonus scheme to encourage unemployed workers to find jobs more quickly using a Mortensen–Pissarides search model in which prices are fixed. A second recent example is the analysis of Heckman et al. (1998e). They consider the evaluation of tuition subsidy programs in a general equilibrium model of human capital accumulation with both schooling and on-the-job training and with heterogeneous skills in which prices are flexible. The first is a model of displacement with fixed prices; the second is a model of substitution.

Both studies demonstrate the problems with, and possibilities for, general equilibrium analysis of the impacts of active labor market programs. They both find important indirect effects of the programs they evaluate. At the same time, both studies demonstrate that the task of finding credible parameters for general equilibrium models is a challenging one. We first consider the analysis of Davidson and Woodbury.

9.2.1. Davidson and Woodbury

The reemployment bonus scheme analyzed by Davidson and Woodbury (1993) accelerates the rate at which unemployed persons offered the bonus find jobs. The bonus is paid to currently unemployed eligible persons with spells below a threshold level who find jobs within a specified time frame. By stimulating aggregate search activity, the bonus may also have macro effects on output and on the search behavior of unsubsidized participants. The higher taxes raised to finance the program may reduce aggregate search activity by the unsubsidized as their return to market activity declines. The higher level of search by the subsidized may discourage search by their unsubsidized competitors in the labor market.

Davidson and Woodbury (1993) consider four classes of workers: (a) unemployment insurance (UI) recipients who are eligible for the bonus if they get hired; (b) UI recipients who are ineligible for a bonus because of the length of their current unemployment spell (the bonus is only paid to persons with an unemployment spell below a certain length); (c) UI recipients who have exhausted their benefits; and (d) jobless workers who were never eligible to receive UI benefits and cannot receive a bonus. They develop an equilibrium model of search assuming that workers are income maximizing and the bonus is offered in the steady state.

Workers eligible for a bonus have an incentive to accelerate their search. Those ineligible for a bonus in the current spell experience two offsetting effects: (a) the competition for jobs increases, making search less profitable and (b) the benefits of being unemployed rise in the next spell because of the bonus. The second effect promotes search because of the eligibility for the program conferred on persons when they eventually secure a job and are at risk for future unemployment. In their simulations these effects cancel out, leaving the search activity of this group unaffected. However, because of enhanced search by those with the subsidy, the rate of job acquisition declines for those currently ineligible for the bonus.

For those who are permanently ineligible, only the first effect operates; as a result they reduce their search activity. This generates displacement. During recessions, the existence of a bonus leads to displacement of non-bonus workers (those permanently ineligible and those whose benefits are exhausted or whose eligibility has expired). Permanently ineligible workers always experience displacement. Davidson and Woodbury estimate that 30–60% of the gross employment effect of the bonus program is offset by displacement of UI-ineligible workers. Microeconomic treatment analyses of program participant employment experiences provide a substantially misleading picture of the effect of the program on society at large. We next turn to a general equililbrium model of an economy with wage flexibility and indirect effects.

9.2.2. Heckman, Lochner and Taber

The typical microeconomic evaluation of tuition policy estimates the response of college enrollment to tuition variation using geographically dispersed cross-sections of individuals facing different tuition rates. These estimates are then used to determine how subsidies to tuition will raise college enrollment. The impact of tuition policies on earnings are evaluated using a schooling-earnings relationship fit on pre-intervention data and do not account for the enrollment effects of the taxes raised to finance the tuition subsidy. Kane (1994) and Cameron and Heckman (1998) exemplify this approach.

The danger in this widely used practice is that what is true for policies affecting a small number of individuals, as studied by social experiments or as studied in the microeconomic "treatment effect" literature, need not be true for policies that affect the economy at large. A national tuition-reduction policy may stimulate substantial college enrollment and will also likely reduce skill prices. However, agents who account for these changes will not enroll in school at the levels calculated from conventional procedures which ignore the impact of the induced enrollment on skill prices. As a result, standard policy evaluation practices are likely to be misleading about the effects of tuition policy on schooling attainment and wage inequality. The empirical question is: how misleading? Heckman et al. (1998e) show that conventional practices in the educational evaluation literature lead to estimates of enrollment responses that are ten times larger than the long-run general equilibrium effects. They improve on current practice in the "treatment effects" literature by considering both the gross benefits of the program and the tax costs of financing the treatment as borne by different groups.

Evaluating the general equilibrium effects of a national tuition policy requires more information than the tuition-enrollment parameter that is the centerpiece of partial equilibrium policy analysis. Policy proposals of all sorts typically extrapolate well outside the range of known experience and ignore the effects of induced changes in skill quantities on skill prices. To improve on current practice, Heckman et al. (1998e) develop an empirically justified rational expectations perfect foresight overlapping-generations general equilibrium framework for the pricing of heterogeneous skills. It is based on an empirically grounded theory of the supply of schooling and post-school human capital, where different schooling levels represent different skills. Individuals differ in learning ability and in initial endowments of human capital. Household saving behavior generates the aggregate capital stock, and output is produced by combining the stocks of different human capitals with physical capital. Factor markets are competitive and there is price flexibility. The framework explains the pattern of rising wage inequality experienced in the United States in the past 30 years. They apply their framework to evaluate tuition policies that attempt to increase college enrollment.

For two reasons, the "treatment effect" framework that ignores the general equilibrium effects of tuition policy is inadequate. First, the parameters of interest depend on who in the economy is "treated" and who is not. Second, these parameters do not measure the full impact of the program. For example, increasing tuition subsidies may increase the earnings of uneducated individuals who do not take advantage of the subsidy. To pay for the

subsidy, the highly educated would be taxed and this may affect their investment behavior. In addition, more competitors for educated workers enter the market as a result of the policy, and their earnings are depressed. Conventional methods ignore the effect of the policy on non-participants operating through changes in equilibrium skill prices as well as Calmfors' tax effect. In order to account for these effects, it is necessary to conduct a general equilibrium analysis.

The analysis of Heckman et al. (1998e) has major implications for the widely used difference-in-differences estimator. If the tuition subsidy changes the aggregate skill prices, the decisions of non-participants will be affected. The "no-treatment" benchmark group is affected by the policy and the difference-in-differences estimator does not identify the effect of the policy for anyone compared to a no-treatment state.[91]

Using their model, Heckman et al. (1998e) simulate the effects on enrollment in college and wage inequality of a revenue-neutral $500 increase in college tuition subsidy on top of existing programs that is financed by a proportional tax. They start from a baseline economy that describes the US in the mid 1980s and that produces wage growth profiles and schooling enrollment and capital stock data that match micro and macro evidence. The partial equilibrium increase in college attendance is 5.3% in the new steady state. This analysis holds skill prices, and therefore college and high school wage rates, fixed – a typical assumption in microeconomic "treatment effect" analyses.

When the policy is evaluated in a general equilibrium setting, the estimated effect falls to 0.46%. Because the college-high school wage ratio falls as more individuals attend college, the returns to college are less than when the wage ratio is held fixed. Rational agents understand this effect of the tuition policy on skill prices and adjust their college-going behavior accordingly. Policy analysis of the type offered in the "treatment effect" literature ignores the responses of rational agents to the policies being evaluated. There is substantial attenuation of the effects of tuition policy on capital and on the stocks of the different skills in their model compared to a partial equilibrium treatment effect model. They demonstrate that their results are robust to a variety of specifications of the economic model.

They also analyze short-run effects. When they simulate the model with rational expectations, the short-run college enrollment effects are also very small, as agents anticipate the effects of the policy on skill prices and calculate that there is little gain from attending college at higher rates. Under myopic expectations, the short-run enrollment effects are much closer to the estimated partial equilibrium effects. With learning on the part of agents, but not perfect foresight, there is still a substantial gap between partial equilibrium and general equilibrium estimates.

Heckman et al. (1998e) also consider the impact of a policy change on discounted earnings and utility and decompose the total effects into benefits and costs, including tax costs for each group, thus isolating Calmfors' tax effect. Table 16 compares outcomes

[91] This problem of spillover effects was first studied by Lewis (1963) who pointed out its implications for estimating the union–non-union wage differential from cross-section and repeated cross-section comparisons.

across two steady states: (a) the benchmark steady state and (b) the steady state associated with the new tuition policy. Given that the estimated schooling response to a $500 subsidy is small, they instead use a $5000 subsidy for the purpose of exploring general equilibrium effects on earnings. (Current college tuition subsidy levels are this high or higher at many colleges in the US.) The row "High School–High School" reports the change in a variety of outcome measures for those persons who would be in high school under either the benchmark or new policy regime; the "High School–College" row reports the change in the same measures for high school students in the benchmark state who are induced to attend college by the new policy; the "College–High School" outcomes refer to those persons in college in the benchmark economy who only attend high school after the policy; and so forth.

By the measure of the present value of earnings, some of those induced to change are worse off. Contrary to the monotonicity assumption built into the LATE parameter discussed in Section 7, and defined in this context as the effect of the tuition subsidy on the earnings of those induced by it to go to college, they find that the tuition policy produces a two-way flow. Some people who would have attended college in the benchmark regime no longer do so. The rest of society also is affected by the policy – again, contrary to the implicit assumption built into LATE that only those who change status are affected by the policy. People who would have gone to college without the policy and continue to do so after the policy are financially worse off for two reasons: (a) the price of their skill is depressed and (b) they must pay higher taxes to finance the policy. However, they now receive a tuition subsidy and for this reason, on net, they are better off both financially and in terms of utility. Those who abstain from attending college in both steady

Table 16
Simulated effects of $5000 tuition subsidy on different groups; steady state changes in present value of lifetime wealth (thousands of 1995 US dollars)[a]

Group (proportion)[b]	After-tax earnings using base tax[c] (1)	After-tax earnings[c] (2)	After-tax earnings net of tuition[c] (3)	Utility[c] (4)
High School–High School (0.528)	9.512	−0.024	−0.024	−0.024
High School–College (0.025)	−4.231	−13.446	1.529	1.411
College–High School (0.003)	−46.711	−57.139	−53.019	−0.879
College–College (0.444)	−7.654	−18.204	0.42	0.42

[a] Source: Heckman et al. (1998e, Table 1).

[b] The groups correspond to each possible counterfactual. For example, the "High School–High School" group consists of individuals who would not attend college in either steady state, and the "High School–College" group would not attend college in the first steady state, but would in the second, etc.

[c] Column (1) reports the after-tax present value of earnings in thousands of 1995 US dollars discounted using the after-tax interest rate where the tax rate used for the second steady state is the base tax rate. Column (2) adds the effect of taxes, column (3) adds the effect of tuition subsidies and column (4) includes the non-pecuniary costs of college in dollar terms.

states are better off in the second. They pay higher taxes, but their skill becomes more scarce and their wages rise. Those induced to attend college by the policy are better off in terms of utility but are not necessarily better off in terms of income. Note that neither category of non-changers is a natural benchmark for a difference-in-differences estimator. The movement in their wages before and after the policy is due to the policy and cannot be attributed to a benchmark "trend" that is independent of the policy.

Table 17 presents the impact of the $5000 tuition policy on the log earnings of individuals with 10 years of work experience for different definitions of treatment effects. The partial equilibrium version given in the first column holds skill prices constant at initial steady state values. The general equilibrium version given in the second column allows prices to adjust when college enrollment varies. Consider four parameters initially defined in a partial equilibrium context. The *average treatment effect* is defined for a randomly selected person in the population in the benchmark economy and asks how that person would gain in wages by moving from high school to college. The parameter *treatment on the treated* is defined as the average gain over their non-college alternative of those who

Table 17
Treatment effect parameters under partial equilibrium and general equilibrium; difference in log earnings for college graduates versus high school graduates at 10 years of work experience[a]

Parameter	Prices fixed[b] (1)	Prices vary[c] (2)	Fraction of sample[d] (%) (3)
Average treatment effect (ATE)	0.281	1.801	100
Treatment on treated (TT)	0.294	3.364	44.7
Treatment on untreated (TOU)	0.270	−1.225	55.3
Marginal treatment effect (MTE)	0.259	0.259	–
LATE[e] 5000 subsidy			
Partial equilibrium	0.255	–	23.6
GE (HS to college) (LATE)	0.253	0.227	2.48
GE (college to HS) (LATER)	0.393	0.365	0.34
GE Net (TLATE)	–	0.244	2.82
LATE[e] 500 subsidy			
Partial equilibrium	0.254	–	2.37
GE (HS to college) (LATE)	0.250	0.247	0.24
GE (college to HS) (LATER)	0.393	0.390	0.03
GE Net (TLATE)	–	0.264	0.27

[a] Source: Heckman et al. (1998e).

[b] In column (1), prices are held constant at their initial steady state levels when wage differences are calculated.

[c] In column (2), we allow prices to adjust in response to the change in schooling proportions when calculating wage differences.

[d] For each row, column (3) presents the fraction of the sample over which the parameter is defined.

[e] The LATE group gives the effect on earnings for persons who would be induced to attend college by a tuition change. In the case of GE, LATE measures the effect on individuals induced to attend college when skill prices adjust in response to quantity movements among skill groups. The partial equilibrium LATE measures the effect of the policy on those induced to attend college when skill prices are held at the benchmark level.

attend college in the benchmark state. The parameter *treatment on the untreated* is defined as the average gain over their college wage received by individuals who did not attend college. The *marginal treatment effect* is defined for individuals who are indifferent between going to college or not. This parameter is a limit version of the LATE parameter under conventional assumptions made in discrete choice theory (Heckman, 1997; Heckman and Vytlacil, 1999a,b). Column 2 presents the general equilibrium version of *treatment on the treated.* It compares the earnings of college graduates in the benchmark economy with what they would earn if no one went to college.[92] The treatment on the untreated is defined analogously by comparing what high school graduates in the benchmark economy would earn if everyone in the population were forced to go to college. The *average treatment effect* compares the average earnings in a world in which everyone attends college versus the earnings in a world in which nobody attends college. Such dramatic policy shifts produce large estimated effects. In contrast, the general equilibrium marginal treatment effect parameter considers the gain to attending college for people on the margin of indifference between attending college and only attending high school. In this case, as long as the mass of people in the indifference set is negligible, partial and general equilibrium parameters are the same.

The final set of parameters Heckman et al. (1998e) consider are versions of the LATE parameter. This parameter depends on the particular intervention being studied and its magnitude. The partial equilibrium version of LATE is defined on the outcomes of individuals induced to attend college, assuming that skill prices do not change. The general equilibrium version is defined for the individuals induced to attend college when prices adjust in response to the policy. The two LATE parameters are quite close to each other and are also close to the marginal treatment effect.[93] General equilibrium effects change the group over which the parameter is defined compared to the partial equilibrium case. For the $5000 subsidy, there are substantial price effects and the partial equilibrium parameter differs substantially from the general equilibrium parameter.

Heckman et al. (1998e) also present partial and general equilibrium estimates for two extensions of the LATE concept: LATER (the effect of the policy on those induced to attend only high school rather than going to college) – Reverse LATE – and TLATE (the effect of the policy on all of those induced to change whichever direction they flow). LATER is larger than LATE, indicating that those induced to drop out of college have larger gains from dropping out than those induced to enter college have from entering. TLATE is a weighted average of LATE and LATER with weights given by the relative proportion of people who switch in each direction.

The general equilibrium impacts of tuition on college enrollment are an order of

[92] In the empirical general equilibrium model of Heckman et al. (1998d), the Inada conditions for college and high school are not satisfied in the aggregate production function and the marginal product of each skill group when none of it is utilized is a bounded number. If the Inada conditions were satisfied, this counterfactual and the counterfactual treatment on the untreated would not be defined.

[93] The latter is a consequence of the discrete choice framework for schooling choices analyzed in the Heckman et al. (1998d) model. Recall our discussion in Section 3.4.

magnitude smaller than those reported in the literature estimating microeconometric treatment effects. The assumptions used to justify the LATE parameter in a microeconomic setting do not carry over to a general equilibrium framework. Policy changes, in general, induce two-way flows and violate the monotonicity – or one-way flow – assumption of LATE. Heckman et al. (1998e) extend the LATE concept to allow for the two-way flows induced by the policies. They present a more comprehensive approach to program evaluation by considering both the tax and benefit consequences of the program being evaluated and placing the analysis in a market setting. Their analysis demonstrates the possibilities of the general equilibrium approach and the limitations of the microeconomic "treatment effect" approach to policy evaluation.

9.3. Summary of general equilibrium approaches

Any policy with a large target population is likely to have general equilibrium impacts. Reliance on microeconomic treatment effect approaches to evaluate such policies produces potentially misleading estimates. Even reducing the Heckman et al. (1998e) estimates by a factor of three to account for learning about future price paths, instead of perfect foresight, produces a sizeable discrepancy between the microeconomic treatment effect estimates and the general equilibrium estimates. Their work and that of Davidson and Woodbury (1993) indicates that the costs of ignoring indirect effects may be substantial. In future evaluations of large scale programs, we urge the use of general equilibrium methods to produce more accurate assessments of the true impacts of the programs being evaluated and to produce a more reliable guide to the distributional impacts of policies.

The cost of this enhanced knowledge is the difficulty in assembling all of the behavioral parameters required to conduct a general equilibrium evaluation. From a long-run standpoint, these costs are worth incurring. Once a solid knowledge base is put in place, a more trustworthy framework for policy evaluation will be available, one that will offer an economically justified framework for accumulating evidence across studies and will motivate empirical research by microeconomists to provide better empirical foundations for general equilibrium policy analyses.

10. A survey of empirical findings

10.1. The objectives of program evaluations

The purpose of government training programs and other active labor market policies is to integrate unemployed and economically disadvantaged workers into the work force either by facilitating their job search, improving their work habits, or augmenting their human capital. In Section 3, we emphasized that program evaluators could assess the success of these programs by their impacts on a variety of outcomes, the choice of which depended on the objectives of policy makers. In practice, the outcomes of greatest interest to program evaluators and to policy makers who fund this research include participants'

labor market outcomes, such as their earnings, employment rates, transition rates out of unemployment and employment, wages, and use of unemployment insurance programs. Participants' non-labor market outcomes, such as their use of social assistance programs, educational attainment, criminal activity, and teen childbearing, are also scrutinized.

The most common outcomes of interest in US program evaluations are annual or quarterly earnings. Positive earnings impacts are often taken as synonymous with increased aggregate output and costs are often ignored. By contrast, in European evaluations the most common outcome of interest is employment. This emphasis reflects an emphasis on programs that reduce longterm unemployment.

Besides examining the impact of active labor market policies on participants' outcomes, another objective of program evaluations is to determine whether these policies constitute worthwhile social investments. The dominant approach followed in the program evaluation literature is to measure the net social benefit of these policies using the change in aggregate output attributable to the program (Heckman and Smith, 1998a). Evaluators estimate this change by subtracting the programs' costs from its discounted stream of benefits. These costs include the operating cost of the program, the cost of education and training expenditures, forgone earnings associated with participants' time in the program, and participants' out-of-pocket expenses for inputs such as transportation and child care.

