

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

HARVARD UNIVERSITY
THE GRADUATE SCHOOL OF ARTS AND SCIENCES



THESIS ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Division

Department Economics

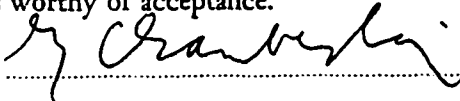
Committee

have examined a thesis entitled

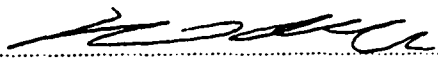
Econometric Methods for Program Evaluation

presented by Rajeev H. Dehejia

candidate for the degree of Doctor of Philosophy and hereby
certify that it is worthy of acceptance.

Signature 

Typed name Gary Chamberlain, Chair

Signature 

Typed name Edward Glaeser

Signature 

Typed name Guido Imbens

Date May 19, 1997

Econometric Methods for Program Evaluation

A thesis presented by

Rajeev Harsha Dehejia

to

The Department of Economics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Economics

**Harvard University
Cambridge, Massachusetts**

May 1997

UMI Number: 9733271

**Copyright 1997 by
Dehejia, Rajeev Harsha**

All rights reserved.

**UMI Microform 9733271
Copyright 1997, by UMI Company. All rights reserved.**

**This microform edition is protected against unauthorized
copying under Title 17, United States Code.**

UMI
300 North Zeeb Road
Ann Arbor, MI 48103

**© 1997 by Rajeev Harsha Dehejia
All rights reserved**

Abstract

This thesis deals with the use of econometric methods to address problems of program evaluation. Program evaluation refers broadly to the assessment of the impact of a program or policy variation, referred to as “the treatment”, on some outcomes of interest; examples of treatments include labor training programs, the adoption of new high-yield seeds, and a change in minimum wage regulations.

The first chapter considers methods that can be used to obtain unbiased estimates of treatment impacts when the available data are from non-experimental studies. Data from randomized experiments are considered ideal, because random assignment into treatment and control implies that simple mean comparisons of outcomes of interest across treatment and control yield unbiased estimates of the treatment impact. Instead when data is gathered non-experimentally, many factors can confound estimates of the treatment impact, a leading instance being sample selection bias. The first chapter demonstrates the use of propensity score methods to adjust in a flexible way for those sources of bias that are attributable to observable differences between the treatment and control groups. Using data from the National Supported Work Program, I demonstrate that propensity score methods succeed in yielding accurate estimates of the treatment effect.

In Chapter 2, using data from the Greater Avenues for Independence (GAIN) experiment, I argue for the use of Bayesian decision theory to set up and solve the decision problems implicitly motivating the program evaluation. There are two advantages to this approach. First, when using standard methods, one concludes that the impact of GAIN is small and statistically insignificant. Instead, in terms of the

decision problem, the impact is economically significant, i.e., *any* risk-averse or -neutral agent prefers the distribution of outcomes under GAIN. Second, I show that a decision-theoretic approach allows us to evaluate hypothetical policies, such as allowing career counselors to assign individuals into the programs, alongside the standard policies of assigning everyone into treatment or control. By assigning only a subset of individuals into treatment, such policies turn out both to be less costly and to yield higher earnings for the participants.

*For my parents and brother, those who have held me,
for all those who are and have been for them –
there is no line, only the two-headed arrow
that threads forward beyond speculation's edge
and back beyond memory and names.*

Table of Contents

Acknowledgment

Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs	1
<i>Abstract</i>	2
1. Introduction	3
2. Review of the Literature	6
3. Selection on Observables and the Role of the Propensity Score	10
4. Estimating the Average Treatment Effect	18
5. The Data	28
6. Results Using the Propensity Score	35
7. Sensitivity Analysis	43
8. Conclusion	47
<i>References</i>	50
<i>Tables</i>	55
A Decision-Theoretic Approach to Program Evaluation	74
<i>Abstract</i>	75
1. Introduction	77
2. The GAIN Program and the GAIN Experiment	80
3. The GAIN Data	83
4. The Individual Choice Problem	86
5. A Model of the Data	88
6. The Individual Decision Problem	99
7. The Social Choice Problem	105
8. What We Learn from GAIN	110
9. Conclusion	117
<i>References</i>	119
<i>Tables</i>	122
<i>Figures</i>	<i>after</i> 136

Acknowledgment

It is a great pleasure to acknowledge all those who have contributed to the work embodied in this thesis.