In some cases, only the direct costs of these programs are likely significant in conventional cost-benefit analyses. The forgone earnings costs of participating in training are less important when evaluating JSA or short WE programs, or for programs targeted toward economically disadvantaged persons who are prone to long spells of unemployment. By contrast, these costs tend to be higher when evaluating a CT program in which individuals acquire skills off-the-job, and their participation in the program causes them to search less intensely for employment. Similarly, these costs are higher for programs serving adult males, especially prime-aged displaced workers, who have well established work histories and who are more likely to be employed in the absence of training.

In practice, conventional cost-benefit analyses usually do not account for several other costs that could reduce the net social benefit of employment and training programs. The first of these costs are the deadweight loss caused by raising taxes to finance training (Browning, 1987). The likely importance of these costs depends on the group being served. These costs should be higher for participants who are not receiving social assistance benefits. Often program evaluations report that the earnings impact of training is offset to some extent by a reduction in social assistance, so that participants' incomes may be little changed as a result of the program (see, e.g., Friedlander et al., 1985). This result implies that the deadweight loss associated with raising taxes to pay for training would be reduced to some extent because of savings in deadweight losses due to reduced taxes required to pay participants' future social assistance benefits.

A second cost usually unaccounted for in program evaluations is the value of participants' reduced leisure time (Greenberg, 1997). In principle, such costs depend on the shape of labor supply curves for different groups of participants. The value of participants' reduced leisure time may be especially significant for economically disadvantaged

women. If these women are the primary child care providers in their households, the social (as well as the private) cost associated with their time away from the home may be significant.

Finally, a third cost usually unaccounted for, especially in US program evaluations, is the cost associated with displacement of non-training participants (Hamermesh, 1971, 1993; Johnson, 1979). As discussed in Section 9, a potentially important policy parameter is the impact of the program on non-participants. If non-participants are displaced from jobs as a result of providing employment and training opportunities to participants, the program may have no impact on aggregate output. In the US, where a larger share of training dollars is spent on CT, the size of the programs compared to the economy is very small, and real wages have been relatively flexible, these costs are relatively small. In such instances, the estimated earnings impacts of the program may closely approximate the impact of the program on aggregate output.

By contrast, many European countries' active labor market polices include substantial expenditures on wage subsidies. These policies in the context of less flexible labor markets suggest that the cost of displacement, substitution and deadweight, as defined in Section 9, can be substantial. Evidence on this is given for the United Kingdom by Begg et al. (1991) and Dolton (1993), for the Netherlands by de Koning (1993) and for Sweden by Forslund and Krueger (1997). See Calmfors (1994) for a general survey.

The benefits from employment and training programs can come from several sources. By design, the discounted earnings impacts should be an important social benefit of most successful programs. In principle, other outcomes also could yield substantial social benefits. These outcomes include the value of output produced by trainees while in training, and the savings in administrative costs because of participants' reduced use of social welfare and of other education and training programs. Further, if improved employment prospects reduce asocial behaviors, society also may benefit from reduced expenditures on the criminal justice system, on substance abuse treatment centers, or on child welfare services. These latter benefits are potentially large for younger, less educated, training participants who are more inclined to engage in such asocial behavior (Mallar et al., 1982; LaLonde, 1995; Heckman and Smith, 1998a).

As shown by Table 18, the primary social benefit reported in most cost-benefit analyses of employment and training programs is the discounted earnings gains. Although this table surveys only a few analyses for economically disadvantaged women, these results are typical of those reported in other studies. Usually, these earnings benefits are one or two orders of magnitude larger than the other measured benefits of these programs. Because of the importance of earnings impacts for conventional cost-benefit analyses, it is important that analysts obtain credible and precise estimates of their magnitude.

The importance of estimated earnings impacts to cost-benefit analyses of employment and training programs highlights an important shortcoming of these analyses. As shown by the last row in Table 18, most program evaluations follow participants only for a couple of years following their entry into the program. Often the earnings impacts during this period

Table 18
Accounting of estimated social benefits and costs per treatment in selected social experiments evaluating employment and training services for female welfare applicants and recipients (1997 US dollars)[a]

	Program/main services provided					
	NSW WE	San Diego CWEP/ JSA/WE	San Diego CWEP/ JSA Only	San Diego SWIM/ JSA/CT	Florida PI/ JSA	NJS All
Benefits						
Increased output from employment (includes earnings and fringe benefit impacts)	24486	3571	2457	2913	757	2066
From projected period only[b]	19084	2101	1161	0	298	0
Value of in-program output	12039	3280	−5	262	NA	NA
Reduced cost of using transfer programs	2160	131	82	53	130	NA
Reduced cost of using other programs (e.g., other education and training programs)	1619	85	74	NA	NA	NA
Costs						
Program operating costs, including JSA or WE	−13850	−968	−857	−866	−417	−1421[d]

Education and training costs	0	0	0	−360	−846	NA[d]
Forgone earnings and fringe benefits[c]	−2341	NA	NA	NA	NA	NA
Participant out-of-pocket expenses (e.g., transportation, child care, clothing costs)	−431	−24	–[e]	–	–	–
Value of reduced non-market time	–	–	–	–	–	–
Displacement of other workers	–	–	–	–	–	–
Deadweight loss from taxes to pay for programs	–	–	–	–	–	–
Net present value of benefits minus costs	21708	3123	1753	2003	−377	645
Observation period in years	2.25	1.00	1.16	5.00	2.00	2.50

[a] Sources: Kemper et al. (1981, Table IV.1, p. 100, Table IV.2, p. 106, Table IV.6, p. 121); Goldman et al. (1986, p. 139, Table 5.4, p. 153, Table 5.8, p. 166); Friedlander and Hamilton (1993, Table 5.1, p. 57); Kemple et al. (1995, Table 7.3, p. 174, Table 7.5, p. 177); Orr et al. (1994, Exhibit 6.2, p. 162).

[b] Projected earnings are based on earnings impacts during the last four quarters of the observation period and are discounted at a rate of 5% per year. Subsequent earnings impacts were assumed to depreciate at a rate of 25% per year.

[d] The −1421 figure for the NJS includes both program operating costs and education and training costs.

[c] For all but the first column, the social costs associated with forgone earnings are embodied in the estimates of the increased output from employment. In these cases this measure is net of the forgone earnings costs of the program.

[e] A "–" denotes costs that were not estimated in the indicated study.

are insufficient to justify the programs' costs. Cost-benefit analyses in this literature customarily project the earnings impacts obtained during the observation period into the future, sometimes for as long as participants' expected working life, and then discount these projected impacts at rates ranging from 0 to 15% (see, e.g., Kemper et al., 1981, pp. 174–177, Table VIII.2). In addition, evaluators sometimes allow these projected impacts to decay through time (see, e.g., the references in Table 18).

As shown by the third row of Table 18, the projected earnings gains can constitute a significant portion of the total earnings gains associated with the program. In the most extreme case, more than three-fourths of the earnings impact used in the cost-benefit analysis of the NSW Demonstration was based on out-of-sample projections. Because the estimated benefits from the reduced use of other social programs by NSW treatments also were based on similar projections, the net social benefit of the NSW Demonstration during the first 27 months is actually negative. This evidence underscores the importance of funding the collection of data that enable longterm evaluations of employment and training programs.

For evaluations that look only at post-program earnings impacts, another potential source of benefits from active labor market policies is the value of the output produced by participants while they were in the program.[94] As shown by the first column of Table 18, this benefit constituted a significant fraction of the total benefit from the NSW program. This result is expected in programs that provide WE compared with those that provide JSA or CT. When training consists of a subsidized job in the public or non-profit sector, evaluators assume that the work performed by participants is valuable to society. Because the NSW program provided relatively longterm WE to a large percentage of participants, the value of in-program output is large compared to other programs. In contrast, the WE in the San Diego CWEP program shown in the second column lasted only a few weeks and was provided to only a small fraction of participants. As a result the value of in-program output was small.

To demonstrate the sensitivity of conventional cost-benefit analyses to assumptions about the costs and benefits of employment and training programs, we reexamine the net social benefits of the WE provided in the NSW Demonstration and of JTPA services provided in the NJS. Since the "final reports" for these two studies were published, there have been two subsequent studies that have followed participants for up to 8 and 5 years, respectively (Couch, 1992; US General Accounting Office, 1996). Both of these studies indicate that the positive shortterm earnings impacts originally reported for adult women in the NSW Demonstration and adults in the NJS persisted, whereas neither program had a significant short- or longer-term impact on youths' earnings.

As shown by Table 19, the estimated net social benefit of treatments' access to WE in the NSW Demonstration are negative for youths, but are sometimes positive for adult

[94] When the in-program period is included in the estimation of program impacts, and participants are paid a market wage, the value of in-program output is implicitly included in the impact estimate because it is reflected in the participants' earnings.

Table 19
Net social returns and internal rates of return: National Supported Work Demonstration (impacts and costs in 1978 US dollars)[a]

Benefit duration	Welfare cost of taxation	AFDC Women IRR[b]	Youth IRR	Annual discount rate	AFDC Women net social benefit	Youth net social benefit
3 years	0.00	<0	<0	0.00	−2152	−1528
				0.05	−2167	−1541
				0.10	−2180	−1553
	0.50	<0	<0	0.00	−3489	−2406
				0.05	−3504	−2419
				0.10	−3517	−2430
	1.00	<0	<0	0.00	−4826	−3283
				0.05	−4841	−3296
				0.10	−4854	−3308
8 years	0.00	0.005	<0	0.00	54	−1463
				0.05	−428	−1482
				0.10	−789	−1499
	0.50	<0	<0	0.00	−1283	−2341
				0.05	−1765	−2359
				0.10	−2126	−2377
	1.00	<0	<0	0.00	−2620	−3218
				0.05	−3102	−3237
				0.10	−3463	−3254
Indefinite	0.00	0.136	<0	0.00	NA[c]	NA
				0.05	4648	−1942
				0.10	961	−1658
	0.50	0.091	<0	0.00	NA	NA
				0.05	3311	−2820
				0.10	−376	−2535
	1.00	0.068	<0	0.00	NA	NA
				0.05	1974	−3697
				0.10	−1713	−3413

[a] Sources: Impact estimates are taken from Couch (1992). Cost estimates are taken from Kemper et al. (1984, Table 8.6). Estimates of the welfare cost of taxation fall within the range given in Browning (1987).

[b] IRR, internal rate of return, the rate of return at which the discounted benefits from the program equal the current costs. Welfare costs of taxation are in dollars of welfare loss per tax dollar.

[c] NA indicates that net social benefits equal positive or negative infinity due to the absence of discounting.

women. However, the table reveals that these estimates are sensitive to the duration of the earnings impacts, the discount rate used in the analysis, and whether the analysis takes into account the deadweight losses associated with the taxes that finance the program. Although the earnings impacts for adult women are positive during the first 3 years,

these impacts by themselves are insufficient to generate positive net social benefits from the program.[95] The importance of the followup study by Couch (1992) is seen in the estimates of the net social benefits from the program when the benefits last for 8 years or indefinitely. In the latter case, estimates based on a variety of plausible assumptions about the appropriate discount rate and the deadweight loss associated with taxes all imply that the estimated net social benefit from the program is positive for AFDC women.

A useful metric for comparing the net social benefits of different active labor market policies to other investments is their internal rate of return (IRR). This measure is the discount rate for which the discounted stream of benefits from the program equals its costs. The IRR allows a comparison between alternative investment projects using a common metric. As shown by the middle columns of Table 19, if we assume that the deadweight loss associated with taxes used to finance the program are 50% or more, the IRR for WE targeted toward adult women is negative for benefit durations of 8 years or less. If the earnings impacts persist indefinitely, the IRR is 9.1%. Thus, for adult women in the NSW, the overall net benefit calculations still depend on projections of earnings gains outside the available data.

Comparing Tables 20 and 21, the cost-benefit analyses indicate that JTPA services generated a substantial net social benefit when targeted toward adults, but none when targeted toward youths. As with the NSW Demonstration, these estimated net social benefits are sensitive to the assumptions underlying the analysis. In the absence of a longterm followup study, we would be less confident about whether JTPA constituted a worthwhile social investment. However, as a result of the followup study, we are more confident that after 5 years the net social benefit per treatment group member ranges from 600 to 2000 and that the IRR are very large. Further, if these gains were to persist indefinitely, it would appear that the JTPA services provided adults in the NJS constituted an extraordinarily successful public investment. By contrast, as shown by Table 21, estimates based on short- and medium-term earnings impacts indicate that JTPA services targeted toward youths constituted a poor social investment. As shown by the last rows of the table, projections of these impacts into the future produce the only positive net social returns for this group. However this result is very tenuous because these projections are based on point estimates for the fifth (followup) year that are not statistically significant.

10.2. The impact of government programs on labor market outcomes

Credible cost-benefit analyses of employment and training programs depend on credible estimates of the costs and benefits of these programs. Because labor market outcomes appear to constitute such an important source of the social benefits from these programs

[95] As we discussed in Section 5, the impact estimates for the NSW Demonstration differ depending on whether the survey earnings data or the administrative earnings data are used, with the estimates based on the survey measures showing a larger positive impact. Because the only longterm followup impact estimates are based on the administrative data (Couch, 1992), we use them in constructing the estimates in Table 19.

Table 20
Net social returns and internal rates of return: National JTPA Study – Adults (impacts and costs in nominal US dollars)[a]

Benefit duration	Welfare cost of taxation	Adult women IRR[b]	Adult men IRR	Annual discount rate	Adult women net social benefit	Adult men net social benefit
3 years	0.00	1.390	>2	0.00	863	1097
				0.05	778	1017
				0.10	702	948
	0.50	0.416	1.787	0.00	485	844
				0.05	400	765
				0.10	324	696
	1.00	0.064	0.689	0.00	107	592
				0.05	22	513
				0.10	−54	443
5 years	0.00	1.610	>2	0.00	1822	1979
				0.05	1589	1766
				0.10	1395	1589
	0.50	0.693	>2	0.00	1443	1726
				0.05	1211	1514
				0.10	1017	1336
	1.00	0.362	0.960	0.00	1065	1474
				0.05	833	1261
				0.10	638	1084
Indefinite	0.00	1.620	>2	0.00	NA[c]	NA
				0.05	7889	6859
				0.10	3891	3607
	0.50	0.738	>2	0.00	NA	NA
				0.05	7510	6607
				0.10	3513	3354
	1.00	0.455	0.985	0.00	NA	NA
				0.05	7132	6354
				0.10	3134	3102

[a] Impact estimates are taken from US General Accounting Office (1996). Cost estimates are taken from Orr et al. (1994). Estimates of the welfare cost of taxation fall within the range given in Browning (1987).

[b] IRR, internal rate of return, the rate of return at which the discounted benefits from the program equal the current costs. Welfare costs of taxation are in dollars of welfare loss per tax dollar.

[c] NA indicates that net social benefits equal positive or negative infinity due to the absence of discounting.

and because these outcomes are relatively easily measured, they are the focus of much of the evaluation literature. There is much less emphasis in the literature on the impact of these programs on non-labor market outcomes.

There have been many surveys of the impact of US programs on labor market outcomes, especially on participants' employment rates and earnings (see, e.g., Perry et al., 1975;

Table 21
Net social returns and internal rates of return: National JTPA Study – Youth (impacts and costs in nominal US dollars)[a]

Benefit duration	Welfare cost of taxation	Female youth IRR[b]	Male youth IRR	Annual discount rate	Female youth net social benefit	Male youth net social benefit
3 years	0.00	<0	<0	0.00	−982	−2196
				0.05	−979	−2145
				0.10	−976	−2101
	0.50	<0	<0	0.00	−1413	−2849
				0.05	−1410	−2798
				0.10	−1407	−2754
	1.00	<0	<0	0.00	−1844	−3502
				0.05	−1841	−3451
				0.10	−1838	−3407
5 years	0.00	<0	<0	0.00	−434	−1158
				0.05	−515	−1281
				0.10	−580	−2027
	0.50	<0	<0	0.00	−865	−1811
				0.05	−946	−1934
				0.10	−1011	−2027
	1.00	<0	<0	0.00	−1296	−2464
				0.05	−1377	−2587
				0.10	−1442	−2680
Indefinite	0.00	0.163	0.163	0.00	NA[c]	NA
				0.05	2995	10880
				0.10	811	3444
	0.50	0.122	0.122	0.00	NA	NA
				0.05	2564	10227
				0.10	380	2791
	1.00	0.098	0.098	0.00	NA	NA
				0.05	2133	9573
				0.10	−51	2138

[a] Impact estimates are taken from US General Accounting Office (1996). Cost estimates are taken from Orr et al. (1994). Estimates of the welfare cost of taxation fall within the range given in Browning (1987).

[b] IRR, internal rate of return, the rate of return at which the discounted benefits from the program equal the current costs. Welfare costs of taxation are in dollars of welfare loss per tax dollar.

[c] NA indicates that net social benefits equal positive or negative infinity due to the absence of discounting.

Grossman et al., 1985; Bassi and Ashenfelter, 1986; Barnow, 1987; Gueron, 1990; LaLonde, 1995; Friedlander et al., 1997). By contrast, there are few surveys of the impacts of these programs operated outside the US (see, e.g., Bradley, 1994; Fay, 1996). Consequently, to address this imbalance we devote a substantial portion of this section to

summarizing what has been learned from evaluations of European programs, and how these studies compare to US evaluations.

Before surveying the results from these evaluations it is helpful to consider what kind of impact on earnings we should expect from public sector employment and training programs. If we believe that the objective of these programs is to augment human capital, the literature on education and earnings provides a useful starting point. This literature indicates that an additional year of schooling is associated with approximately a 10% increase in the typical worker's earnings (Ashenfelter and Rouse, 1995). In many countries, the return to schooling is smaller, as it also has been in the past in the United States. The cost of a year of education includes direct instructional expenditures, any forgone earnings, and other inputs from the family and the community. Formal schooling is usually more intensive and costly than public sector employment and training programs. As a result, it would be surprising if such programs, which usually last far less than a year, consistently led to larger increases in earnings than an additional year of schooling. By analogy, the relatively few programs that are more intensive and costly than a year of schooling should generate larger earnings gains (Heckman et al., 1993; LaLonde, 1995). Accordingly, a training program costing several hundred dollars or even a few thousand dollars per participant would likely lead to annual earnings gains of at most several hundred dollars. Earnings gains much larger than this would suggest that these programs generate large social returns compared to formal schooling and to other investments in general. In this vein, the results in Table 20 showing very high estimated internal rates of return from JTPA are unexpectedly large.

The evidence both from North American and European studies indicates that government employment and training programs have at best a modest positive impact on adult earnings. Further, when longer term followup data are available, these gains do not always persist. The evidence suggests that the gains, when they occur, are more likely the result of an increased probability of employment than of increased wages. Indeed, the case for these programs increasing participants' subsequent hourly wages remains weak. The finding that earnings gains are in large measure the result of increased employment rates raises the question of how active labor market polices affect non-participants through displacement. In the US especially, this issue has received relatively little empirical attention (but, see Davidson and Woodbury, 1993, 1995).

Among youths, the evidence is mixed between the two continents. In the US, studies consistently report that these programs have no impact (or sometimes even a negative impact) on youths' earnings. By contrast, in Europe, studies of less economically disadvantaged youths find that these programs sometimes substantially raise employment rates, because they raise transition rates out of unemployment. At the same time, however, other studies (sometimes of participants in the same program) report no effect on employment. Such results suggest either that there is substantial heterogeneity in impacts among cohorts or that these impacts possess the same sensitivity to econometric specification that we documented for the US CETA studies. In any case, as with adults there is little evidence that these programs raise wages.

Part of the reason that there is little evidence on the relation between government training programs and wages has to do with the quality of data available for program evaluations. Hourly wage data are unavailable in many program evaluations, especially those conducted in the United States. Further, when these measures are available, the sample sizes often are too small to estimate wage impacts with precision. From a human capital perspective, the wage gains associated with these programs should be small. The experiences with the NSW Demonstration and NJS illustrate this problem. In the former program the wages of adult women in the treatment group were approximately 8% higher than those in the control group; in the later program wages of adults in the treatment group were 2–3%% higher than those in the control group (Masters and Maynard, 1981; Orr et al., 1994). Neither of these impacts were statistically significant at conventional levels. However, given the costs of these programs, these point estimates are not surprising. Indeed, they compare favorably to estimates of the wage impacts associated with a year of community college schooling in the US (Kane and Rouse, 1993).

At the same time, this characterization of the empirical results from the program evaluation literature masks substantial heterogeneity in the estimated impacts which vary widely among programs, among field offices and among different demographic and skill groups. In many instances, the evidence suggests that training either had no effect or may have lowered earnings, while in other cases the impacts are so large that programs such as JTPA appear to generate substantial internal (and private) rates of return. Indeed, for economically disadvantaged adult women residing in the US, a case can be made that these programs consistently have been a productive social investment, whose returns are larger than those from formal schooling. For other groups this conclusion clearly does not hold. In particular, there appears to be a weak tendency in the literature suggesting that the earnings impacts and the net social returns from many active labor market policies, particularly those that provide training, are smaller for the least skilled participants.

10.3. The findings from US social experiments

As explained in Section 5, an unusual characteristic of the empirical literature on active labor market policies is that it includes a relatively large number of both experimental and non-experimental studies. However, because treatment non-participation and control group substitution are often substantial, the parameter measured in experimental studies is the effect of the "intention to treat" and not the impact of "treatment on the treated" (Heckman et al., 1998f). Dropping out similarly afflicts non-experimental studies, and contamination bias is the counterpart to control group substitution. Accordingly, although the estimates reported in the experimental literature are usually thought to be different from those in the non-experimental literature, it is easy to exaggerate the differences. Nonetheless, because the estimates reported in both literatures do not adjust for these biases, and because the incidence of the various biases may differ between experimental and non-experimental studies, it is likely that different parameters are

estimated in these diverse literatures. For these reasons, we survey the two literatures separately.

Provided the assumptions discussed in Section 5 hold, social experiments yield easily computed and widely understood estimates of the "the intention to treat" on the treatments' outcomes. As shown by Table 22, collectively, the US experimental evaluations provide some compelling evidence that the opportunity to receive these services sometimes can improve participants' employment prospects and that the resources spent on these services can pass a standard cost-benefit test. The most consistent evidence in this regard is found for adult women.[96] As shown by Table 22, the earnings gains received by adult women assigned to the treatment group are (i) usually modest in size ranging from a few hundred dollars to more than one thousand dollars, annually, (ii) often persist at least for several years without signs of decay, (iii) arise from a variety of intended treatments, and (iv) sometimes appear to be remarkably cost effective, at least before the deadweight costs of taxation, displacement and substitution effects are taken into account. Further, although the opportunity to receive job search assistance appears to be the most cost-effective service in the sense that it has the highest IRR, more expensive WE and training programs result in larger absolute earnings gains.

Because of substantial treatment non-participation and control group substitution, the impact of these services on those who actually received them is generally larger than indicated by the experimental estimates reported in Table 22. The exceptions are the NSW and AFDC Homemaker-Health Care demonstrations. As explained in Section 5, the NSW provided relatively longterm WE. The AFDC Homemaker Demonstrations trained economically disadvantaged women to provide in-home care to the disabled and the elderly (Bell and Reesman, 1987). Participation rates in these relatively expensive treatments were high and similar services were generally unavailable to the controls (Masters and Maynard, 1981, p. 148; Bell and Reesman, 1987, p. 14). Therefore, in these two studies the experimental impacts can reasonably be interpreted as approximating the impact of the "treatment on treated."

As suggested by the number of studies surveyed in Table 22, there have been fewer experimental evaluations of the impacts of employment and training programs for adult men and especially for youths. As a result, the evidence based on social experiments is more fragmented. Nevertheless, the evidence suggests that programs that offer training can raise the earnings of economically disadvantaged adult males, but programs that focus on

[96] In keeping with the emphasis of US policy on reducing reliance on social assistance, most social experiments have tested the impact of employment and training services on individuals who were applying for or receiving social assistance or welfare (AFDC). The number of these experiments proliferated during the 1980s after the federal government authorized states to operate as demonstration projects community work experience programs (CWEP) for their welfare population. In several states, officials implemented an experimental design in a few welfare offices by mandating that only a random sample of the eligible population participate in JSA, CWEP, or other employment related activities (Goldman et al., 1986). Because the vast majority of social assistance recipients are single female household heads, this has meant that most of the experimental evidence relates to economically disadvantaged adult women. These experimental results were influential in shaping US welfare policy during the late 1980s (Greenberg and Wiseman, 1992).

JSA or WE appear to be ineffective or sometimes worse. Earnings impacts of the San Diego CWEP program, the Baltimore Options program, and the NSW Demonstration were small or negative for disadvantaged adult men. By contrast, the impacts reported in programs that offered training opportunities, San Diego-SWIM program, GAIN, and the NJS, were larger and statistically significant. In particular, the NJS found that economically disadvantaged adult men experienced earnings gains similar to those achieved by adult women (Orr et al., 1994, p. 82).

The evidence from experimental evaluations for youths is not encouraging. As shown by the last panel of Table 22, the results suggest that the array of services currently offered do little to raise youth employment and earnings. For example, the prolonged WE provided to disadvantaged high school dropouts in the NSW Demonstration had no effect on their earnings during the 8 years after the treatment was offered (Couch, 1992). Similarly, the JOBSTART demonstration, which provided disadvantaged youths with services similar to those offered by the comprehensive Job Corps program, but without the residential living centers, did not generate significantly higher earnings for the treatments during the 4-year followup period (Cave et al., 1993). Finally, the NJS finds no evidence that youth served by JTPA benefit from its relatively low cost training services. In fact the shortterm point estimates for the males were actually negative.

Another finding highlighted in Table 22 is the correspondence between earnings impacts and employment impacts. In most cases large earnings impacts are accompanied by significant impacts on employment rates. Moreover, in most of these studies analysts measure employment rates at the quarterly level and information on hours of work are unavailable. When such measures are available, hours impacts also can be a significant source of earnings gains (see, e.g., the NSW Demonstration, Hollister et al., 1984). Indeed, there are only two cases in the table for which the long-run earnings impacts are significant, but not the impact on employment rates. This evidence underscores the concern that because access to government employment and training programs raises earnings through higher employment rates, displacement of non-participants may mitigate the net social benefits reported for these treatments in conventional cost-benefit analyses.

The experimental impacts reported in Table 23 indicate that the impact of the opportunity to participate in particular employment and training services varies substantially among demographic groups. The WE services provided in the NSW demonstration were effective for adult women, but not youths; the WE provided in the San Diego CWEP demonstration was more effective for female welfare applicants than for their male counterparts. The JSA and training experiences provided in the San Diego SWIM demonstration also had a larger impact on women than on men.[97] Finally, the NJS reported striking differences between the impact of JTPA services on adults and youths. These results raise the issue of the importance of impact heterogeneity in this literature.

[97] These differences in experimental impacts are not the result of differing participation rates in the programs by women and men. In the San Diego SWIM Demonstration participation rates in programs services among female (i.e., AFDC-FG) and male (AFDC-U) treatments were nearly the same. Male controls were less likely than female controls to obtain the same services elsewhere (see Freidlander and Hamilton (1993, p. 22, Table 3.1).

Table 22
Impacts from US social experiments evaluating employment and training programs (difference between the treatments' and controls' mean employment rates and earnings in 1997 US dollars)[a]

Demographic group/ services tested/study	Average net costs[b]	Impacts on employment rates and earnings			
		Employment rate last quarter[c]	Earnings[d] Year 1/2	Earnings Year 3/4/5	% of control earnings
A. Economically disadvantaged females					
Job search assistance					
Arkansas WORK	244	6.2*	339*	487*	31
Louisville (WIN-1)	206	5.3*	425*	643*	18
Cook County, IL	231	1.2	12	NA	1
Louisville (WIN-2)	340	14.2*	679*	NA	43
San Diego – CWEP	891	−0.7	402*	NA	8
Food Stamp E & T	180	−2.5	−90	NA	−3
Minnesota – MFIP[e]	NA	14.5*	921*	NA	30
Job search assistance and work experience					
West Virginia	388	−1.0	25	NA	4
Virginia ES	631	4.6*	106	387*	11
San Diego – CWEP	690	3.8*	1120*	NA	23
Baltimore Options	1407	0.4	231	764*	17
Job search assistance and CT or OJT services					
Maine TOPS	2972	1.1	433*	1720*	36
San Diego SWIM	964	0.3	509*	180	15
New Jersey	1165	NA	874*	NA	14
GAIN (JOBS):	3757	5.9[f]	339*	740*	25
Alameda (Oakland)	6036	6.0*	266	901*	37
Los Angeles	6356	1.9	−5	178	9
Riverside	1753	7.5*	1173*	1176*	40
San Diego	2099	2.7*	445*	830*	23
MFSP San Jose (CET)	5132	8.6*	1470*	NA	25
MFSP Other Sites	4525	1.2	400	NA	6

Table 22 (*continued*)

Demographic group/ services tested/study	Average net costs[b]	Impacts on employment rates and earnings			
		Employment rate last quarter[c]	Earnings[d] Year 1/2	Earnings Year 3/4/5	% of control earnings
Florida PI (JOBS)	1339	0.4	93	NA	3
Work experience and training					
National Supported Work	8614	7.1	657	1062	43
AFDC Homemaker	8371	NA	2135*	NA	NA
NJS (JTPA)	1028	NA	691*	441*	7
Recommended for CT	1690	NA	359	NA	NA
Recommended for OJT	643	NA	747*	NA	NA
B. Economically disadvantaged males					
Job search assistance					
San Diego – CWEP	931	0	−325	NA	5
Job search assistance and work experience					
San Diego – CWEP	597	−1.2	−461*	NA	−8
West Virginia	210	NA	−234	NA	−6
Baltimore	1014	NA	−2564	NA	−27
Job search assistance and CT or OJT services					
San Diego – SWIM	747	2.1	704*	−305	−5
GAIN (JOBS):	3202	4.5[f]	489*	413*	11
Alameda (Oakland)	NA	7.9	69	615	18
Los Angeles	4885	9.9*	330*	1033*	22
Riverside	2361	3.8*	970*	389	10
San Diego	2251	−1.3	308	−300	−6
Work experience and training					
NSW-Ex-Offenders	10797	−0.9	100	NA	4

NSW-Ex-Addicts	12150	17.2*	86	706[g]	32
NJS (JTPA)	660	NA	668*	357	7
Recommended for CT	900	NA	895	NA	NA
Recommended for OJT	753	NA	1249*	NA	NA
C. Economically disadvantaged youths[h]					
Work experience and training					
National Supported Work	9314	0.3	−79	−79	−4
JOBSTART	6403	−0.9	−721	523	8
NJS (JTPA)					
Females	1116	NA	133	246	4
Males[i]	1731	NA	−553	852	11

[a] Sources: LaLonde (1995, p. 159); Greenberg and Wiseman (1992, pp. 52–53, 56); Gueron (1990, pp. 92–93); Goldman et al. (1986, pp. 54, 102, 241); Friedlander and Hamilton (1993, pp. 57, 117–118, 133–134); Friedlander et al. (1995, p. xx); Riccio et al. (1994, pp. 254, 267, 316–328, 350–361); Hollister et al. (1984, pp. 148, 181, Table 6.2); Orr et al. (1994, pp. 64, 65, 82, 104, 131, 151, 162, 165, 166); US General Accounting Office (1996, pp. 20–21, Tables ii.1, ii.2, ii.3, and ii.4); Puma and Burstein (1994, pp. 322–23, 325); Knox et al. (1997, Table 5.3). An asterisk indicates that the impact is statistically significant at the 10% level.

[b] Average net costs are the incremental costs of providing services to the members of the treatment group.

[c] "Employment rate last quarter" refers to the difference between treatments' and controls' employment rates during the last quarter of the followup period for which data was available.

[d] The earnings impacts are annual (or annualized) differences between the treatments' and controls' mean earnings during the first or second year (Year 1/2) and during the third, fourth, or fifth year (Year 3/4/5).

[e] Figures are for longterm welfare recipients only. Two other components of this program included both threat of sanctions and financial incentives for welfare recipients to find work.

[f] Measure of ever employed during the last year of followup instead of the last quarter.

[g] Standard error not available.

[h] These studies also examined the impacts on arrests. In the NSW demonstration, the percentage of treatments ever arrested during the study's first 27 months was 8.8 percentage points (22%) less than the controls (Maynard, 1980, p. 138, Table VI.5); In the JOBSTART demonstration the number of treatments' arrested was 2.6 percentage points (21%) lower than the number of controls during the first year of the study. During the 4-year study the percentage of arrests among both experimental groups was the same (Cave et al., 1993, p. 195, Table 6.7); In the NJS, the percentage of treatments arrested was higher than for the controls; the percentage of males with no prior arrests (since age 16) before the study who were subsequently arrested was 7.1 percentage points (38%) larger for the treatments than for the controls (Orr et al., 1994, p. 117, Exhibit 4.22).

[i] Sample of male non-arrestees.

Just as this impact heterogeneity is found among different demographic groups, it also is often found among different sites in the same study. When experimental impact estimates for the same program are available for different sites, it is common to find that the impacts vary among sites. For example, as shown by Table 23, the results from the GAIN program and Minority Female Single Parent Demonstration (MFSP) reveal substantial variation in impacts among sites. Similar variation in experimental impacts also is reported among the 10 sites in the NSW Demonstration and the 16 sites in the NJS (see Maynard, 1980, p. 83; Masters and Maynard, 1981, p. 85; Heckman and Smith, 1998b). At the very least, this evidence of heterogeneity in impacts among sites raises the question of the external validity of these evaluations, i.e., whether their results can be extended to other settings. For policy purposes it is important to know whether the differences in site impacts arise from differences in the skills of program operators and trainers, program organization, or the characteristics of those who are served.

The experimental evidence can shed some light on how heterogenous the impacts are among those served by these programs. An important question in this regard is whether government training programs generate different returns for participants depending on their observed and (to the econometrican) unobserved skills. If returns are smaller for the least skilled, then policy makers would be faced with the difficult question of whether to reallocate expenditures toward less "needy" participants. In 1981, US policy makers in fact made the opposite decision when they directed that employment and training expenditures be targeted to a more economically disadvantaged population (Barnow, 1987). An important policy question is whether this decision improved or worsened the returns from these social programs.

Neither the experimental nor the non-experimental evidence provides a clear answer to the question of whether the impacts of these programs vary with participants' skills. But the experimental evidence does suggest that the least able participants among the low-skilled populations served by these programs benefit the least from them, especially when the programs provide CT and OJT opportunities. To illustrate these points, Table 23 presents the experimental impacts by the prior skills of participants for several social experiments. The measures of skill differ among studies, but as indicated by the controls' earnings during the followup period, these differing measures of skill correctly identify individuals likely to perform poorly in the labor market. In the GAIN and NJS studies more skilled persons benefited more from access to the program's services than did less skilled persons. However, as the table demonstrates, in some programs, such as the NSW and the San Diego CWEP Demonstrations, the least skilled experienced larger gains. Significantly, these programs provided treatments with WE. As explained in Section 2, the purpose of this service is to provide a job experience to individuals with poor employment histories so that they can develop acceptable "work habits." By design, therefore, it might be expected that this service would provide greater benefit to less skilled participants than to more skilled participants who already possess such skills.

Table 23
Experimental impacts of employment and training programs on earnings by prior skills of participants (impacts in nominal US dollars)[a]

Evaluation/total followup period in years/skill measure	Controls' earnings	Impact on earnings[b]	Percentage impact
A. *Economically disadvantaged female household heads*			
NSW/2.25 years			
9–11 years of school	324[c]	181*	52
HS Graduate	633[c]	72	11
San Diego CWEP/1.5 years			
Not employed during prior year	1474	1066*	72
Employed during prior year	4640	347	7
San Diego SWIM/5 years			
HS Drop-out	8783	1654	19
HS Graduate	18135	2405*	13
Florida Project Independence/2 years			
Never employed during prior 36 months	2117	318	15
HS Drop-out and worked 12/36 months	2904	209	7
HS Graduate and worked 12/36 months	6538	314	5
California GAIN/3 years			
a. Alameda Co. (Oakland):			
Assessed to need basic education	3826	610	16
Does not require basic education	8142	2947*	36
b. Los Angeles:			
Assessed to need basic education	3809	107	3
Does not require basic education	8142	1147*	14
c. Riverside:			
Assessed to need basic education	4408	2595*	59
Does not require basic education	9206	3950*	43

Table 23 (*continued*)

Evaluation/total followup period in years/skill measure	Controls' earnings	Impact on earnings[b]	Percentage impact
d. San Diego:			
Assessed to need basic education	5837	572	10
Does not require basic education	11026	3040*	28
Minority Female Single Parent			
Demonstrations/1 year			
a. Atlanta, Georgia - AUL:			
HS Drop-out	3967	576	15
HS Graduate	5948	−280	−5
b. San Jose, California - CET:			
HS Drop-out	4656	1068*	23
HS Graduate	5364	1368*	26
c. Providence, Rhode Island - OIC:			
HS Drop-out	3272	408	13
HS Graduate	4608	72	2
National JTPA Study/2.5 years			
HS Drop-out	9379	878	9
HS Graduate	13484	1152*	8
Received welfare for >2 years	8056	2255*	28
Never received welfare	14513	563	4
Never worked	6887	788	11
Earned <4 in last job	10979	943	9
Earned >4 in last job	14528	1626*	11
B. *Economically disadvantaged male household heads*			
NSW-Ex-addicts/3 years			
9–11 years of school	442[c]	142	32
HS Graduate	458[c]	320*	70
NSW-Ex-offenders/3 years			
9–11 years of school	596[c]	95	16
HS Graduate	622[c]	126	20
San Diego CWEP - JSA only/1.5 years			
Received welfare for >2 years	6911	1187	17

Table 23 (*continued*)

Evaluation/total followup period in years/skill measure	Controls' earnings	Impact on earnings[b]	Percentage impact
Never received welfare	7487	−364	−5
San Diego CWEP - JSA/WE/1.5 years			
Received welfare for >2 years	5724	1398	24
Never received welfare	7852	−280	−4
San Diego SWIM/5 years			
HS Drop-out	19329	−679	−4
HS Graduate	24645	3041	12
California GAIN/3 years			
a. Riverside:			
Assessed to need basic education	9398	555	6
Does not require basic education	11274	3461*	31
b. San Diego:			
Assessed to need basic education	5837	−515	−5
Does not require basic education	11026	1453	10
National JTPA Study/2.5 years			
HS Drop-out	14520	1353	9
HS Graduate	20018	918	4
Never worked	14368	−2104	−15
Earned <4 in last job	14268	245	2
Earned >4 in last job	19353	1647*	9
C. *Economically disadvantaged male youths*			
JOBSTART/4 years			
Not employed during prior year	20164	−1893	−9
Employed during prior year	24729	707	3
Arrested since age 16	20344	1553*	8
Not Arrested since age 16	23183	−921	−4
National JTPA Study/2.5 years (non-arrestees)			
HS Drop-out	14394	−1064	9

Table 23 (*continued*)

Evaluation/total followup period in years/skill measure	Controls' earnings	Impact on earnings[b]	Percentage impact
HS Graduate	19605	−484	4
Never worked	11052	587	5
Earned <4 in last job	16143	−1198	−7
Earned >4 in last job	19056	−1727	−9

[a] Sources: NSW: Masters and Maynard (1981, pp. 89–90); Hollister et al. (1984, pp. 154, 183); San Diego CWEP: Goldman et al. (1986, pp. 92, 126); San Diego SWIM: Friedlander and Hamilton (1993, pp. xxix and xxxi); Florida Project Independence: Kemple et al. (1995, p. 136); California GAIN: Riccio et al. (1994, pp. 137–138, 217–218); Minority Female Single Parent Demonstration: Rangarajan et al. (1992, Volume IV, pp. 37–41); National JTPA Study: Orr et al. (1994, pp. 135–137, 154); JOBSTART: Cave et al. (1993, pp. 156–163). HS, high school. An asterisk indicates that the impact is significant at the 10% level.