My thesis committee has been a tireless source of support. Gary Chamberlain and Guido Imbens have patiently helped me to develop not only the ideas embodied in this thesis, but also my understanding of econometrics. They offered their insight and encouragement throughout, without which this thesis would not have been possible. Edward Glaeser has been a constant source of advice and stimulus. Caroline Hoxby, Larry Katz, and Donald Rubin have read, commented on, and helped me immeasurably to improve my work.

I also have benefited from the academic support of Andreu Mas-Colell, whose concern and advice have been a source of strength, and of Dale Jorgenson and Eric Maskin, whose advice helped me find my way.

Many thanks are due to Sadek Wahba, who co-authored the first chapter of this thesis, and was instrumental to the second chapter by bringing the GAIN data to my attention. I have benefited both intellectually and personally from our collaboration.

I owe a great debt to the many who helped to create a congenial and stimulating intellectual atmosphere during my stay at Harvard. I would like to thank: Amartya Sen, for the opportunity to be involved with the Political Economy Lecture Series for the last three years; the Humble Lunch Committee (Jamil Baz, Ashish Garg, Tokuo Iwaisako, Athanasios Vamvakidis, and Sadek Wahba) for helping to create a relaxed and sympathetic oasis in our weekly lunches; Ellen DiPippo, Carrie Daniels, Stephanie Smith, and Kathy Wahl for their role in steering me through the various stages of my

degree; Terry Burham, Massimo Morelli, and Roberta Gatti for concern and interest in my work; and Kei Hirano for his sympathetic ear.

I have enjoyed the invaluable and irreplaceable support of many outside Harvard. First and foremost among these are my parents, Harsha Dehejia and Sudha Dehejia, and my brother, Vivek Dehejia, for their endless sympathy, support, and interest, and for lending me their strength. To name some is to exclude others, but with this apology it is also a pleasure to thank: Francesco Caselli and Thea Chiarini, Roberto Censolo, Sven Feldmann, Anselmo Tabbitt, José Tavares, Suzy Wahba, and Bruce Watson.

Finally there are those who deserve acknowledgment, but cannot be named, who are too many, too vast. My debt to these is profound.

Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs*

* Co-authored with Sadek Wahba

Abstract

The need to use randomized experiments in the context of manpower training programs and in analyzing causal effects more generally has been a subject of much debate. The paper draws on methods for causal inference developed in the statistics literature that extend the theory of classical randomized experiments to a non-experimental context when selection is on observable characteristics. In the context of selection bias, we use the propensity score –the probability of receiving treatment given observed characteristics – to reduce the dimensionality problem for a non-parametric estimation of the treatment effect. The paper provides a detailed discussion of the implementation of propensity score methods. We make use of the National Supported Work (NSW) Demonstration Program used by Lalonde (1986) to contrast the non-parametric estimates of the treatment effect with more conventional regression-based estimates such as those evaluated in Lalonde. The results demonstrate that the method closely replicates the experimental training effect, thus addressing the concerns raised about non-experimental techniques. The paper suggests that the techniques could be a useful complement to standard econometric tools for estimating causal relations.

1. Introduction

An important question when analyzing causal effects is how well non-experimental techniques of causal inference perform relative to experimental evaluations. For example, how accurately can a researcher hope to estimate the effect of a manpower training program on earnings in an observational study?¹ The question itself is not a new one. In economics the need to use the classical statistical methodology of randomized experiments in the context of manpower training programs and for causal effects in general, is addressed, *inter alia*, by Ashenfelter (1978), Ashenfelter and Card (1985), Burtless and Orr (1986) and more recently by Burtless (1995). Lalonde (1986) is the first study to evaluate directly standard econometric procedures used to evaluate training programs. He examines a randomized experiment (the National Supported Work Demonstration, NSW) from which he obtains an unbiased estimate of the training effect and then compares the experimental result to those obtained from the standard techniques. Specifically, a range of parametric selection models (estimated using least squares regressions, instrumental variables, and the Heckman two-step procedure) is applied to the observations that received training and a set of control observations constructed from population survey data sets (CPS and PSID). The results are then compared to the benchmark experimental estimate. The conclusion in Lalonde (1986), which has been very influential in labor economics and the evaluation of social

¹ The importance of classical experiments in explaining causal relations in econometrics goes beyond the case of training programs, used here as one possible application. See Cox (1992), Leamer (1978), and Pratt and Schlaifer (1988) for various perspectives on the role of randomization in economic analysis.

programs (e.g., Katz [1992]), is that these econometric techniques generally fail to replicate the experimentally determined results.