[b] Earnings impacts are the difference between treatments' and controls' nominal earnings during the entire followup period given in years next to the name of the program.

[c] Subgroup impacts in the NSW studies in Masters and Maynard (1981) and Hollister et al. (1984) are reported in terms of monthly hours. The figures in the table refer to the period during the last 9 months followed in the study multiplied by 9.

10.4. The findings from non-experimental evaluations of US programs

The experimental evaluations provide evidence that the opportunity to participate in employment and training programs (i) can improve the employment prospects of low skilled persons, and (ii) has markedly varying impacts on different demographic and skill groups. Non-experimental evaluations more often estimate the treatment on the treated parameter although partial participation and dropping out are an important part of ongoing programs as well (Heckman et al., 1998f). Patterns have emerged from these studies that are consistent with and reinforce the findings from the experimental literature.

These patterns exist despite the controversy about the sensitivity in non-experimental estimates and its implications for policy analysis, suggesting that the problems raised by the proponents of the experimental method may be exaggerated. As discussed earlier, the most striking result of non-experimental evaluations of US employment and training programs is the variability in the estimated impacts of training. Not only do the effects vary among different cohorts, but even when program evaluators assess the same cohort, they often arrive at substantially different estimates of the training effect. This sensitivity is one of the most important lessons from this literature and, as we discuss below, it is a lesson that emerges to some extent from the European experience as well. A dramatic illustration of this assertion is the evaluation of the US CETA program. As shown by Table 24, the impact estimates from six evaluations of the 1976 CETA cohort range from −$1553 to $1638 for male participants and from $24 to $2669 for female participants. Not surprisingly, one group of evaluators involved in these studies concluded that

Table 24
The impact of US Federal Government employment and training programs on participants' earnings (increase in post-program annual earnings in 1997 US dollars)[a]

Study	Training cohort[b]	Men[c] (whites/minorities)	Women (whites/minorities)
A. *Non-experimental estimates for economically disadvantaged adult participants*			
Ashenfelter (1978)	1964 MDTA	910/631	2111/1868
Kiefer (1979)	1969 MDTA	−2026/−2244	1905/2621
Gay and Borus (1980)	1969–1972 MDTA	152/161	1373/377
Cooley et al. (1979)	1969–1971 MDTA	1395	2038
Westat (1984)	1976 CETA	−12/−255	983/801
Bassi (1983)	1976 CETA	61/−1055	1286/2669
Dickinson et al. (1986)	1976 CETA	−1553	24
Geraci (1984)	1976 CETA	0	2026
Bloom/McLaughlin (1982)	1976 CETA	364	1844
Ashenfelter/Card (1985)	1976 CETA	1638	2220
Dickinson et al. (1986)	1/76–6/76 CETA	−1031	546
Westat (1984)	1977 CETA	1128/1480	1201/1711
Bassi et al. (1984)	Welfare		
	1977 CETA	1419/−231	2014/1529
Bassi et al. (1984)	Non-welfare		
	1977 CETA	170/546	1650/1783
B. *Non-experimental estimates for displaced workers*			
Bloom (1990)[d]	1984–1985 JTPA, Texas	973	1659
Decker and Corson (1995)	2/88 - 7/88 TAA	−1000	NA
	2/89 - 7/89 TAA	1713	NA
C. *Non-experimental estimates for economically disadvantaged youth participants*			
Cooley et al. (1979)	1969–1971 MDTA	1492	728
Gay and Borus (1980)	1969–1972 Job Corps	−261/180	−1555/−394
Mallar et al. (1982)	1977 Job Corps	2354/2621	NA
Dickinson et al. (1986)	1976 CETA	−1347	449
Bryant and Rupp (1987)	1976 CETA-WE	73(combined)	
Bryant and Rupp (1987)	1976 CETA-WE	1274(combined)	
Bassi et al. (1984)	1977 CETA	−1225/−1614	97/315

[a] Sources: LaLonde (1995, p. 157, Table 1); Barnow (1987, pp. 182–185); Ashenfelter (1978, Tables 4 and 6); Bloom (1990, p. 141, Table 7.6); Kiefer (1979, Table 6.1); Cooley et al. (1979, Table 2); Bassi (1983, Tables 4.3, 4.7, 4.8, and 4.9); Ashenfelter and Card (1985, pp. 658–659); Mallar (1978, Table 1).

[b] MDTA refers to programs funded under the Manpower Development and Training Act, 1962; CETA refers to programs funded under the Comprehensive Employment and Training Act, 1973; JTPA refers to programs funded under the Job Training Partnership Act, 1982; TAA refers to programs funded as part of the Trade Adjustment Assistance Program.

[c] The sets of estimates for each sex refer to the training effect for whites and minorities, respectively.

[d] The Bloom (1990) study was an experimental evaluation. The estimates in the table adjust for non-participation of treatment group members as described in Bloom (1984).

> [al]though these evaluations have all been based on the same datasets, they have produced an extremely wide range of estimated program impacts. In fact, depending on the particular study chosen, one could conclude that CETA programs were quite effective in improving the post-program earnings of participants or, alternatively, that CETA programs reduced the post-program earnings of participants relative to comparable non-participants (Dickinson et al., 1987, pp. 452–453).

Further, different studies of the impact of specific CETA employment and training services exhibit the same variability as the overall program estimates presented in Table 24 (see, e.g., Barnow, 1987, pp. 182–183, Table 3). Five of the six studies summarized in that table also examine the impacts of classroom instruction, on-the-job training, work experience, and public service employment on the earnings of 1976 CETA participants. For example, the estimated effects of OJT for white women in this cohort range from −$295 to $2310 per year. The range of estimates for WE is even larger. Negative training effects are common, but so are large positive impacts.

As discussed in detail in Section 8.4, an important factor contributing to the variability in these non-experimental estimates are differences among analysts' methods of matching. We noted that decisions to match on pre-program earnings at different times substantially affect the estimates. As noted in our discussion on the fallacy of alignment, the problem that arises in these studies is that substantial bias may result when evaluators create comparison groups by matching on serially correlated pre-program outcomes. Matching on such variables alters the properties of the unobservables in the comparison sample in ways that do not guarantee that it will mimic the unobservables of trainees during the post-training period. The bias induced by this practice in the CETA studies can account for their sharply different estimates. Nevertheless, when the estimates from studies most susceptible to this practice are eliminated, the qualitative evidence from the CETA studies is consistent with the experimental evidence from the NJS.

A practical implication of the sensitivity of impact estimates to alternative econometric methods, both experimental and non-experimental, is that cost-benefit analyses of active labor market policies are very fragile. To see the implications of this sensitivity for cost-benefit analyses, consider the following example. Suppose two evaluations of the same program each report that the impacts persist for exactly 8 years. However, the annual impact reported by the first study is $300 per year, while the impact reported by the second study is $700. Assume training costs $2000 per participant. As shown by Table 24, the range of these impacts is consistent with those in the literature. As discussed in Section 2, these costs are typical of government programs. The first evaluation implies that the internal rate of return of the program is 5%, while the second evaluation implies that it is 30%. Readers persuaded by the analysis in the first evaluation would conclude that the program constituted a marginal social investment, whereas those persuaded by the second evaluation would conclude that the program was very productive. This example underscores the importance for policy making of the underlying econometric methodology used

in program evaluations. Modest differences in estimated impacts can have dramatic effects on calculations of the net social benefit of government programs.

Despite the well-documented sensitivity of non-experimental estimates, certain patterns emerge from the non-experimental literature. Government employment and training programs raise the earnings of economically disadvantaged adult women. As shown by Table 24, the estimated impacts are all positive, and many are large relative to the incomes of this population. Further, these impacts are often substantial compared to the costs of these programs which we described in Section 2. Significantly, these results are consistent with the findings in the experimental literature for adult women. In other words, the experimental evaluations, which mostly came after the non-experimental evaluations in time, have led to the same qualitative policy conclusions.

Turning to the impacts for adult males, we observe that they are often smaller and less consistently positive than the impacts for adult women. Accordingly, these estimates suggest that the internal rates of return from these programs are likely lower for males. To illustrate this point consider Ashenfelter's (1978) study of the 1964 MDTA cohort. He reported that CT raised minority males' earnings by \$631 and minority females' earnings by \$1868. The training cost \$8600 (Ashenfelter, 1978, p. 56). If these estimated impacts persisted for the remainder of trainees' working lives, the IRR to training would only be 6% for men, but 22% for women. Because these direct costs include a stipend paid to the trainees, these calculations understate the true IRR. However, they do suggest that these programs constitute a very productive social investment when targeted toward adult women.

As indicated by our discussion of cost-benefit analyses of government programs, these calculations are only suggestive. Many additional considerations besides the earnings impacts affect the IRR of these programs and whether the net benefits are larger when servicing one demographic group compared to another. An important consideration in this regard is the length of the followup period used in the analysis. Ashenfelter's study is relatively unusual in that it followed participants for 5 years after they left the program. By contrast in the CETA studies the followup period usually was less than 2 years. Accordingly, estimates of the IRR from these programs depend crucially on how far into the future analysts project positive shortterm impacts. A second consideration in these IRR calculations is that the foregone earnings cost of participating in training is ignored. These costs are usually larger for adults males. As a result, the gap between the IRR for US programs targeted toward males and females is probably larger than is suggested by the foregoing calculations.

Another factor that may distort simple IRR calculations based on earnings impacts and measures of the average direct cost of training arises because program administrators tend to assign males and females to different services. In the US, males are much more likely to be assigned to receive OJT, which is a less costly service, whereas female participants are more likely to be assigned to receive CT (National Commission for Employment Policy, 1987; Sandell and Rupp, 1988). This practice explains why in Table 20 the internal rates of return estimated for the male participants in the NJS were larger than for females, even

though the earnings impacts shown in Table 22 for the two groups were similar. The males in the NJS were more often assigned to the OJT treatment stream, so the direct costs of servicing them were lower. In the absence of separate measures of the direct costs of these services for males and females, calculations of the IRR of these programs understate the gains from servicing males.

Turning to the non-experimental evaluations of programs for youths, we find that their evidence is also consistent with the results in the experimental literature. The estimated impacts usually are close to zero or even negative. Only one evaluation, that of the US Job Corps program by Mallar et al. (1982), reported substantial positive impacts for youths. However, the earnings impacts during the 4-year followup period are far from sufficient to cover the cost of the program. The modest internal rates of return that have been estimated for this program result from the extrapolation of earnings impacts into the future and from reductions in criminal activity (LaLonde, 1995, p. 164, Table 3). Significantly, these impacts on crime are based on fragile estimates of lower arrest rates for murder (Donohue and Siegelman, 1998). In addition, the comparison group used in this study consisted of non-participants similar to the participants in terms of observable characteristics but drawn from different local labor markets. As explained in Section 8.2, there is now substantial evidence that this approach yields biased estimates of the impact of training. As a result we believe that neither the experimental or non-experimental literatures provide much evidence that employment and training programs improve US youths' labor market prospects.

Over the years both experimental and non-experimental evaluations of government training programs have focused largely on the economically disadvantaged rather than on displaced workers. This focus is in keeping with the emphasis of US employment and training policy on reducing the reliance of low-income persons on various forms of social assistance. Although some of the adult participants in the MDTA and CETA programs would be classified as displaced under the current policy, there have been no separate evaluations of training for displaced workers under these programs that are comparable to those surveyed in Table 24.

As a result, much less is known about the impact of employment and training programs on the earnings of displaced workers. Much of our understanding of how training affects this more advantaged group comes from several demonstrations conducted during the 1980s (Leigh, 1990) as well as from an evaluation of a special program for persons determined to have been displaced by competition from foreign producers (Corson et al., 1993) (see Table 24). Like the MDTA and CETA evaluations, the non-experimental evaluations of these demonstrations find considerable variability in the impact of these training services on different cohorts of displaced workers. But two substantive findings seem clear. First, as is the case for economically disadvantage adults, JSA also is a cost-effective service for displaced workers (Bloom, 1990; Corson et al., 1993). Participants receiving this service have higher earnings because they find jobs sooner than similarly skilled non-participants. Second, participants who have the opportunity to receive CT or OJT derive only modest or no additional benefit from these services.

Given the different objectives of government programs, it also is important to understand how training affects the separate components of earnings, such as employment rates, part-time/full-time status, and hourly wages. A shortcoming of US non-experimental evaluations is that the outcome studied has almost always been annual or quarterly earnings. The CETA and MDTA studies surveyed in Table 24 use annual administrative earnings. These data contain no measures of hours or wages. Further, the employment measure is relatively crude; it reports whether an individual worked in a "covered" job for pay during the year (Card and Sullivan, 1988). Finally, information on the duration of employment or unemployment spells is unavailable. Consequently, by contrast to evaluations of European programs, little is known from non-experimental evaluations of ongoing programs about their impact on employment rates, transition rates out of unemployment or wages. This lack of information makes it difficult to determine whether training raises worker productivity or leads to more stable employment. Much of our knowledge on how US programs affect such outcomes comes from non-experimental evaluations using data from social experiments (see, e.g., Ham and LaLonde, 1996; Eberwein et al., 1997).

10.5. The findings from European evaluations

The European training evaluations are distinct from the US evaluations in several ways. First, they began later in time and only recently has the number become significant. By contrast, the output of such evaluations done by US academics slowed starting in the mid-1980s, although many evaluations continue to be performed by social science consulting firms. This difference in timing results partly from the timing of expanded expenditures on these programs.

Second, European evaluations, particularly those performed outside of the Nordic countries, usually do not use the longitudinal methods commonly used in academic evaluations in the US. Instead, the underlying models are cross-sectional in nature, and control for biases resulting from individual self-selection using parametric methods discussed in Section 7.4. When these evaluations report separate estimates of the impact of training, including and excluding controls for self-selection into training, the estimates controlling for selection usually yield similar or larger estimated impacts than those produced without such controls. As shown by Table 25, this result is seen in evaluations in Austria, Ireland, Norway, Sweden, and the UK. Several authors have noted this finding and have concluded that cross-sectional estimators that fail to account for self-selection into training likely understate the impact of European training programs.

The studies in Sweden and Denmark are generally distinct from other European studies because of their use of longitudinal data and corresponding econometric methods. A factor accounting for this difference is the availability of high quality earnings data from the national "registers." This source of administrative data can yield very large datasets with relatively long panels. For example, the sample used by Westergard-Nielsen (1993) contained more than 30,000 observations covering an 8-year period. This large sample was undoubtedly important in his being able to precisely estimate wage impacts on the

Table 25
Estimated impacts of Canadian and European job training programs (impacts on employment and earnings outcomes)

Country/study author(s)	Outcome studied[a]	Estimator[b]	Program/cohort[c]	Impacts[c]
Austria				
Zweimuller and Winter-Ebmer (1996)	Unemployment risk	Probit selection	ARB-CT: males,1986	0 −0.40*
Canada				
Park et al. (1993)	Annual earnings	Diff-in-diff	Canadian jobs strategy	
			CT, 1988	0.09
			CT, 1989	−0.20
			Job Entry, 1988	0.24
			Job Entry, 1989	0.18
			OJT1, 1988	0.06
			OJT1, 1989	−0.11
			OJT2, 1988	0.26*
			OJT2, 1989	−0.01
Denmark				
Jensen et al. (1993)/ Westergard-Neilsen (1993)	Unemployment rate	Panel	AMU: Adults, 1976/1988	0
			Poor recent job history	(−)*
	Log hourly wages		Males	0.01*
			Skilled males	0.01
			Unskilled males	0.01*
			Females	0.00
France				
Thierry and Sollogoub (1995)	Employment hazard	MLE	YTP: OJT	(−)*
Bonnal et al. (1997)	Unemployment hazard	MLE	YTP: Males < 26, 1986–1988	
			Without diploma:	
			CT	+*
			WE	0
			OJT	+*

			With certificate:	
			CT	+*
			WE	(−)*
			OJT	+*
	Employment hazard	MLE	Without diploma:	
			CT	(−)*
			WE	0
			OJT	(−)*
			With certificate:	
			CT	(−)*
			WE	+*
			OJT	(−)*
Germany				
Kraus et al. (1997)	Non-employment hazard to stable employment	MLE	AFG: 1992–1994	
			CT-males/females	+*
			OJT-males	0
			OJT-females	+*
Lechner, (1996, 1997)	Unemployment rate	Matching	AFG: 1991–1993	0.01
	Monthly earnings			0.04
Ireland				
Breen (1988, 1991)	Employment rate	Probit	AnCO/FAS: Youth, 1981–1982	
			After exit	0.17*
			After 1 year	0.06
		Selection	After exit	0.25*
			After 1 year	0.04
		Probit	WEP/Teamwork: Youth, 1982–1986	
			After exit	0.23*
			After 1 year	0.26*
		Selection	After exit	0.77*
			After 1 year	0.18
O'Connell and McGinnity (1997)	Employment rate	Probit	Youth < 23, 1992	

Table 25 (*continued*)

Country/study author(s)	Outcome studied[a]	Estimator[b]	Program/cohort[c]	Impacts[c]
			CT	0.16*
			OJT	0.21*
			WE	0.00
	Weekly wage	OLS	CT	0.01*
			OJT	0.02*
			WE	0.00
The Netherlands				
Ridder (1986)	Employment hazard	MLE	E & T Progams: 1979–1981	
			>35 years	0
			<35 years, WE	(−)*
			<35 years, OJT/CT	(−)
	Unemployment hazard	MLE	E & T Progams: 1979–1981	
			>35 years	(−)*
			<35 years, WE	(−)
			<35 years, OJT/CT	(−)
de Koning et al. (1991)	Unemployment hazard	Matching/MLE	CVV-CT:	
			Blue collar trades	+*
			Clerical courses	0
de Koning (1993)	Unemployment rate	OLS	VMA/JOB-OJT:	
			Youth < 25 (JOB)	(−)
			Adults (VMA)	(−)*
Norway				
Torp et al. (1993)	Employment rate	Experiment	CT: All treatments,1991	0.03
		Probit	Training completers only	(−)
		Selection		(−)*
Sweden				
Delander (1978)/ Björklund and Regner (1996)	Employment rate	Experiment	ES/Intensified JSA: 1975	
			Job seekers in Eskilstuna	0.13*

	Monthly earnings			0.06
Engstrom et al. (1988)/ Björklund (1993)	Unemployment hazard	MLE	ES: 1983 Displaced workers	0
Björklund (1993, 1994)	Employment rate	OLS	AMS-CT: 1976–1980 16–64 years/unemployed	0.05
		Panel		0.08*
	Log hourly wage	OLS	16–64 years/unemployed	−0.05
		Selection		0.05
		Panel		0.10*
Edin (1988)	Log weekly earnings	Panel	AMS-CT: 1977 Displaced workers	−0.09*
Axelsson (1989)/ Björklund (1993)	Annual earnings (in %)	Panel	AMS-CT: 1981	0.22*
Ackum (1991)	Log hourly wages	OLS	AMS-CT: 1981 Youths < 25	−0.02
		Selection		−0.01
		Panel		−0.05
Andersson (1993)	Annual earnings (in %)	Match/OLS	AMS-CT: 1989–1990 1989 cohort	−0.05*
			1990 cohort	−0.15*
		Match/panel	1989 cohort	−0.02
			1990 cohort	−0.13*
Regner (1996)	Annual earnings (in %)	Match/panel	AMS-CT: 1990 1989 male cohort	0.10
			1990 male cohort	−0.10
			1989 youth cohort	−0.06
			1990 youth cohort	−0.26*
Harkman et al. (1996)	Employment rate	Match/probit	AMS-CT: All, 1993	−0.01
		Match/selection		0.09
	Log hourly wages	Match/OLS		0.02
		Match/Selection		0.05
United Kingdom				
Main and Raffe (1983)	Employment rate	Probit	YOP-Scotland: 1978 Males	0.06
			Females	0.14*

Table 25 (*continued*)

Country/study author(s)	Outcome studied[a]	Estimator[b]	Program/cohort[c]	Impacts[c]
Main (1985)	Employment rate	Probit	YOP-Scotland: 1980	
			Males:	
			All	0.04*
			Disadvantaged[d]	0.03*
			Females:	
			All	0.08*
			Disadvantaged[d]	0.07*
Whitfield and Bourlakis (1991)	Employment rate	Probit	YTS-I:	0.04*
	Log hourly wage	Selection		−0.03
Main and Shelly (1990)	Employment rate	Probit	YTS-I-Scotland: All	0.15*
			Disadvantaged[d]	0.11*
	Log hourly wage	Selection	YTS-I-Scotland:	
			Advantaged[d]	0.20
			Disadvantaged[d]	0.32
Main (1991)	Employment rate	Probit	YTS-I-Scotland:	
			Advantaged[d]	0.14
			Disadvantaged[d]	0.19*
O'Higgins (1994)	Employment rate	Probit	YTS-I: All	0.08*
			Disadvantaged[d]	0.04*
		Selection	YTS-I: All	0.21*
			Disadvantaged[d]	0.09*
			Females	0.28*
Green et al. (1996)	Log hourly wage	Selection	YTS-II:	
			YTS+ certification	0.19
			YTS, no certification	0.02
			YTS, other (e.g., CT)	0.29
Dolton et al. (1992)	Employment rate	Selection	YTS-II: no prior OJT	(−)
			Prior/current OJT	+
	Log hourly wage	Selection	Males age 16 in 1985–1986:	
			WE/OJT	0.05

			CT/OJT	0
			Females age 16 in 1985–1986:	
			WE/OJT	−0.05
			CT/OJT	−0.03
Dolton et al. (1994a)	Unemployment hazard	MLE	YTS-II:	
			To all jobs:	
			Males	(−)*
			Females	(−)*
			To all jobs net of time in YTS:	
			Males	+*
			Females	+*
			To "stable" jobs:	
			Males	0
			Females	+*
Dolton et al. (1994b)	Log hourly wage	Selection	YTS-II: Disadvantaged[d]	
			Males	0.26
			Females	−0.08
White and Lakey (1992)	Employment rate	Experiment	Restart: 1989	0.04*
Payne et al., (1996)	Employment rate	Matching	ET: Adults, 1991–1992	0.22*
	Log hourly wages	Selection		0
	Employment rate	Matching	EA: Adults, 1991–1992	0.04
	Log hourly wages	Selection		0
Dolton and O'Neill (1996a,b)	Unemployment hazard	MLE	Restart: 1989	
			To all jobs	+*
			To stable jobs	0
			To training schemes	+
			To "signing off" UI	+*

[a] Characterizes the class of estimator used in the analysis. "Probit" refers to a univariate probit or logit model that includes one or more indicator variables for training status. "Selection" refers to a parametric econometric model of selection bias that controls for correlation between unobservables in the outcome and participation equations (e.g., Heckman, 1976, 1979; Lee, 1983). "Diff-in-diff" denotes a difference-in-differences estimate that may control for covariates in a regression. "Panel" refers to estimators that exploit longitudinal data, which include difference-in-differences, fixed effects, or autoregressive estimators. "Experiment" refers to the mean difference between outcomes for experimental treatment and control groups. "MLE" refers to standard maximum likelihood methods used for event history data (e.g., Heckman and Singer, 1984; Lancaster, 1990).

[b] Description of employment and training programs: *AMU*: CT usually provide participants with 2–4 weeks of full-time instruction at government training centers. Courses targeted toward both employed and unemployed. Courses partly financed through a payroll tax on employees. *AMS*: CT (Arbetsmarknad-

sutbildning) provided in training centers to unemployed in Sweden. Classroom training programs usually last less than 17 weeks. *AnCO/FAS*: Classroom training lasting less than 6 months. It is sometimes subcontracted to outside or external providers. *ARB*: (Arbeitmarktverwaltung) CT in Austria. *Canadian Jobs Strategy*: consists of several options. CT refers to the feepayers option that allows unemployed adults who receive benefits to enroll at their own expense in approved full-time CT for a period not exceeding 1 year while they continue to receive benefits. Job search requirement waived while in training. Basic skills training not available under "feepayer" option until after 1991. The "Job Entry" option is designed for out-of-school youths and women who have been out of the labor force for three or more years. Separate components for youth and women designed to ease transition into labor force. "OJT1" refers to the "Job Development" option designed for longterm unemployed. "OJT2" refers to "skill shortages" option designed for the unemployed who where "not job ready", who did not meet criteria for other Canadian programs, and who administrators thought would benefit from the program. OJT may last up to 3 years. *CVV*: Vocational Training Centres provide adult participants with vocational CT in blue collar and clerical occupations. *EA*: UK Employment Action program which provides participants with subsidized employment in non-profit or public sector jobs. *East German AFG*: CT programs subsidized under the Work Support Act (Arbeitsforderungsgesetz). *ES*: (Employment Service) services that can include job search assistance (JSA), career counseling, mobility grants, etc. *ET*: UK Employment Training program which provides participants with CT and some OJT opportunities. *Labor Market Training/CT*: Vocational courses provided by Norwegian Directorate of Labour, educational institutions, and private firms. Average duration of such classes was 18 weeks in 1991, although approximately 40% of persons took part in a class lasting 40 or more weeks. *Restart*: April 1987–present, Employment Service provides counseling to all unemployed after 6, 12, and 24 months of unemployment. Assesses claimant's job search behavior and offers advice. May suspend benefits to claimants who are not available for work, who decline offers of assistance, or who fail to attend scheduled interviews. In the Restart experiment, a random sample of unemployed was not required to attend an interview until its 12th month of unemployment. The "treatments" were required to attend the 6 month interview. *YOP*: Youth Opportunities Program 1978–1983. *YTS-I*: Youth Training Scheme 1983–1986 in England and Wales, unless indicated otherwise. *YTS-II*: Youth Training Scheme 1986–1989 in England and Wales, unless indicated otherwise. *VMA/JOB*: provides longterm unemployed adults (VMA) and youth (JOB) with subsidized jobs from private employers. *WEP/Teamwork*: WEP refers to work experience program that provides temporary subsidized jobs with private employer; Teamwork provides the same in a volunteer or community organization. Trainees may be retained by employer when subsidy ends. *YTP*: Youth Training Programs in France.

[c] Impacts measure the percentage effect of the program on earnings or wages and the percentage point impact on employment or unemployment. Both impacts are expressed in decimal form. Results from evaluations of programs on hazard rates out of employment or unemployment are expressed in terms of the effect on the sign of the impact. An asterisk denotes that the impact is statistically significant at the 5% level.

[d] Disadvantaged refers to participants with poor academic qualifications and who reside in local labor markets with high unemployment rates. Advantaged refers to participants with relatively strong academic qualifications (four or more O grades) among non-college bound youth and who reside in local labor markets with low unemployment rates.

order of 1%. The evaluations based on Swedish data usually use a smaller number of observations because they study a random sample of participants and non-participants from the registers, cover a smaller cohort of participants, are limited to a certain geographic section of the country, or discard many non-participants when they create a "matched" comparison group. In Sweden, the ability to use administrative records to match participants to non-participants from the same labor market likely improves the quality of these evaluations. Such matching was impossible in US evaluations that used large administrative datasets (Ashenfelter, 1978, for MDTA; Barnow, 1987, for CETA). Other studies that make use of administrative data include Zweimuller and Winter-Ebmer (1996) for Austria, Ridder (1986) for the Netherlands, and Bonnal et al. (1997) for France. These latter two studies evaluate the effects of training in the context of event history models of labor force dynamics (Flinn and Heckman, 1982).

Evaluations of employment and training programs in the United Kingdom generally use existing general survey data. For example, the evaluations by Whitfield and Bourlakis (1991) and O'Higgins (1994) use the first cohort of the England and Wales Youth Cohort Study (YCS). This survey was administered in three successive years starting in May 1985 to persons who completed their compulsory education during the 1983–1984 academic year.[98] A factor affecting the quality and precision of the estimates in these studies is attrition from the sample. Among those in the first cohort of the YCS only 40% of the original sample responded to all three "sweeps." Similar attrition is reported in existing survey data used in evaluations of the east German programs (Kraus et al., 1997). These experiences underscore the problem of sample attrition when using survey data that does not arise in administrative data obtained from national registers such as those used in the Danish and Swedish studies.

Despite concerns about attrition and the quality of survey responses, an advantage of these survey data is that they contain a much richer set of baseline characteristics on participants and non-participants than is usually available from administrative data sources. For example the UK data contain detailed information on how well the respondent had done in school, including the number of "O" and "A" levels obtained. In addition, it is possible to obtain from both participants and non-participants detailed information on training provided privately by employers. This type of information has generally not been available to evaluators of US programs (but see Gritz, 1993; Heckman and Roselius, 1994). Moreover, these data have enabled evaluators in the UK to look for evidence of heterogeneity in training effects using a wide array of variables that usually has not been available to US evaluators. Finally, these datasets contain local labor market identifiers and as a result several studies have accounted for this variable in their analyses.

A third difference between European and US evaluations is the concentration of these studies on youths. The studies for Austria, Denmark, and Sweden usually include both

[98] The studies by Main and Shelly are based on the comparable Scottish Young Peoples Surveys. The study by Dolton et al. (1992) uses the third cohort of the YCS, which contains individuals who completed their compulsory education during the 1985–1986 academic year.

adults and youths, but nearly all the other studies summarized in Table 25 focus on youth or very young adults. This difference in emphasis reflects policy concerns in Europe about youth unemployment, as compared to policy concerns in the US about the economically disadvantaged of all ages. An advantage of the youth focus of European evaluations is that they provide an opportunity to assess the impact of public sector training interventions on a much less disadvantaged population of youths than is possible in US evaluations. However, surveying the results in the table provides no consistent indication whether these interventions are more or less effective for youth, nor whether more disadvantaged youth benefit more or less from these programs.

A fourth difference between European and US evaluations is that European evaluations place much greater emphasis on measuring the impact of training on hourly wages. As indicated above, this difference reflects the common use of administrative data in US evaluations and the fact that these data almost never contain measures of wages or hours worked. From the perspective of assessing the impact of active labor market policies on human capital accumulation and worker productivity, the European studies potentially shed more light on these questions than is possible in the US studies.

Turning to the estimated impacts presented in the table, we first observe that of the three social experiments conducted in Europe, two tested the impact of employment services along the lines of JSA offered in the United States. Both studies report results that are consistent with those in the US, namely that despite their low costs, access to these services significantly raises employment rates. In the Swedish experiment, unemployed participants received an average of 7.5 h of additional job search assistance compared to 1.5 h received by the control group. Nine months later, the treatments' employment rate was 13 percentage points higher than that of the controls. In the British Restart experiment, a random sample of individuals who had been unemployed for exactly 6 months were assigned to a control group and excused from receiving the 15–25 min interview and counselling session normally required at that time. By contrast, the treatments risked losing their benefits if they failed to attend the interview or demonstrate that they were available for work. Although they could voluntarily request such an interview, the controls were allowed to wait until the next regularly scheduled interview after their twelfth month of unemployment. After 1 year, those assigned to the control group had employment rates that were 4 percentage points lower than those in the treatment group, and for males this impact persisted for at least 5 years (Dolton and O'Neill, 1996b, 1997; Robinson, 1996).[99] The one non-experimental evaluation of JSA was a study of Swedish displaced workers by Engstrom et al. (1988), who found that these services had no significant impact on employment rates.

Among the evaluations summarized in Table 25, we do not observe any pattern that

[99] The original sample contained 8925 persons of which 582 were assigned to a control group. Of the original sample, 5200 persons completed the first 6 month followup survey, of which 323 were controls. Dolton and O'Neill (1996a) found no evidence that this attrition was correlated with a person's experimental status. Dolton and O'Neill matched these survey responses to administrative data (JUVOS) from the Employment Service.

leads us to conclude that any one active labor market policy consistently yields greater employment impacts than any other. Instead, the European evaluations often reveal large and statistically significant effects of any one of these policies on employment rates. This finding is seen directly in the Irish study of Breen (1991), the Swedish study by Björklund (1989), the UK studies by Main and Raffe (1983), Main (1985, 1991), Main and Shelly (1990), and O'Higgins (1994), and indirectly in the Austrian study by Zweimuller and Winter-Ebmer (1996), the French study by Bonnal et al. (1997), and the Dutch study by Ridder (1986). As shown by the table, the estimated employment impacts exceed 10 percentage points in several of these studies.

At the same time, other studies such as the Danish and Norwegian evaluations, the Swedish study by Harkman et al. (1996), and the UK studies by Dolton et al. (1992, 1994b) report much smaller and sometimes even negative impacts of these programs on employment. Although the variability in the impact estimates among studies is reminiscent of the experience with the US CETA evaluations, it is important to observe that these studies are of different cohorts and in some cases of different programs.

Whereas it is common for European evaluations to report that training has significant impact on employment rates, it is relatively uncommon for them to report the same for log wages. In several studies, the point estimates of the impact of training are extremely large, but they are not statistically significant. The largest statistically significant impact reported in the table is by Björklund (1994) who finds that during the late 1970s labor market training in Sweden may have raised hourly wages by 10%. At the same time, he is careful to observe that this finding is sensitive to the econometric method used in the analysis. Moreover, this finding raises the question posed above in Section 10.2 of whether it is plausible that 17 weeks of CT – the standard in Sweden – could result in such a large impact on a trainee's wages. After all, during this period, the impact of a year of formal schooling as measured by a conventional Mincerian wage equation was as low as 2% (Harkman et al., 1996).

In light of this consideration, the other instance of an evaluation reporting a statistically significant impact of training on log wages is more plausible. The Danish study found that 2–4 weeks of vocational classroom training raised the subsequent hourly wages of unskilled male workers by approximately 1%. The point estimate for skilled males was the same, but it was not statistically significant. The point estimates for females were approximately equal to zero, but also not statistically significantly different from 1%. As indicated above, the reason why this study could estimate these impacts so precisely, especially for the males, is because the authors' sample was extremely large.

Although, many of the point estimates of the impact of training on wages are positive, there also are several studies that find either no or negative effects of training on wages. Besides the Danish study referred to above, Whitfield and Bourlakis (1991) and Dolton et al. (1994a) report similar findings for youth in the UK as do Ackum (1991) and Regner (1996) in Sweden. In Sweden several studies also report that training has either no or negative impacts on monthly earnings.

Accordingly, there is little compelling evidence that European active labor market

policies have had a positive impact on participants' wages. By contrast, we have already observed that the case for positive employment effects from these policies is stronger, although there is as yet no consensus on this question. Even if there were a compelling consensus, the question remains whether these employment impacts correspond to an increase in aggregate output or are offset to some extent by displacement of non-participants (Johnson, 1979). Because of the size of these programs as documented in Section 2, because of the emphasis in many European countries on OJT, and because earnings gains from these programs likely are generated through higher employment rates, cost-benefit analyses based on the impact estimates presented in Table 25 probably overstate the net social benefit derived from active labor market policies in Europe.

11. Conclusions

This chapter has examined the effectiveness of active labor market policies and the methods used to evaluate their effectiveness. When these programs are effective they make economically disadvantaged persons less poor, and modestly increase the probability of employment among the unemployed. But the gains from existing programs are not sufficiently large to lift many out of poverty nor to significantly reduce unemployment rates. Further, because these gains, when they occur, appear to arise from increased employment rates instead of wages, they likely overstate the human capital-enhancing benefits of these policies. In Europe, especially, evidence that these programs also result in the displacement of non-participants indicates that the net social benefits of active labor market policies are substantially smaller than are indicated by the impacts from conventional program evaluations.

The evidence we summarize also suggests that it is unlikely that even a substantial increase in government-funded training services will significantly improve the skills in the work force. As indicated above, this finding should not be surprising, because most of these programs cost only a few thousand dollars or less per participant. Although European programs often are more expensive, these costs include stipends paid to participants which do not represent investments in human capital. To expect such programs to raise participants' subsequent annual earnings by several thousand dollars would imply that these social investments consistently have an extraordinary rate of return. A 10% rate of return is high in this literature. Even granting it, a thousand dollars invested in a poor person would only raise annual earnings by $100 per year. A more realistic view of the returns to public-sector-sponsored training would suggest that this type of impact requires an investment that is more than an order of magnitude greater than what is currently being spent on low income and dislocated workers (Heckman et al., 1993).