In this paper we draw on methods for causal inference developed in the statistics literature (Rubin [1974, 1977, 1978], Rosenbaum and Rubin [1983a], and reviewed in Holland [1986]) to exploit fully information contained in observable covariates. The approach extends the theory of classical randomized experiments (Fisher [1935] and Neyman [1935]) to a non-experimental context, using the key assumption of selection on observable characteristics. In the context of selection bias, we use the propensity score -- the probability of receiving treatment -- to reduce the dimensionality problem for a non-parametric estimation of the treatment effect. The paper provides a detailed discussion of how to estimate the propensity score and its use in estimating the treatment effect. The approach relates to previous work in econometrics (Barnow, Cain, and Goldberger [1980], and Goldberger [1972a,b]), but one important difference is that it recognizes explicitly the importance of defining the assignment mechanism and the pre-treatment variables that determine assignment. Furthermore, the approach highlights the need for overlap in the distribution of the covariates that determine selection for treated and control observations, even if assignment is fully understood. We contrast our non-parametric estimates of the treatment effect with more conventional regression-based estimates such as those evaluated in Lalonde (1986). The results demonstrate that the method closely replicates the experimental training effect thus addressing the concerns raised about non-experimental techniques.

The paper argues that the approach is flexible and self-diagnostic, allowing the researcher to assess the comparability of the distributions of the treated and control units and then correct the bias attributable to observable characteristics. For example, recent studies based on treatment-outcome models have examined alternative assumptions behind the selection process (e.g., Card and Sullivan [1988] who examine the effect of training on employment; Heckman et al., [1995] who estimate the effect of the JTPA training program on earnings; and Manski et al., [1992] who examine the impact of family structure on school enrollment). Even in settings where the selection on observables assumption is not adequate, the techniques described in the paper should be seen as an important and practical complement to econometric methods such as: instrumental variables that depend on well specified exclusion restrictions (e.g., Angrist [1990] and Imbens and Angrist [1994]); and assumptions on the distribution of unobserved characteristics (Heckman [1979]).

The paper is organized as follows. Section 2 briefly reviews the literature on causal effects and selection bias in economics. Section 3 identifies the treatment effect under the causal effect model, and Section 4 discusses estimation procedures for the propensity score and the treatment effect. In Section 5, we give an overview of the NSW experiment and summarize Lalonde's results. In Section 6, we implement the approach of Sections 3 and 4, and in Section 7, we discuss the sensitivity of the results to the methodology. Section 8 concludes the paper.

2. Review of the Literature

The econometric formulation of a causal relation such as evaluating the impact of a training program on earnings can be represented by the following model of self-selection (e.g., Maddala [1983] and more recently Heckman [1990]):

$$\begin{aligned} Y_{i1} &= \alpha_1 + X_i\mu_1 + u_{i1} && \text{for participants} \\ Y_{i0} &= \alpha_0 + X_i\mu_0 + u_{i0} && \text{for non-participants,} \end{aligned} \quad (1)$$

where Y_{i1} and Y_{i0} are the outcome variables of interest which are a linear function of a vector of observable characteristics X and some error term. The following participation decision rule,

$$\begin{aligned} T_i &= \begin{cases} 1 & \text{if } T_i^* \geq 0 \\ 0 & \text{if } T_i^* < 0 \end{cases} \\ T_i^* &= \pi_0 + X_i\pi + v_i, \end{aligned} \quad (2)$$

determines whether individual i participates in the program ($T_i=1$) or not ($T_i=0$), T^* being a linear function of observable characteristics X . The error terms are distributed as:

$$\begin{pmatrix} u_{i1} \\ u_{i0} \\ v_i \end{pmatrix} \Bigg| X \text{ i.i.d. } N(0, \Sigma), \quad (3)$$

where Σ is the variance matrix, some of the covariance terms being defined according to specific assumptions outlined below. The case that has received the most attention in the training literature, and more generally in evaluating causal effects, is a constant additive treatment effect which defines the observed outcome variable Y_i as follows,

$$Y_i = \delta + Y_{i0}, \quad (4)$$

where δ is the effect due to treatment. The equation for the observed outcome variable can then be written as:

$$\begin{aligned}
 Y_i &= \beta_0 + \delta T_i + X_i \beta_1 + \varepsilon_i \\
 \text{where} \\
 Y_i &= T_i Y_{i1} + (1 - T_i) Y_{i0} \\
 \beta_0 &= \alpha_0 \\
 X_i \beta_1 &= X_i \mu_0 = X_i \mu_1 \\
 \delta &= \alpha_1 - \alpha_0 \\
 \varepsilon_i &= u_{i1} = u_{i0}.
 \end{aligned} \quad (5)$$