A major focus of this chapter has been on the methodological lessons learned from 30 years of evaluation activity in the United States and their relevance for the conduct of future evaluations. For brevity, we have left several important issues for discussion elsewhere. In this chapter, we have focused on identifying mean outcomes and in particular the

mean impact of treatment on the treated. Heckman and Smith (1998a) and Heckman et al. (1997c) discuss conditions for recovering distributions of impacts and present evidence on the empirical importance of heterogeneity in impacts in assessing programs. They demonstrate the value of knowing the distribution of program impacts in evaluating the modern welfare state. Heckman et al. (1999) present evidence from the NJS data that persons act on their idiosyncratic response to training, so that the theoretical possibility that we have discussed in this essay is practically important for empirical work in evaluating programs.

We summarize the methodological lessons discussed in this chapter as follows: First, a major development in the field of program evaluation is recognition of the *multiplicity* of the parameters of interest in evaluating employment and training programs. This multiplicity is a consequence of well-documented heterogeneity in the impact of even a single training program. Recognition of this heterogeneity in response among participants and of the possibility that agents participate in programs, at least in part, on the basis of their idiosyncratic responses to them, fundamentally alters intuitions about, and formal properties of, standard econometric estimators. Different parameters require different identifying assumptions, as we demonstrate in our discussion of the conditions for IV to identify "treatment on the treated" rather than LATE in the presence of response heterogeneity. When responses to treatment are heterogeneous, the case for using fixed effect or instrumental variables methods to estimate the parameters commonly sought in evaluation analysis becomes much weaker. Even the case for social experiments has to be qualified significantly if persons enroll in programs at least in part on the basis of their own idiosyncratic response to training.

A second major lesson that flows in part from the first is that the choice of an evaluation method depends on the question being asked in the evaluation and on the economic model generating participation and outcomes. Because both questions and models vary among programs and economic environments, there is no "method of choice" for conducting evaluations. This conclusion is at odds with segments of the current literature which treat matching or, more commonly, fixed effects methods, difference-in-differences or IV as cure-alls for selection problems. Proper choices among alternative experimental and nonexperimental methods should be dictated by the economics of the problem, their relevance to the data in hand, and the evaluation question being addressed. The nature and range of questions being asked by policy makers and researchers make it impossible for a rigorously justified "consensus" to emerge about the proper choice of an estimator to evaluate a social program that is valid in all contexts. All methods for evaluating social programs are based on identifying assumptions that are difficult to test unless additional data about the unobservables in a given study are collected.

There is no universally correct way to construct the counterfactuals needed to evaluate the training programs of the welfare state. Even social experiments are valid only under special assumptions about behavior. We have discussed the interplay between theory, data, and the questions being addressed in an evaluation and how each affects the choice of an estimator. We also have shown that many widely used evaluation strategies – such as choosing comparison groups to make participant and comparison group preprogram earn-

ings histories as alike as possible – only "work" under certain conditions and under other conditions may produce substantially misleading assessments of the program being evaluated.

A third major lesson is that evidence that different estimators produce different estimates, while disappointing, does not necessarily indicate that non-experimental methods fail to measure the appropriate counterfactual. Different estimators solve the selection problem under different assumptions. Only if there is no selection problem and there is no model misspecification problem would all estimators produce the same estimate, up to sampling variation. Robustness studies that show that all methods produce the same estimate only reveal that there is no selection bias.

A fourth major lesson follows from a reexamination of the evidence and issues raised in LaLonde's (1986) paper and Fraker and Maynard's (1987) paper on evaluating non-experimental evaluations. These papers concluded that "...policymakers should be aware that available non-experimental evaluations of employment and training programs may contain large and unknown biases resulting from specification errors" (LaLonde, 1986, p. 617). Nevertheless, some people interpret this work as having proved that conventional econometric program evaluation and model selection procedures are unreliable and cannot be used to produce valid program evaluations. Advocates of social experiments (e.g., Stormsdorfer et al., 1985) and advocates of the robust bounding and sensitivity analyses we briefly survey in Section 7.8 routinely cite it in defense of their methods.

In this chapter, we have reexamined the inferences from this work by drawing on more recent research of Heckman et al. (1996b, 1997a, 1998b). We find that once certain basic principles of data quality are adhered to, selection bias, rigorously defined, is only a small contributor to the bias from using non-experimental data that LaLonde reports in his paper. A far more important bias arises from comparing non-comparable people.

The sources of non-comparability in his study arise from (i) using different surveys or data sources to measure the outcomes and background characteristics of participants and comparison group members; (ii) using participants and comparison group members from different local labor markets; and (iii) using individuals mismatched on personal characteristics. Comparing comparable people goes a long way toward reducing the bias in non-experimental methods reported by LaLonde. This shifts the emphasis in program evaluation away from specifying econometric methods for selection bias and toward more careful construction and weighting of comparison groups. It suggests that in the future, non-experimental comparison groups should be selected to balance the support of the regressors in the comparison group to make it comparable to that in the treatment group. For matching, classical selection bias estimators and non-parametric difference-in-differences estimators, it suggests making the supports of the probability of selection, $P(X)$, coincide in the treatment and comparison groups. This principle should guide both data collection efforts (where stratified sampling of non-participants may be useful) and the analysis of existing datasets.

We also have shown that no econometric or statistical cure-all fixes the problem of fundamentally bad data. Heckman et al. (1998b) demonstrate that econometric selection

estimators and a non-parametric version of difference-in-differences "work" reasonably well for an averaged version of the treatment on the treated parameter when a good comparison group is available. Even the bias from matching is not large. No non-experimental method is particularly effective when a bad comparison group is all that is available. The solution to the evaluation problem lies in both the method *and* the data. The literature on evaluating job training programs has focused largely on methods and not issues of data, taking a passive approach to data collection.

A fifth lesson is that non-experimental evaluations are not necessarily significantly less expensive than experimental evaluations. The low cost of previous non-experimental evaluations resulted from reliance on existing data sources. The importance of high quality data for constructing comparison groups means that credible non-experimental evaluations are likely to be expensive. Existing general survey data and administrative data, which are inexpensively obtained, often contain either too few participants or non-participants, or contain too little information on demographic characteristics or on labor force dynamics. This information has been shown to be important for conducting better non-experimental evaluations and is usually obtained only by collecting costly new survey data. The high cost of previous social experiments results not from administering randomization, but from data collection, careful documentation of the implementation of the program, analysis, and dissemination of reports. These costs are not unique to social experiments, but arise in any careful program evaluation.

A sixth major lesson that emerges from the recent literature is the advantage of using non-parametric econometric methods for program evaluations. The non-parametric approach instructs analysts to compare comparable people. Systematically applied, the non-parametric approach avoids the use of potentially misleading functional forms in constructing counterfactuals.

A seventh major lesson is a better understanding of the benefits and limitations of social experiments. Under ideal conditions, experiments enable us to bypass the need to carefully specify an econometric model or to determine which variables belong in the model. They offer an easily explained procedure for estimating the impact of social programs. In addition, they provide an important benchmark for learning about non-experimental models. Further, even when the ideal conditions are violated, the experimental design enables analysts to obtain a comparison group whose distribution of characteristics is likely similar to those of individuals in the treatment group. Under less than ideal conditions, analysts have to rely on non-experimental methods to estimate parameters of policy interest, but can do so using a better quality comparison group than they could obtain from existing data sources.

Even under ideal conditions, however, the means that can be constructed from a social experiment either by randomizing out people accepted into the program, or randomizing eligibility, identify only a few of the many parameters that can be defined when responses to treatment are heterogeneous, and which are of practical interest to policy makers and social scientists seeking to evaluate active labor market policies. When analysts estimate an evaluation parameter that is not the direct product of the experiment, they must rely on

the same non-experimental methods discussed in Section 7 (Heckman, 1992; Ham and LaLonde, 1996).

The modern case for social experiments usually seeks to recover only one well-defined parameter. This objective is in contrast to the older case that motivated the Negative Income Tax experiments. The older case sought to conduct experiments to recover estimates of the parameters of well-posed economic models that provide the basis for policy analyses of hypothetical programs different from those evaluated by the experiment producing the estimates. Samples generated under the new model for social experiments produce evidence that does not accumulate in the same way as evidence accumulated under the old model, because there is no common basis for comparing the "treatment effects" from one experiment with those from another. Given the nature of the choice-based, endogenously stratified sampling rules used to produce the data used in recent social experiments, it is difficult to use these data to estimate policy-invariant structural parameters that can be used to evaluate a wide variety of programs never previously implemented. Social experiments produced from randomizing out people who applied and were accepted into the program produce knowledge that does not accumulate within the context of economic models unless elaborate non-experimental methods are used to correct for endogenous stratification.

We also have presented evidence on how experiments work in practice. Nearly all social experiments operate in a less than ideal environment and as a result often produce estimates that are not easily interpreted. They are much less effective in evaluating ongoing programs, as illustrated by our discussion of the National JTPA study, than they are in evaluating a new program never previously put in place and for which there are no good substitutes, such as the National Supported Work Demonstration. We draw on the work of Heckman et al. (1998a), who provide evidence that when persons randomized out of the program can find close substitutes for it, the parameter obtained from an experiment differs substantially from the parameter of interest to program evaluators and policy analysts.

An eighth major lesson is that when programs are implemented on a large scale, they may change the prices and opportunities facing everyone in the population. The micro-economic treatment effect literature ignores the effects of programs on the interactions among agents. A convincing evaluation requires embedding the treatment effect framework in a social setting. Drawing on the research of Heckman et al. (1998e) and Davidson and Woodbury (1993), we demonstrate that displacement and general equilibrium effects may be sizeable. The lessons from the treatment effect literature that ignores social interactions can be quite misleading. The challenge in estimating these general equilibrium effects is the challenge of estimating credible general equilibrium models. However, unless the challenge is met, or the social interactions are documented to be unimportant, the output of micro treatment effect evaluations will provide poor guides to public policy.

We conclude this chapter with our recommendations for conducting evaluations based on our best current knowledge. They are: (1) carefully define the parameter of interest; different parameters require different identifying assumptions; (2) compare comparable people; (3) using better data in modeling participation decisions and labor market

outcomes helps a lot. In particular, it is important to measure outcome variables in the same way for participants and non-participants and to draw the treatment and comparison groups from the same local labor markets. In addition, recent evidence suggests that labor force status dynamics represent an important determinant of participation in job training programs; (4) there is no universally "correct" experimental or non-experimental estimator that applies in all contexts. The overwhelming reliance on IV, fixed effects or difference-in-differences and matching estimators in recent research lacks theoretical and empirical justification. In LaLonde's (1986) study, fixed effect estimators produced the most unstable estimates. Evaluators should use economic theory, the available data and prior information to guide the choice of non-experimental estimators, carefully state the conditions under which counterfactual states are generated, and defend their plausibility; (5) expect different estimators to produce different estimates unless there is no selection problem; (6) use experimental methods when possible in evaluating demonstrations of employment and training strategies whose services are not available elsewhere in the community, but collect enough data to test the identifying assumptions that justify experiments. When an experimental design is used to evaluate an ongoing program, analysts should be prepared to use non-experimental methods to answer many important policy questions; (7) the validity of partial equilibrium, microeconomic approaches needs to be confirmed. The estimates from the micro economic treatment effect literature may be very misleading. A more satisfactory approach accounts for the impact of a policy on the interactions of agents in a market economy.

References

Aakvik, A. (1998), "Estimating the employment effects of education for disabled workers in Norway", Unpublished manuscript (University of Chicago).

Ackum, S. (1991), "Youth unemployment, labour market programs and subsequent earnings", Scandinavian Journal of Economics 93 (4): 531–543.

Amemiya, T. (1985), Advanced econometrics (Harvard University Press, Cambridge, MA).

Anderson, K., R. Burkhauser, J. Raymond and C. Russell (1991), "Mixed signals in the Job Training Partnership Act", Growth and Change 22 (3): 32–48.

Anderson, K., R. Burkhauser and J. Raymond (1993), "The effect of creaming on placement rates under the Job Training Partnership Act", Industrial and Labor Relations Review 46 (4): 613–624.

Andersson, H. (1993), "Choosing among alternative nonexperimental methods for estimating the impact of training: new Swedish evidence", Unpublished manuscript (Swedish Institute for Social Research, Stockholm University).

Andrews, D. and M. Schafgans (1998), "Semiparametric estimation of a sample selection model", Review of Economic Studies 56 (3).

Ashenfelter, O. (1978), "Estimating the effect of training programs on earnings", Review of Economics and Statistics 6 (1): 47–57.

Ashenfelter, O. (1979), "Estimating the effect of training programs on earnings with longitudinal data", in: F. Bloch, ed., Evaluating manpower training programs (JAI Press, Greenwich, CT) pp. 97–117.

Ashenfelter, O. and D. Card (1985), "Using the longitudinal structure of earnings to estimate the effect of training programs", Review of Economics and Statistics 67 (3): 648–660.

Ashenfelter, O. and C. Rouse (1995), "Schooling, intelligence and income in America: cracks in the bell curve", Unpublished manuscript (Princeton University).

Axelsson, R. (1989), "Svensk arbetsmarknadsutbildning - en kvantitativ analys av dess effekter", Umeå economic studies (Umeå University).

Balestra, P. and M. Nerlove (1966), "Pooling cross section and time series data in the estimation of a dynamic model: the demand for natural gas", Econometrica 34 (1): 585–612.

Balke, A. (1995), "Probabilistic counterfactuals: semantics, computation, and applications", Technical report R-242 (UCLA).

Balke, A. and J. Pearl (1993), "Nonparametric bounds on causal effects from partial compliance data", Technical report R-199 (UCLA).

Balke, A. and J. Pearl (1997), "Bounds on treatment effects from studies with imperfect compliance", Journal of the American Statistical Association 92 (439): 1171–1176.

Baltagi, B. (1995), Econometric analysis of panel data (Wiley, New York).

Barnow, B. (1987), "The impact of CETA programs on earnings: a review of the literature", Journal of Human Resources 22: 157–193.

Barnow, B., G. Cain and A. Goldberger (1980), "Issues in the analysis of selectivity bias", in: E. Stromsdorfer and G. Farkas, eds., Evaluation studies, Vol. 5 (Sage Publications, Beverly Hills, CA) pp. 290–317.

Barron, J., D. Black and M. Lowenstein (1989), "Job matching and on-the-job training", Journal of Labor Economics 7 (1): 1–19.

Barron, J., M. Berger and D. Black (1997), "How well do we measure training?" Journal of Labor Economics 15 (3): 507–528.

Bassi, L. (1983), "The effect of CETA on the post-program earnings of participants", Journal of Human Resources 18 (Fall): 539–556.

Bassi, L. (1984), "Estimating the effect of training programs with non-random selection", Review of Economics and Statistics 66 (1): 36–43.

Bassi, L. and O. Ashenfelter (1986), "The effects of direct job creation and training programs on low-skilled workers", in: S. Danziger and D. Weinberg, eds., Fighting poverty (Harvard University Press, Cambridge, MA) pp. 133–151.

Bassi, L., M. Simms, L. Burnbridge and C. Betsey (1984), "Measuring the effect of CETA on youth and the economically disadvantaged", Final report prepared for the US Department of Labor under contract no. 20-11-82-19 (The Urban Institute, Washington, DC).

Begg, I., A. Blake and B. Deakin (1991), "YTS and the labour market", British Journal of Industrial Relations 29 (2): 223–236.

Bell, S. and C. Reesman (1987), AFDC Homemaker-Home Health Aide demonstrations: trainee potential and performance (Abt Associates, Cambridge, MA).

Bell, S., L. Orr, J. Blomquist and G. Cain (1995), Program applicants as a comparison group in evaluating training programs (W.E. Upjohn Institute for Employment Research, Kalamazoo, MI).

Bera, A., C. Jarque and L. Lee (1984), "Testing the normality assumption in limited dependent variable models", International Economic Review 25 (3): 563–578.

Berry, D. and B. Fristedt (1985), Bandit problems (Chapman and Hall, London).

Björklund, A. (1989), "Evaluations of training programs: experiences and suggestions for future research", Discussion paper no. 89 (Wissenschaftszentrum, Berlin).

Björklund, A. (1993), "The Swedish experience", in: K. Jensen and P.K. Madsen, eds., Measuring labour market measures (Ministry of Labour, Copenhagen, Denmark) p. 243–263.

Björklund, A. (1994), "Evaluations of Swedish labor market policy", International Journal of Manpower 15 (5, part 2): 16–31.

Björklund, A. and R. Moffitt (1987), "Estimation of wage gains and welfare gains in self-selection models", Review of Economics and Statistics 69 (1): 42–49.

Björklund, A. and H. Regner (1996), "Experimental evaluation of European labour market policy", in: G.

Schmid, J. O'Reilly and K. Schömann, eds., International handbook of labour market policy and evaluation (Edward Elgar, Aldershot, UK) pp. 89–114.

Bloom, H. (1984), "Accounting for no-shows in experimental evaluation designs", Evaluation Review 82 (2): 225–246.

Bloom, H. (1990), Back to work: testing reemployment services for displaced workers (W.E. Upjohn Institute for Employment Research, Kalamazoo, MI).

Bloom, H. and M. McLaughin (1982), CETA training programs: do they work for adults? Joint report (CBO-NCEP, Washington, DC).

Bloom, H., L. Orr, G. Cave, S. Bell and F. Doolittle (1993), The National JTPA Study: title II-A impacts on employment and earnings at 18 months (Abt Associates, Bethesda, MD).

Bonnal, L., D. Fougere and A. Serandon (1997), "Evaluating the impact of French employment policies on individual labour market histories", Review of Economic Studies 64 (4): 683–713.

Bound, J., C. Brown, G. Duncan and W. Rodgers (1994), "Evidence on the validity of cross–sectional and longitudinal labor market data", Journal of Labor Economics 12 (3): 345–368.

Bradley, S. (1994), "The Youth Training Scheme: A critical review of the evaluation literature", International Journal of Manpower 16 (4): 30–56.

Breen, R. (1988), "The work experience program in Ireland", International Labour Review 127 (4): 429–444.

Breen, R. (1991), "Assessing the effectiveness of training and temporary employment schemes: some results from the youth labour market", The Economic and Social Review 22 (3): 177–198.

Brown, R. (1979), Assessing the effects of interview nonresponse on estimates of the impact of supported work (Mathematica Policy Research, Princeton, NJ).

Browning, E. (1987), "On the marginal welfare cost of taxation", American Economic Review 77 (1): 11–23.

Bryant, E. and K. Rupp (1987), "Evaluating the impact of CETA on participant earnings", Evaluation Review 11: 473–492.

Burtless, G. (1995), "The case for randomized field trials in economic and policy research", Journal of Economic Perspectives 9 (2): 63–84.

Butler, R. and J. Heckman (1977), "Government's impact on the labor market status of black Americans: a critical review", in: Equal rights and industrial relations (Industrial Relations Research Association, Madison, WI) pp. 235–281.

Cain, G. (1975), "Regression and selection models to improve nonexperimental comparisons", in: C. Bennett and A. Lumsdaine, eds., Evaluation and experiment (Academic Press, New York) pp. 297–317.

Calmfors, L. (1994), "Active labor market policy and unemployment - a framework for the analysis of crucial design features", OECD Economic Studies 22 (1): 7–47.

Cameron, S. and J. Heckman (1998), "Life cycle schooling and dynamic selection bias: models and evidence for five cohorts of American males", Journal of Political Economy 106 (2): 262–333.

Campbell, D. and J. Stanley (1963), "Experimental and quasi-experimental designs for research on teaching", in: N. Gage, ed., Handbook of research on teaching (Rand McNally, Chicago, IL) pp. 171–246.

Campbell, D. and J. Stanley (1966), Experimental and quasi-experimental designs for research (Rand McNally, Chicago, IL).

Card, D. (1995), "Earnings, schooling and ability revisited", Working paper no. 4832 (NBER, Cambridge, MA).

Card D. and D. Sullivan (1988), "Measuring the effects of CETA participation on movements in and out of employment", Econometrica 56 (3): 497–530.

Cave, G., H. Bos, F. Doolittle and C. Toussaint (1993), JOBSTART: final report on a program for school dropouts (Manpower Demonstration Research Corporation, New York).

Chamberlain, G. (1984), "Panel data", in: Z. Griliches and M. Intriligator, eds, Handbook of econometrics (North-Holland, Amsterdam) pp. 1248–1318.

Chickering, D. and J. Pearl (1996), "A clinician's tool for analyzing non-compliance", in: Proceedings of the National Conference on Artificial Intelligence (AAAI-96) (Morgan Kauffman, Boston, MA) pp. 1269–1276.

Chipman, J. and J. Moore (1976), "Why an increase in GNP need not imply an improvement in potential welfare", Kyklos 29: 391–418.

Cochran, W. and D. Rubin (1973), "Controlling bias in observational studies", Sankyha 35: 417–446.
Cooley, T., T. McGuire and E. Prescott (1979), "Earnings and employment dynamics of manpower trainees: an exploratory econometric analysis", in: R. Ehrenberg, ed., Research in labor economics, Vol. 4, Suppl. 2 (JAI Press, Greenwich, CT) pp. 119–147.
Cook, T. and D. Campbell (1979), Quasi-experimentation: design and analysis issues for field settings (Houghton-Mifflin, Boston, MA).
Corson, W., P. Decker, P. Gleason and W. Nicholson (1993), International trade and worker dislocation: evaluation of the trade adjustment assistance program (Mathematica Policy Research, Princeton, NJ).
Cosslett, S. (1983), "Distribution-free maximum likelihood estimator of the binary choice model", Econometrica 51 (3): 765–782.
Couch, K. (1992), "New evidence on the long-term effects of employment and training programs", Journal of Labor Economics 10 (4): 380–388.
Davidson, C. and S. Woodbury (1993), "The displacement effect of reemployment bonus programs", Journal of Labor Economics 11 (4): 575–605.
Davidson, C. and S. Woodbury (1995), "Wage subsidies for dislocated workers", Unpublished manuscript (W.E. Upjohn Institute for Employment Research, Kalamazoo, MI).
Decker, P. and W. Corson (1995), "International trade and worker displacement: evaluation of the trade adjustment assistance program", Industrial and Labor Relations Review 48 (4): 758–774.
de Koning, J. (1993), "Measuring the placement effects of two wage subsidy schemes for the long term unemployed", Empirical Economics 18: 447–468.
de Koning, J., M. Koss and A. Verkaik (1991), "A quasi-experimental evaluation of the vocational training centre for adults", Environment and Planning C: Government and Policy 9: 143–153.
Delander, L. (1978), "Studier kring den arbetsformedlande verksamheten" (Studies of the Swedish Employment Office) in SOU: 60.
Devine, T. and J. Heckman (1996), "The structure and consequences of eligibility rules for a social program", in: S. Polachek, ed., Research in labor economics, Vol. 15 (JAI Press, Greenwich, CT) pp. 111–170.
Dickinson, K., T. Johnson and R. West (1984), "An analysis of the impact of CETA programs on participants' earnings", Report prepared for the US Department of Labor under contract no. 20-06-82-21 (SRI International, Menlo Park, CA).
Dickinson, K., T. Johnson and R. West (1986), "An analysis of the impact of CETA on participants' earnings", Journal of Human Resources, 21: 64–91.
Dickinson, K., T. Johnson and R. West (1987), "An analysis of the sensitivity of quasi-experimental net estimates of CETA programs", Evaluation Review 11: 452–472.
Dolton, P. (1993), "The economics of youth training in Britain", Economic Journal 103 (420): 1261–1278.
Dolton, P. and D. O'Neill (1996a), "Unemployment duration and the Restart effect: some experimental evidence", Economic Journal 106 (435): 387–400.
Dolton, P. and D. O'Neill (1996b), "The Restart effect and the return to full-time stable employment", Journal of the Royal Statistical Society Series A 159 (2): 275–288.
Dolton, P. and D. O'Neill (1997), "The long-run effect of unemployment monitoring and work search programs: some experimental evidence from the U.K.", Unpublished monograph (University of Newcastle-upon-Tyne).
Dolton, P., G. Makepeace and J. Treble (1992), "Public- and private-sector training of young people in Britian", in: L. Lynch, ed., Training and the private sector (University of Chicago Press, Chicago, IL) pp. 261–281.
Dolton, P., G. Makepeace and J. Treble (1994a), "The Youth Training Scheme and the school-to-work transition", Oxford Economic Papers 46 (4): 629–657.
Dolton, P., G. Makepeace and J. Treble (1994b), "The wage effect of YTS: evidence from YCS", Scottish Journal of Political Economy 41 (4): 444–453.
Donohue, J. and P. Siegelman (1998), "Allocating resources among prisons and social programs in the battle against crime", Journal of Legal Studies 27 (1): 1–43.
Doolittle, F. and L. Traeger (1990), Implementing the National JTPA Study (Manpower Demonstration Research Corporation, New York).

Eberwein, C., J. Ham and R. LaLonde (1997), "The impact of classroom training on the employment histories of disadvantaged women: evidence from experimental data", Review of Economic Studies 64 (4): 655–682.

Edin, P.-A. (1988), "Individual consequences of plant closures", PhD dissertation (Uppsala University).

Engstrom, L., K. Lofgren and O. Westerlund (1988), "Intensified employment services, unemployment duration and unemployment risks", Economic studies no. 186 (Umeå University).

Farber, H. and R. Gibbons (1994), "Learning and wage dynamics", Unpublished manuscript (Princeton University).

Fay, R. (1996), "Enhancing the effectiveness of active labour market policies: evidence from programme evaluations in OECD Countries", Occasional papers no. 18 (Labour market and social policy, OECD, Paris).

Fechner, G. (1860), Elemente der psychophysik (Breitkopf and Härtel, Leipzig, Germany).

Finifter, D. (1987), "An approach to estimating the net earnings impact of federally subsidized employment and training programs", Evaluation Review 11 (4): 528–547.

Fisher, R. (1935), Design of experiments (Hafner, New York).

Flinn, C. and J. Heckman (1982), "New methods for analyzing structural models of labor force dynamics", Journal of Econometrics 18 (1): 115–168.

Forslund, A. and A. Krueger (1997), "An evaluation of the Swedish active labor market policy: new and received wisdom", in: R. Freeman, R. Topel and B. Swedenburg, eds, The welfare state in transition (The University of Chicago Press for NBER, Chicago, IL).

Fraker, T. and R. Maynard (1987), "The adequacy of comparison group designs for evaluations of employment-related programs", Journal of Human Resources 22 (2): 194–227.

Friedlander, D. and G. Hamilton (1993), The saturation work initiative model in San Diego: a five-year follow-up study (Manpower Demonstration Research Corporation, New York).

Friedlander, D and P. Robbins (1995), "Evaluating program evaluations: new evidence on commonly used nonexperimental methods", American Economic Review 85 (4): 923–937.

Friedlander, D., G. Hoertz, J. Quint and J. Riccio (1985), Arkansas, the demonstration of state work/welfare initiatives: the final report on the WORK program in two counties (Manpower Demonstration Research Corporation, New York).

Friedlander, D., D. Greenberg and P. Robins (1997), "Evaluating government training programs for the economically disadvantaged", Journal of Economic Literature 35 (4): 1809–1855.

Gay, R. and M. Borus (1980), "Validating performance indicators for employment and training programs", Journal of Human Resources 15: 29–48.

Geraci, V. (1984), "Short-term indicators of job training program effects on long-term participant earnings, Report prepared for US Department of Labor under contract no. 20-48-82-16.

Glynn, R. and D. Rubin (1986), "Selection modeling versus mixture modeling", in: H. Wainer, ed., Drawing inferences from selected samples (Springer-Verlag, Berlin).

Goldberger, A. (1972), "Selection bias in evaluating treatment effects", Discussion paper no. 123-172 (Institute for Research on Poverty, University of Wisconsin).

Goldman, B., D. Friedlander and D. Long (1986), California, the demonstration of state work/welfare initiatives: final report on the San Diego job search and work experience demonstration (Manpower Demonstration Research Corporation, New York).

Gramlich, E. and B. C. Ysander (1981), "Relief work and grant displacement in Sweden", in: G. Eliasson, B. Holmlund and F. Stafford, eds., Studies in labor market behavior: Sweden and the United States (The Industrial Research Institute, Stockholm, Sweden) pp. 139–166.

Green, F., M. Hoskins and S. Montgomery (1996) "The effects of company training, further education and the youth training scheme on the earnings of young employees", Oxford Bulletin of Economics and Statistics 58 (3) 469–488.

Greenberg, D. (1997), "The leisure bias in cost–benefit analyses of employment and training programs", Journal of Human Resources 32 (2): 413–439.

Greenberg, D. and M. Wiseman (1992), "What did the OBRA demonstrations do?" in: C. Manski and I.

Garfinkel, eds., Evaluating welfare and training programs (Harvard University Press, Cambridge, MA) pp. 25–75.
Gritz, M. (1993), "The impact of training on the frequency and duration of employment", Journal of Econometrics 57 (1–3): 21–51.
Grossman, J., R Maynard and J. Roberts (1985), Reanalysis of the effects of selected employment and training programs for welfare recipients (Mathematica Policy Research, Princeton, NJ).
Gueron, J. (1990), "Work and welfare: lessons on employment programs", Journal of Economic Perspectives 4 (1): 79–98.
Hahn, J., P. Todd and W. van der Klaauw (1998), "Estimation of treatment effects with a quasi-experimental regression-discontinuity design with application to evaluating the effect of federal antidiscrimination laws on minority employment in small U.S. firms", Unpublished manuscript (University of Pennsylvania).
Ham J. and R. LaLonde (1990), "Using social experiments to estimate the effect of training on transition rates", in: J. Hartog, G. Ridder and J. Theeuwes, eds., Panel data and labor market studies (North-Holland, Amsterdam) pp. 157–172.
Ham J. and R. LaLonde (1996), "The effect of sample selection and initial conditions in duration models: evidence from experimental data", Econometrica 64 (1): 175–205.
Hamermesh, D. (1971), Economic aspects of manpower training programs (Lexington Books, Lexington, MA).
Hamermesh, D. (1993), Labor demand (Princeton University Press, Princeton, NJ).
Harberger, A. (1971), "Three basic postulates for applied welfare economics", Journal of Economic Literature 9 (3): 785–797.
Harkman, A., F. Jansson and A. Tamas (1996), "Effects, defects, and prospects – an evaluation of labor market training in Sweden", Unpublished manuscript (Research Unit, Swedish National Labour Market Board).
Haveman, R. and D. Saks (1985), "Transatlantic lessons for employment and training policy", Industrial Relations 24 (2): 20–36.
Heckman, J. (1976), "Simultaneous equations models with continuous and discrete endogenous variables and structural shifts", in: S. Goldfeld and R. Quandt, eds., Studies in nonlinear estimation (Ballinger, Cambridge, MA).
Heckman, J. (1978), "Dummy endogenous variables in a sumultaneous equations system", Econometrica 46 (4): 931–959.
Heckman, J. (1979), "Sample selection bias as a specification error", Econometrica 47 (1): 153–161.
Heckman, J. (1980), "Addendum to sample selection bias as a specification error", in: E. Stromsdorfer and G. Farkas, eds., Evaluation studies review annual, Vol. 5 (Sage, San Francisco, CA) pp. 970–995.
Heckman, J. (1990), "Varieties of selection bias", American Economic Review 80 (2): 313–318.
Heckman, J. (1992), "Randomization and social policy evaluation", in: C. Manski and I. Garfinkel, eds., Evaluating welfare and training programs (Harvard University Press, Cambridge, MA) pp. 201–230.
Heckman, J. (1996), "Randomization as an instrumental variable", Review of Economics and Statistics 78 (2): 336–341.
Heckman, J. (1997), "Instrumental variables: a study of implicit behavioral assumptions in one widely used estimator", Journal of Human Resources, 32 (3): 441–461.
Heckman, J. (1998a), "The economic evaluation of social programs", in: J. Heckman and E. Leamer, eds., Handbook of econometrics, Vol. 5 (Elsevier, Amsterdam), in press.
Heckman, J., ed. (1998b), Performance standards in a government bureaucracy: analytical essays on the JTPA performance standards system (W.E. Upjohn Institute for Employment Research, Kalamazoo, MI).
Heckman, J. (1998c), "A unified matching and weighting framework for all evaluation estimators", Unpublished manuscript (University of Chicago).
Heckman, J. and G. Borjas (1980), "Does unemployment cause future unemployment? Definitions, questions and answers from a continuous time model of heterogeneity and state dependence", Economica 47 (187): 247–283.
Heckman, J. and B. Honoré (1990), "The empirical content of the Roy model", Econometrica 58 (5): 1121–1149.
Heckman, J. and J. Hotz (1989), "Choosing among alternative methods of estimating the impact of social

programs: the case of manpower training", Journal of the American Statistical Association 84 (408): 862–874.

Heckman, J. and T. MaCurdy (1986), "Labor econometrics", in: Z. Griliches and M. Intriligator, eds., Handbook of econometrics (North-Holland, Amsterdam) pp. 1917–1977.

Heckman, J. and R. Robb (1982), "The longitudinal analysis of earnings", Unpublished manuscript (University of Chicago).

Heckman, J. and R. Robb (1985a), "Alternative methods for evaluating the impact of interventions", in: J. Heckman and B. Singer, eds., Longitudinal analysis of labor market data (Cambridge University Press for Econometric Society Monograph Series, New York) pp. 156–246.

Heckman, J. and R. Robb (1985b), "Alternative methods for evaluating the impact of interventions: an overview", Journal of Econometrics 30 (1,2): 239–267.

Heckman, J. and R. Robb (1986a), "Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes", in: H. Wainer, ed., Drawing inferences from selected samples (Springer-Verlag, Berlin) pp. 63–107.

Heckman, J. and R. Robb (1986b), "Alternative identifying assumptions in econometric models of selection bias", in: G. Rhodes, ed. Advances in econometrics, Vol. 5 (JAI Press, Greenwich, CT) pp. 243–287.

Heckman, J. and R. Roselius (1994), "Evaluating the impact of training on the earnings and labor force status of young women: better data help a lot", Unpublished manuscript (University of Chicago).

Heckman, J. and G. Sedlacek (1985), "Heterogeneity, aggregation and market wage functions: an empirical model of self-selection in the labor market", Journal of Political Economy 98 (6): 1077–1125.

Heckman, J. and B. Singer (1984), "A method for minimizing the impact of distributional assumptions in econometric models for duration data", Econometrica 52 (2): 271–320.

Heckman, J. and J. Smith (1993), "Assessing the case for randomized evaluation of social programs", in: K. Jensen and P.K. Madsen, eds., Measuring labour market measures (Ministry of Labour, Copenhagen, Denmark) pp. 35–96.

Heckman, J. and J. Smith (1995), "Assessing the case for social experiments", Journal of Economic Perspectives 9 (2): 85–100.

Heckman, J. and J. Smith (1998a), "Evaluating the welfare state", in: S. Strom, ed., Econometrics and economics in the 20th century: the Ragnar Frisch centenary (Cambridge University Press for Econometric Society Monograph Series, New York).

Heckman, J. and J. Smith (1998b), "The sensitivity of experimental impact estimates: evidence from the National JTPA Study", in: R. Freeman and L. Katz, eds., Youth employment and unemployment in the OECD countries (University of Chicago Press for NBER, Chicago, IL) in press.

Heckman, J. and J. Smith (1998c), "The performance of performance standards: the effects of JTPA performance standards on efficiency, equity and participant outcomes", Unpublished manuscript (University of Chicago).

Heckman, J. and J. Smith (1998d), "The determinants of participation in a social program: evidence from the job training partnership act", Unpublished manuscript (University of Chicago).

Heckman, J. and J. Smith (1998e), "The sensitivity of nonexperimental evaluation estimators: a simulation study", Unpublished manuscript (University of Chicago).

Heckman, J. and J. Smith (1999), "The pre-program dip and the determinants of program participation in a social program: implications for simple program evaluation strategies", Economic Journal, in press.

Heckman, J. and P. Todd (1994), "Interpreting standard measures of selection bias", Unpublished manuscript (University of Chicago).

Heckman, J. and E. Vytlacil (1999a), "Local instrumental variables and latent variable models for identifying and bounding treatment effects", Proceedings of the National Academy of Sciences USA 96: 4730–4734.

Heckman, J. and E. Vytlacil (1999b), "The relationship between treatment parameters within a latent variable framework", Economics Letters, in press.

Heckman, J. and K. Wolpin (1976), "Does the contract compliance program work? An analysis of Chicago data", Industrial and Labor Relations Review 19: 415–433.

Heckman, J., R. Roselius and J. Smith (1993), "U.S. education and training policy: a re-evaluation of the

underlying assumptions behind the 'new consensus'", in: A. Levenson and L. Solomon, eds., Labor markets, employment policy and job creation (Milken Institute for Job and Capital Formation, Santa Monica, CA) pp. 83–121.

Heckman, J., M. Khoo, R. Roselius and J. Smith (1996a), "The empirical importance of randomization bias in social experiments: evidence from the national JTPA study", Unpublished manuscript (University of Chicago).

Heckman, J., H. Ichimura, J. Smith and P. Todd (1996b), "Sources of selection bias in evaluating social programs: an interpretation of conventional measures and evidence on the effectiveness of matching as a program evaluation method", Proceedings of the National Academy of Sciences USA 93 (23): 13416–13420.

Heckman, J., J. Smith and C. Taber (1996c), "What do bureaucrats do? The effects of performance standards and bureaucratic preferences on acceptance into the JTPA program", in: G. Libecap, ed., Reinventing government and the problem of bureaucracy, Vol. 6, Advances in the study of entrepreneurship, innovation and economic growth (JAI Press, Greenwich, CT) pp. 191–218.

Heckman, J., H. Ichimura and P. Todd (1997a), "Matching as an econometric evaluation estimator: evidence from evaluating a job training program", Review of Economic Studies 64 (4): 605–654.