In this model, selection bias when estimating the treatment effect occurs when there is a dependence between the assignment and the error term due to the dependence of ε_i and v_i (see Heckman and Robb [1985], who refer to this selection as selection on unobservables). Solutions to the selection problem range from exclusion restrictions, where some variables are excluded from one of the two equations (1 and 2) to obtain instrumental variables, to assumptions on the joint distribution of the unobservables.

Most of the recent attention has typically focused on these types of restrictions to identify the treatment effect (see Manski [1995] for a recent review of the issues). This paper focuses on evaluating econometric methods that rely on the selection on observable characteristics assumption. The selection on observable characteristics assumption can be expressed as:

$$\begin{aligned} &\text{If } \varepsilon_i \perp\!\!\!\perp v_i, \text{ then } T_i \perp\!\!\!\perp \varepsilon_i | X_i \\ &\text{which leads to} \\ &E(Y_i | T_i, X_i) = \beta_0 + \delta T_i + X_i \beta_1 \end{aligned} \tag{6}$$

The case of selection on observable characteristics was considered first by Goldberger (1972a) and further developed in Barnow, Cain, and Goldberger (1980). Under the above assumptions, equation (6) can be estimated using least squares to obtain an unbiased estimate of the treatment effect δ .

The previous estimation of the treatment effect under selection on observable characteristics assumed a constant additive treatment effect specified in equation (4). One of the conclusions reached by Lalonde (1986) was that *as it stands* the model does not perform well when tested against an experimental data set. However linearity and additivity of the treatment effect are not necessary assumptions and indeed, following Goldberger (1972b) one can relax these restrictions. Under the selection on observables assumption we can rewrite equation (5) as:

$$E(Y_i|T_i, X_i) = \varphi_0 + \delta T_i + X_i \varphi_1 + T_i X_i \varphi_2$$

where

$$\varphi_0 = \alpha_0 \tag{7}$$

$$\delta = \alpha_1 - \alpha_0$$

$$\varphi_1 = \mu_0$$

$$\varphi_2 = (\mu_1 - \mu_0).$$

Equation (7) defines more than one treatment effect. In non-experimental settings where data on the control group is either a self-selected sample or, in some cases (as in Lalonde's paper) is drawn from an altogether different population, the distribution of the observable characteristics between treated and control units need not overlap very much. In this case, estimating treatment effects through models such as equation (5) amounts to extrapolating between two very different groups. If the groups are sufficiently different such an extrapolation can be extremely misleading, as will be demonstrated.

Another related dimension along which equations (5) and (7) can be relaxed is through higher order and interaction terms of the covariates X_i . Although, in principle it is a flexible approach for estimating the treatment effect, estimating such a model when X_i is multi-dimensional (and includes many continuous variables) is an econometric (non-parametric) problem of a high order of difficulty.² Simply saturating a regression with higher order and interaction terms would quickly exhaust the number of observations available, and which interaction terms to include (or exclude) is an

² Note that even if all variables are dichotomous with k variables, the total number of interaction terms is 2^k . With dichotomous X variables, the function simplifies to matching observations on the covariates. For an example as well as more discussion see Angrist (1995).

issue that increases in complexity as the number of possible terms increases exponentially (see [Hardle \[1990\]](#) on the so-called curse of dimensionality). A systematic method to make these choices is required.

The methodology we follow in this paper pursues the selection on observables approach. In particular, by expressing the causal relation in terms that recognize explicitly the assignment mechanism, the pre-treatment covariates that determine assignment, and the overlap in the distribution of covariates between the treatment and control groups, we estimate the treatment effect with relatively weak assumptions on functional form and distribution. We present this method in the next section.

3. Selection on Observables and the Role of the Propensity Score

3.1 Causality and the Role of Randomization

To define the framework used in this paper we first formalize the notion of causality. A cause is viewed as a manipulation or treatment which brings about a change in the variable of interest as compared to some baseline called the control. The fundamental problem of estimating a causal effect from a given sample is that for any unit the variable of interest is observed under either the treatment or control, but never both. The role of randomization is to allow for unbiased estimation of the average value of the variable of interest over the whole population under both the treatment and control.³

³ An important assumption is that the conditional expectation of the outcome for unit i does not depend on the treatment status of other units. Otherwise we would have to condition throughout on the entire vector of treatment assignments. This is referred to in the statistical literature as the stable unit treatment value assumption (SUTVA) ([Holland 1986](#) and [Rubin 1978](#)) or more generally known as the contamination problem.