Heckman, J., L. Lochner, J. Smith and C. Taber (1997b), "The effects of government policies on human capital investment and wage inequality", Chicago Policy Review 1 (2): 1–40.

Heckman, J., J. Smith and N. Clements (1997c), "Making the most out of programme evaluations and social experiments: accounting for heterogeneity in programme impacts", Review of Economic Studies 64 (4): 487–535.

Heckman, J., N. Hohmann and J. Smith with M. Khoo (1998a), "Substitution and dropout bias in social experiments: evidence from an influential social experiment", Quarterly Journal of Economics, in press.

Heckman, J., H. Ichimura, J. Smith and P. Todd (1998b), "Characterizing selection bias using experimental data", Econometrica 66: 1017–1098.

Heckman, J., H. Ichimura and P. Todd (1998c), "Matching as an econometric evaluation estimator", Review of Economic Studies 65 (2): 261–294.

Heckman, J., L. Lochner and C. Taber (1998d), "Explaining rising wage inequality: explorations with a dynamic general equilibrium model of labor earnings with heterogeneous agents", Review of Economic Dynamics 1 (1): 1–64.

Heckman, J., L. Lochner and C. Taber (1998e), "General equilibrium treatment effects: a study of tuition policy", American Economic Review 88 (2): 381–386.

Heckman, J., J. Smith and C. Taber (1998f), "Accounting for dropouts in evaluations of social programs", Review of Economics and Statistics 80 (1): 1–14.

Heckman, J., J. Smith and P. Todd (1999), "The evaluation problem", Unpublished manuscript (University of Chicago).

Heinrich, C. (1998), "The role of performance standards in JTPA program administration and service delivery at the local level", in: J. Heckman, ed., Performance standards in a government bureaucracy: analytical essays on the JTPA performance standards system (W.E. Upjohn Institute for Employment Studies, Kalamazoo, MI) in press.

Holland, P. (1986), "Statistics and causal inference", Journal of the American Statistical Association 81 (396): 945–960.

Holland, P. (1988), "Causal inference, path analysis and recursive structural equation models", in: C. Clogg, ed., Sociological methodology (American Sociological Association, Washington, DC) pp. 449–484.

Hollister, R. and D. Freedman (1988), "Special employment programmes in OECD countries", International Labour Review 127 (3): 317–334.

Hollister, R., P. Kemper and R. Maynard (1984), The National Supported Work demonstration (University of Wisconsin Press, Madison, WI).

Honoré, B. and E. Kyriazidou (1998), "Panel data discrete choice models with lagged dependent variables", Unpublished manuscript (University of Chicago).

Hotz, V.J. (1992), "Designing an evaluation of the Job Training Partnership Act", in: C. Manski and I. Garfinkel, eds., Evaluating welfare and training programs (Harvard University Press, Cambridge, MA) pp. 76–114.

Hsiao, C. (1986), Analysis of panel data (Cambridge University Press for Econometric Society Monograph Series, Cambridge, UK).

Hutchinson, G. and A. Church (1989), "Wages, unions, the Youth Training Scheme and the Young Workers Scheme", Scottish Journal of Political Economy 36 (2): 160–182.

Ichimura, H. (1993), "Semiparametric least squares (SLS) and weighted SLS estimation of single-index models", Journal of Econometrics 58 (1,2): 71–120.

Imbens, G. and J. Angrist (1994), "Identification and estimation of local average treatment effects", Econometrica 62 (4): 467–476.

Imbens, G. and T. Lancaster (1996), "Case-control studies with contaminated controls", Journal of Econometrics 71 (1,2): 145–160.

Jensen, P., P. Pederson, N. Smith and N. Westergard-Nielson (1993), "The effects of labor market training on wages and unemployment: some Danish results", in: H. Bunzel, P. Jensen and N. Westergard-Nielson, eds., Panel data and labour market dynamics, contributions to economic analysis no. 222 (North Holland, Amsterdam) pp. 311–331.

Johnson, G. (1979), "The labor market displacement effect in the analysis of the net impact of manpower training programs", in: F. Bloch, ed., Evaluating manpower training programs, Suppl. 1 (JAI Press, Greenwich, CT) pp. 227–254.

Johnson, G. and R. Layard (1986), "The natural rate of unemployment: explanation and policy", in O. Ashenfelter and R. Layard, eds., Handbook of labor economics, Vol. 2 (North-Holland, Amsterdam) pp. 921–999.

Johnson, G. and J. Tomola (1977), "The fiscal substitution effects of alternative approaches to public service employment", Journal of Human Resources 12 (1): 3–26.

Kane, T. (1994), "College entry by blacks since 1970: the role of college costs, family background and the return to education", Journal of Political Economy 102 (5): 878–912.

Kane, T. and C. Rouse (1993), "Labor market returns to two- and four-year college", American Economic Review 85 (3): 600–614.

Kemper, P., D. Long and C. Thornton (1981), The supported work evaluation: final cost benefit analysis (Manpower Demonstration Research Corporation, New York).

Kemper, P., D. Long and C. Thornton (1984), "A benefit–cost analysis of the supported work experiment", in: R. Hollister, P. Kemper and R. Maynard, eds., The National Supported Work demonstration (University of Wisconsin Press, Madison, WI) pp. 239–285.

Kemple, J., D. Friedlander and V. Fellerath (1995), Florida's project independence: benefits, costs and two-year impacts of Florida's JOBS program (Manpower Demonstration Research Corporation, New York).

Kiefer, N. (1979), The economic benefits of four employment and training programs (Garland Publishing, New York).

Knox, V., P. Auspos, J. Hunter-Manns, C. Miller and A. Orenstein (1997), Making welfare work: 18-month impacts of Minnesota's family investment program (Manpower Demonstration Research Corporation, New York).

Kornfeld, R. and H. Bloom (1996), "Measuring the impacts of social programs on the earnings and employment of low income persons: do UI wage records and surveys agree?" Unpublished manuscript (Abt Associates, Bethesda, MD).

Kraus, F., P. Puhani and V. Steiner (1997), "Employment effects of publically financed training programs – the East German experience", Discussion paper no. 97-33 (Zentrum für Europäische Wirtschaftsforshung).

Laffont, J. (1989), Fundamentals of public economics (MIT Press, Cambridge, MA).

LaLonde, R. (1984), "Evaluating the econometric evaluations of training programs with experimental data", Working paper no. 183 (Industrial Relations Section, Princeton University).

LaLonde, R. (1986), "Evaluating the econometric evaluations of training programs with experimental data", American Economic Review 76 (4): 604–620.

LaLonde, R. (1995), "The promise of public sector-sponsored training programs", Journal of Economic Perspectives 9 (2): 149–168.

LaLonde, R. and R. Maynard (1987), "How precise are evaluations of employment and training programs: evidence from a field experiment", Evaluation Review 11: 428–451.

Lancaster, T. (1990), Econometric analysis of transition data (Cambridge University Press for Econometric Society Monograph Series, Cambridge, UK).

Lechner, M. (1996), "An evaluation of public-sector-sponsored continuous vocational training programs in East Germany", Unpublished manuscript (Universität Mannheim).

Lechner, M. (1997), "Earnings and employment effects of continuous off-the-job training in East Germany after unification", Unpublished manuscript (Universität Mannheim).

Lee, L. (1983), "Generalized econometric models with selectivity", Econometrica 51 (2): 507–512.

Leigh, D. (1990), Does training work for displaced workers? (W.E. Upjohn Institute for Employment Research, Kalamazoo, MI).

Leigh, D. (1995), Assisting workers displaced by structural change (W.E. Upjohn Institute for Employment Research, Kalamazoo, MI).

Lewbel, A. (1998), "Semiparametric qualitative response model estimation with instrumental variables and unknown heteroskedasticity", Unpublished manuscript (Brandeis University).

Lewis, H.G. (1963) Unionism and relative wages (University of Chicago Press, Chicago, IL).

MaCurdy, T. (1982), "The use of time series processes to model the error structure of earnings in a longitudinal data analysis", Journal of Econometrics 18 (1): 83–114.

Main, B. (1985), "School leaver unemployment and the Youth Opportunitues Programme in Scotland", Oxford Economic Papers 37 (3): 426–447.

Main, B. (1991), "The effects of the Youth Training Scheme on employment probability", Applied Economics 23 (2): 367–372.

Main, B. and D. Raffe (1983), "Determinants of employment and unemployment among school leavers: evidence from the 1979 survey of scottish school leavers", Scottish Journal of Political Economy 30 (1): 1–17.

Main, B. and M. Shelly (1990), "The effectiveness of the Youth Training Scheme as a manpower policy", Economica 57 (228): 495–514.

Mallar, C. (1978), "Alternative econometric procedures for program evaluations: illustrations from an evaluation of Job Corps", Proceedings of the American Statistical Association: 317–321.

Mallar, C., S. Kerachsky, C. Thornton and D. Long (1982), Evaluation of the economic impact of the Job Corps program: third follow-up report (Mathematica Policy, Princeton, NJ).

Manski C. (1995), The identification problem in the social sciences (Harvard University Press, Cambridge, MA)

Manski C. and S. Lerman (1977), "The estimation of choice probabilites from choice-based samples", Econometrica 45 (8): 1977–1988.

Manski, C. and D. McFadden (1981), "Alternative estimators and sample designs for discrete choice analysis", in: C. Manski and D. McFadden, eds., Structural analysis of discrete data with econometric applications (MIT Press, Cambridge, MA) pp. 1–50.

Masters, S. and R. Maynard (1981), The impact of supported work on long-term recipients of AFDC benefits (Manpower Demonstration Research Corporation, New York).

Matzkin, R. (1992), "Nonparametric and distribution-free estimation of the binary threshold crossing and the binary choice models", Econometrica 60 (2): 239–270.

Matzkin, R. (1993), "Nonparametric identification and estimation of polychotomous choice models", Journal of Econometrics 58 (1,2): 137–168.

Maynard, R. (1980), The impact of supported work on young school dropouts (Manpower Demonstration Research Corporation, New York).

McLennan, A. (1991), "Binary stochastic choice", in: J. Chipman, D. McFadden and M. Richter, eds., Preferences, uncertainty and optimality: essays in honor of Leonid Hurwicz (Westview Press, Boulder, CO) pp. 187–202.

Mincer, J. (1962), "On-the-job training: costs, returns and some implications", Journal of Political Economy 70 (5): 50–79.
Mincer, J. (1993), "Investment in U.S. education and training", Discussion paper no. 671 (Columbia University).
Moffitt, R. (1992), "Evaluation methods for program entry effects", in: C. Manski and I. Garfinkel, eds., Evaluating welfare and training programs (Harvard University Press, Cambridge, MA).
National Commission for Employment Policy (1987), The Job Training Partnership Act (US Government Printing Office, Washington, DC).
Neyman, J. (1935), "Statistical problems in agricultural experiments", The Journal of the Royal Statistical Society 2 (2) (Suppl.): 107–180.
O'Connell, P. and F. McGinnity (1997), "What works, who works? The employment and earnings effects of active labour market programmes among young people in Ireland", Work, Employment and Society 11 (4): 639–661.
O'Higgins, N. (1994), "YTS, employment and sample selection bias", Oxford Economic Papers 46 (4): 605–628.
OECD (1993), "Active labour market policies: assessing macroeconomic and microeconomic effects", in: Employment outlook (OECD, Paris) pp. 39–67.
OECD (1996), Employment outlook (OECD, Paris).
Orr, L., H. Bloom, S. Bell, W. Lin, G. Cave and F. Doolittle (1994), "The National JTPA Study: impacts, benefits and costs of title II-A", Produced for the US Department of Labor under contract no. 99-6-0803-77-068 (Abt Associates, Bethesda, MD).
Park, N., W.C. Riddell and R. Power (1993) An evaluation of UI-sponsored training (Evaluation Branch, Human Resources Development Canada).
Payne, J., S. Lissenburg and M. White (1996), Employment training and employment action: an evaluation by the matched comparison method (Policy Studies Institute, London).
Perry, C., R. Anderson, R. Rowan and H. Northrupt (1975), The impact of government manpower programs (University of Pennsylvania Press, Philadelphia, PA).
Powell, J. (1994), "Estimation of semiparametric models", in: R. Engle and D. McFadden, eds., Handbook of econometrics, Vol. 4 (North-Holland, Amsterdam) pp. 2443–2521.
Puma, M. and N. Burstein (1994), "The national evaluation of the Food Stamp employment and training program", Journal of Policy Analysis and Management 13 (2): 311–330.
Quandt, R. (1972), "Methods for estimating switching regressions", Journal of the American Statistical Association 67 (338): 306–310.
Quandt, R. (1988), The economics of disequilibrium (Basil Blackwell, Oxford, UK).
Quint, J., B. Fink and S. Rowser (1994), New chance: interim findings on a comprehensive program for disadvantaged mothers and their children (Manpower Demonstration Research Corporation, New York).
Raaum, O. and H. Torp (1997), "Labour market training in Norway – effect on earnings", Report no. 46/97 (Stiftelsen for samfunns– ognaeringslivsforskning).
Rangarajan, A., J. Burghardt and A. Gordon (1992), Evaluation of the minority single parent demonstration, Vol. I: summary (Mathematica Policy Research, Princeton, NJ).
Rao, C.R. (1965), "On discrete distributions arising out of methods of ascertainment", in: G.P. Patil, ed., Classical and contagious discrete distributions (Statistical Publication Society, Calcutta).
Rao, C.R. (1986), "Weighted distributions", in: S. Feinberg, ed., A celebration of statistics (Springer-Verlag, Berlin).
Regner, H. (1996), "A nonexperimental evaluation of manpower training in Sweden", Unpublished manuscript (Stockholm University).
Regner, H. (1997), Training at the job and training for a new job: two Swedish studies (Swedish Institute for Social Research, Stockholm, Sweden).
Riccio, J., D. Friedlander and S. Freedman (1994), GAIN: benefits, costs and three-year impacts of a welfare-to-work program (Manpower Demonstration Research Corporation, New York).
Ridder, G. (1986), "An event history approach to the evaluation of training, recruitment and employment programmes", Journal of Applied Econometrics 11: 109–126.

Robins, J. (1989), "The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies", in: L. Sechrest, H. Freeman and A. Mulley, eds., Health service research methodology: a focus on AIDS (US Public Health Service, Washington, DC) pp. 113–159.

Robinson, P. (1996), "The role and limits of active labour market policy", Working paper RSC no. 96/27 (European Union Institute).

Rosenbaum, P. (1995) Observational studies (Springer-Verlag, Leipzig, Germany).

Rosenbaum, P. and D. Rubin (1983), "The central role of the propensity score in observational studies for causal effects", Biometrika 70 (1): 41–55.

Roy, A. (1951), "Some thoughts on the distribution of earnings", Oxford Economic Papers 3: 135–146.

Royden, H. (1968), Real analysis, 2nd edition (MacMillan Press, New York).

Rubin, D. (1974), "Estimating causal effects of treatments in randomized and non–randomized studies", Journal of Educational Psychology 66: 688–701.

Rubin, D. (1978), "Bayesian inference for causal effects: the role of randomization", Annals of Statistics 6 (1): 34–58.

Rubin, D. (1979), "Using multivariate matched sampling and regression adjustment to control bias in observational studies", Journal of the American Statistical Association 74: 318–328.

Sandell, S. and K. Rupp (1988), "Who is served in JTPA programs: patterns of participation and intergroup equity", US National Commission for Employment Policy RR-88-03.

Smith, J. (1992), "The JTPA selection process: a descriptive analysis", Unpublished manuscript (University of Chicago).

Smith, J. (1994), "A note on estimating the relative costs of experimental and non-experimental evaluations using cost data from the National JTPA Study", Unpublished manuscript (University of Chicago).

Smith, J. (1997a), "Measuring earnings dynamics among the poor: evidence from two samples of JTPA eligibles", Unpublished manuscript (University of Western Ontario).

Smith, J. (1997b), "Measuring earnings levels among the poor: evidence from two samples of JTPA eligibles", Unpublished manuscript (University of Western Ontario).

Smith, J. and F. Welch (1986), Closing the gap: forty years of economic progress for blacks (RAND, Santa Monica, CA).

Stormsdorfer, E., R. Boruch, H. Bloom, J. Gueron and F. Stafford (1985), Recommendations of the Job Training Longitudinal Survey Research Advisory Panel to the Office of Strategic Planning and Policy Development (US Department of Labor, Washington, DC).

Sudman, S. and N. Bradburn (1982), Asking questions (Jossey-Bass, San Francisco, CA).

Thierry, P. and M. Sollogoub (1995), "Les politiques francaises d'emploi en faveur des jeunes. Une evaluation econometrique", Revue-Economique 46 (3): 549–559.

Topel, R. and M. Ward (1992), "Job mobility and the careers of young men", Quarterly Journal of Economics 107: 439–480.

Torp, H., O. Raaum, E. Heraes and H. Goldstein (1993), "The first Norwegian experiment", in: K. Jensen and P.K. Madsen, eds., Measuring labour market measures (Ministry of Labour, Copenhagen, Denmark) pp. 97–140.

Trochim, W. (1984), Research design for program evaluation: the regression-discontinuity approach (Sage, Newbury Park, CA).

Trott, C. and J. Baj (1993), "An analysis of repeating in JTPA in Illinois", Report prepared for the Illinois Department of Commerce and Community Affairs (Northern Illinois University).

USGAO (1991), "Job Training Partnership Act: racial and gender disparities in services", Report no. GAO/HRD-91-148 (US General Accounting Office).

USGAO (1996), "Job Training Partnership Act: long-term earnings and employment outcomes", Report no. GAO/HEHE-96-40 (US General Accounting Office).

van der Klaauw, W. (1997), "A regression-discontinuity evaluation of the effect of financial aid offers on college enrollment", Unpublished manuscript (New York University).

Vytlacil, E. (1999), "Independence, monotonicity and latent variable models: an equivalence result", Unpublished manuscript (Univeristy of Chicago).

Weitzman, M. (1979), "Optimal search for the best alternative", Econometrica 47 (3): 641–654.

Westat (1981), "Continuous longitudinal manpower survey net impact report no. 1: Impact on 1997 earnings of new FY 1976 CETA enrollees in selected program activities", Report prepared for the US Department of Labor under contract no. 23-24-75-07.

Westat (1984), "Summary of net impact results", Report prepared for US Department of Labor under contract no. 23-24-75-07.

Westergard-Nielsen, N. (1993), "The effects of training: a fixed effect model", in: K. Jensen and P.K. Madsen, eds., Measuring labour market measures (Ministry of Labour, Copenhagen, Denmark) pp. 167–200.

White, M. and J. Lakey (1992), The Restart effect: does active labour market policy reduce unemployment? (Policy Studies Institute, London).

Whitfield, K. and C. Bourlakis (1991), "An empirical analysis of YTS, employment and earnings", Journal of Economic Studies 18 (1): 42–56.

Zweimuller, J. and R. Winter-Ebmer (1996), "Manpower training programmes and employment stability", Economica 63 (249): 113–130.