Formally, let i index the units under consideration. Then as in the previous section Y_{i1} is the value of the variable of interest when the unit i is subject to treatment (1), and Y_{i0} is the value of the variable of interest when the unit is exposed to treatment (0) (called control). As compared to Y_i , defined in (5) as the observed value of the outcome variable, only one of Y_{i0} or Y_{i1} is observed for any i . The treatment effect for a single unit, τ_i , is defined by: $\tau_i \equiv Y_{i1} - Y_{i0}$. The treatment effect of interest is the expected treatment effect over the population, hence:

$$\begin{aligned} \tau &\equiv E(\tau_i) \\ &= E(Y_{i1}) - E(Y_{i0}) \\ &= E(Y_{i1}|T_i = 1) \cdot p(T_i = 1) + E(Y_{i1}|T_i = 0) \cdot p(T_i = 0) \\ &\quad - [E(Y_{i0}|T_i = 0) \cdot p(T_i = 0) + E(Y_{i0}|T_i = 1) \cdot p(T_i = 1)], \end{aligned} \quad (8)$$

where $T_i=1$ ($= 0$) if the i -th unit was exposed to treatment (control). The problem of unobservability is summarized in the fact that we can only estimate $E(Y_{i1}|T_i=1)$ and $E(Y_{i0}|T_i=0)$, hence:

$$\tau^e = E(Y_{i1}|T_i = 1) - E(Y_{i0}|T_i = 0) \quad (9)$$

which, by simply comparing equations (8) and (9), indicates that τ^e is potentially a biased estimate of τ . Intuitively, if the treated and control units systematically differ in their characteristics, then in observing only the treated group do we not correctly estimate Y_{i1} for the whole population ($E(Y_{i1}) \neq E(Y_{i1}|T_i = 1)$) and likewise using the

control group for Y_{i0} . The role of randomization is precisely that the outcome variables are independent of assignment to treatment or control, so that:

$$Y_{i1}, Y_{i0} \perp\!\!\!\perp T_i \quad (10)$$

and therefore

$$E(Y_{i1}|T_i = 1) = E(Y_{i1}|T_i = 0), E(Y_{i0}|T_i = 0) = E(Y_{i0}|T_i = 1).$$

Thus, the treated and control groups do not systematically differ from each other in Y_{i1} and Y_{i0} (ignorable treatment assignment) making the conditioning on T_i in the expectation unnecessary, and yielding:

$$\tau^e = \tau = E(Y_{i1}|T_i = 1) - E(Y_{i0}|T_i = 0) = E(Y_i|T_i = 1) - E(Y_i|T_i = 0) \quad (11)$$

3.2 The Role of the Propensity Score

The extension of the classical randomized framework to a non-experimental setting when selection to a treatment occurs on observable characteristics is attributed to Rubin (1974, 1977, 1978) and referred to as the “potential outcomes” approach (since one estimates what would have been the outcome had the unit not received the treatment while observing only one of the outcomes). The analytical contribution of the approach is in identifying a transparent set of assumptions that parallel those of a classical experiment enabling the researcher to obtain unbiased estimates of the treatment effect.

In non-experimental studies, data is typically available only on a treated group made up of a systematic sub-sample of the population (e.g., volunteers). The control group is either a systematic sub-sample of the population (e.g., those who did not volunteer) or it may not have been collected alongside the treatment group and may have to be created by turning to other data sets. In the context of medical records, data are often maintained on patients who were not chosen for treatment; likewise in economics, potential controls are often available through periodic population surveys. In these cases, it is the treated group which is the population of interest.⁴ The treatment effect is then defined as:

$$\tau_{T=1} = E(Y_{11}|T_i = 1) - E(Y_{10}|T_i = 1). \quad (12)$$

However, equation (12) is not identified since Y_{10} is never observed for units with $T_i=1$. Estimating $\tau_{T=1}$ by equation (9) will yield a biased estimate since it can be rewritten as:

$$\tau^* = \tau_{T=1} + [E(Y_{10}|T_i = 1) - E(Y_{10}|T_i = 0)] \quad (9')$$

⁴ A less natural but consistent case would be to use a treated group to estimate the causal effect for a given control population of interest. Note that in the setting of a randomized experiment (10), the treatment effect for the treated population is identical to the treatment effect for the untreated population: $\tau_{T=1} = \tau_{T=0} = (E(Y_1|T_i=1) - E(Y_1|T_i=0))$.

The earnings of the control group sample may not be representative of the treated individuals' earnings had they not received training. The last term would drop only if treatment was randomly assigned. Identification is possible under the assumption of "ignorable assignment conditional on covariates", i.e., assignment to treatment or control is a (stochastic) function of a vector of (observable) covariates. In such a case, conditional on the vector X , the assignment mechanism is like a randomized experiment (Rubin [1977]):

Proposition 1: *If for each unit we observe a vector of covariates X_i and*

$$Y_{i,1}, Y_{i,0} \perp\!\!\!\perp T_i | X_i, \forall i,$$

then:

$$\begin{aligned} \tau_{T=1} &\equiv E(Y_{i1} | T_i = 1) - E(Y_{i0} | T_i = 1) \\ &= E_x \{ E(Y_i | X_i, T_i = 1) - E(Y_i | X_i, T_i = 0) | T_i = 1 \}, \end{aligned}$$

where $Y_i = T_i Y_{i1} + (1 - T_i) Y_{i0}$.

Proof:

$$Y_{i1}, Y_{i0} \perp\!\!\!\perp T_i | X_i$$

$$\Rightarrow E(Y_{i1} | X_i, T_i = 1) = E(Y_{i1} | X_i, T_i = 0) = E(Y_{i1} | X_i),$$

and similarly for Y_{i0} , which allows us to write:

$$\begin{aligned}
\tau_{T=1} &= E(Y_n|T_i = 1) - E(Y_{i0}|T_i = 1) \\
&= E_X \left\{ E(Y_n|X_i, T_i = 1) - E(Y_{i0}|X_i, T_i = 1) \middle| T_i = 1 \right\} \\
&= E_X \left\{ E(Y_n|X_i, T_i = 1) - E(Y_{i0}|X_i, T_i = 0) \middle| T_i = 1 \right\} \\
&= E_X \left\{ E(Y_i|X_i, T_i = 1) - E(Y_i|X_i, T_i = 0) \middle| T_i = 1 \right\} \\
&= E_X \left\{ (\tau_{T=1,X}) \middle| T_i = 1 \right\},
\end{aligned}$$

where $\tau_{T=1,X} \equiv E(Y_i|X_i, T_i = 1) - E(Y_i|X_i, T_i = 0)$.

Intuitively, this assumes that conditional upon observed covariates X , assignment to treatment T , can be taken to be random (conditionally ignorable treatment assignment); comparing two individuals with the same observable characteristics, one of whom was treated and one of whom was not is like comparing those two individuals in a randomized experiment. Taken literally, the notion of conditioning corresponds to matching or grouping the observations according to the value of the covariate X . The proposition is equivalent to the model in equation (7) in which all covariates X interact with the treatment effect. Note however the difference in emphasis; in this approach the assignment mechanism must be explicitly defined as well as the covariates that determine assignment. Furthermore, the treatment of interest is defined in terms of potential outcomes for treated individuals: what would have been the outcome had they not received the treatment?

The limitation of Proposition 1 is in the estimation of the treatment effect: it relies on a sufficiently simple set of discrete covariates to keep the task of conditioning on the exact value of X a tractable exercise. If there are k dichotomous covariates, the

number of possible combinations will be 2^k . As the number of variables increase, the number of matching cells will increase exponentially. Rosenbaum and Rubin (1983a) suggest the use of the propensity score, the probability of receiving a treatment given a set of covariates, to reduce the dimension problem associated with implementing the conditioning strategy following Proposition 1:

Proposition 2: Let $p(X_i)$ be the probability of unit i having been assigned to treatment, defined as $p(X_i) = Pr(T_i=1|X_i) = E(T_i|X_i)$, where $0 < p(X_i) < 1, \forall i$. Then:

$$(Y_{1i}, Y_{0i}) \perp\!\!\!\perp T_i | X_i$$

implies

$$(Y_{1i}, Y_{0i}) \perp\!\!\!\perp T_i | p(X_i).$$

Proof:

$$\begin{aligned} E(T_i | Y_1, Y_0, p(X)) &= E_X \{E(T_i | Y_1, Y_0, X) | Y_1, Y_0, p(X)\} \\ &= E_X \{E(T_i | X) | Y_1, Y_0, p(X)\} \\ &= E_X \{p(X) | Y_1, Y_0, p(X)\} \\ &= p(X). \end{aligned}$$

Hence,

$$\Rightarrow T \perp\!\!\!\perp Y_1, Y_0 | p(X).$$

Corollary 2.1: $\tau_{|T=1} = E_{p(X)} \{E(Y_i | T_i = 1, p(X_i)) - E(Y_i | T_i = 0, p(X_i)) | T_i = 1\}.$

Proof:

$$\begin{aligned}
\tau|_{T=1} &= E(Y_{11}|T_i = 1) - E(Y_{10}|T_i = 1) \\
&= E_{p(x)} \left\{ E(Y_{11}|T_i = 1, p(X_i)) - E(Y_{10}|T_i = 1, p(X_i)) \mid T_i = 1 \right\} \\
&= E_{p(x)} \left\{ E(Y_{11}|T_i = 1, p(X_i)) - E(Y_{10}|T_i = 0, p(X_i)) \mid T_i = 1 \right\} \\
&= E_{p(x)} \left\{ E(Y_i|T_i = 1, p(X_i)) - E(Y_i|T_i = 0, p(X_i)) \mid T_i = 1 \right\} \\
&\equiv E_{p(x)} \left\{ \tau|_{T=1, p(X)} \mid T_i = 1 \right\}.
\end{aligned}$$

Thus, the conditional independence result extends to the use of the propensity score, as does by immediate implication our result on the computation of the conditional treatment effect, now $\tau|_{p(x)}$. The achievement of the score is that it reduces the dimensionality of the problem substantially, requiring only matches on a univariate scale rather than in a space of dimension equal to the dimension of X .

Proposition 2 essentially reduces the exercise of estimating the treatment effect to estimating the following two non-parametric functions:

$$E(Y_{11}|p(X_i)) = E(Y_{11}|T = 1, p(X_i)) \quad (13)$$

$$E(Y_{10}|p(X_i)) = E(Y_{10}|T_i = 0, p(X_i)), \quad (14)$$

which would be univariate non-parametric regressions if the propensity score were known. Since we are estimating the average treatment effect for the treated population and Y_{11} for the treated population is known, in effect we do not have to estimate equation (13). The next section discusses how to estimate the propensity score in a

multi-dimensional, non-parametric regression and then to use the estimated score in a uni-dimensional non-parametric regression for the treatment effect. The contention is that this is an easier approach than taking on the full non-parametric regression implied by Proposition 1, or alternatively by equation (7).⁵

4. Estimating the Average Treatment Effect

4.1 Estimating the Propensity Score

The first step in estimating the treatment effect is to estimate the propensity score. Any standard probability model can be used, e.g., logit or probit. It is important to remember that the role of the score is only in reducing the dimensions of the conditioning, and, as such, it has no behavioral assumptions attached to it. For ease of estimation, most applications in the statistics literature have concentrated on a logistic equation:

$$\Pr(T_i = 1 | X_i) = \frac{e^{\lambda h(X_i)}}{1 + e^{\lambda h(X_i)}}, \quad (15)$$

where T_i is the treatment status, and $h(X_i)$ is made up of linear and higher order terms of the covariates on which we condition to obtain an ignorable treatment assignment.⁶

⁵ The use of the propensity score in estimating the treatment effect has been discussed briefly in econometric contexts (Heckman and Robb 1986:100-104), but to our knowledge has not been explored fully in an economic context. For a recent exception see Heckman et al., (1995).

⁶ Because we allow for higher order terms in X , this choice is not very restrictive. By re-arranging and taking logs, we obtain: $\ln(\Pr(T_i=1|X_i)/\Pr(T_i=0|X_i)) = \lambda h(X_i)$. A Taylor series expansion allows us an arbitrarily precise approximation. See also Rosenbaum and Rubin (1983a).

If the propensity score is used as a mechanism to reduce the dimensions of the estimation problem, it is also true that in estimating the score through equation (15) the choice of which interaction or higher order term to include is determined solely by the need to condition fully on the observable characteristics that make up the assignment mechanism. This is formally represented in the following proposition due to Rosenbaum and Rubin (1983a).

Proposition 3:

$$X \perp\!\!\!\perp T \mid p(X).$$

Proof: From the definition of $p(X)$ in proposition 2:

$$E(T_i | X_i, p(X_i)) = E(T_i | X_i) = p(X_i).$$

Though elementary, Proposition 3 is fundamental in providing a framework to validate estimates of the propensity score and hence in choosing which higher order and interaction terms to use. For equal values of the propensity score, the covariates are also on average balanced, i.e., observations with the same propensity score also have the same distribution of covariates; as in a random experiment, observations assigned to a treatment by the flip of a coin would vary individually across treatment but not on average for a large sample. This provides an easy diagnostic for how well the score has been estimated.

The algorithm we propose works as follows. Starting with a parsimonious logistic function with linear covariates to estimate the score, rank all observations by

the estimated propensity score (from lowest to highest). Divide the observations into strata such that within each stratum or block the difference in score for treated and control observations is insignificant (a t-test on difference of means of treated and control is a criterion followed in this algorithm). Proposition 3 tells us that within each stratum the distribution of the covariates should be approximately the same across the treated and control groups once the score is controlled for. Within each stratum, we can test for statistically significant differences between the distribution of covariates for treated and control units; operationally, t-tests on differences in the first moments are often sufficient but a joint F-test for the difference in means for all the variables within each block could also be performed.⁷ When the covariates are not balanced within a particular block, the block may be too coarsely defined; recall that Proposition 3 in fact deals with observations with an identical propensity score. The solution adopted is to divide the block into finer blocks and test again for no difference in the distribution of the covariates within the finer blocks. If however some covariates remain unbalanced for many blocks, the score may be poorly estimated, which suggests that additional terms (interaction or higher order terms) of the unbalanced covariates should be added to the logistic specification to control for these characteristics better. This procedure is repeated for each given block until covariates are balanced.⁸ The algorithm is summarized in Figure 1.

⁷ More generally one can also consider higher moments or interactions, but usually there is little difference in the results.

⁸ Cochran (1968) and Rosenbaum and Rubin (1985a) show that under certain restrictions, including normality of covariate distribution, five equal size blocks reduce 95 percent of the bias. Although these results can be taken as a benchmark for the number of blocks that would reduce most of the bias from differences in the distribution of the covariates, ultimately the blocking is a function of the overlap between the distribution of the score for treated and control samples as well as sample size. It is also

Figure 1 - A Simple Algorithm for Estimating the Propensity Score

- Start with a parsimonious logit function to estimate the score.
- Sort data according to estimated propensity score (ranking from lowest to highest).
- Stratify all observations such that estimated propensity scores within a stratum for treated and control are close (no significant difference); e.g. start by dividing observations in blocks of equal score range (0-0.2, ..., 0.8-1).
- Statistical test: difference-in-means for all covariates of treated and control in all blocks are not significant from zero at relevant confidence level.
 1. If covariates are balanced between treated and control observations for all blocks, stop.
 2. If covariate i is not balanced for some blocks; divide block into finer blocks and re-evaluate.
 3. If covariate i is not balanced for all blocks, modify logit by adding interaction terms and/or higher order terms of covariate i , and re-evaluate.

A key property of this estimation procedure is that it uses a well-defined criterion to determine which interaction terms to use in the estimation, namely those terms which balance the covariates. It makes also no use of the outcome variable, and embodies one of the specification tests proposed by Lalonde (1986) and others in the context of evaluating the impact of training on earnings, namely to test for the

evident that the simple stratification procedure adopted here is not the only way of determining the number of blocks that balance the covariates; other non-parametric techniques could be used such as kernel or nearest neighbor (see Hardle 1990).

regression adjusted difference in the earnings prior to treatment.⁹ Once the propensity score is estimated the treatment effect can be obtained in a number of ways.

4.2 Estimating Treatment Effect by Stratifying on the Score

Equations (13) and (14) that define the treatment effect are estimated as univariate non-parametric regressions, amounting to methods that use the propensity score non-linearly. The first estimator of the treatment effect adopts the stratification procedure used for estimating the propensity score to estimate in turn the treatment effect. The procedure described in Figure 1 generates a series of blocks, within which the score is approximately equal for all units, and within each block assignment to treatment is ignorable (Proposition 2 and Corollary 2.1). Within each block, the treatment effect, $E(Y_{i1}|T_i=1,p(X_i))-E(Y_{i0}|T_i=0,p(X_i))$, is the difference of two expectations that are a function of observables. Estimating $\tau|_{T=1,p(\alpha)}$ requires only point estimates of each term, and like a randomized experiment the difference of means is an (approximately) unbiased estimator. Stratifying on the score ensures an overlap in the distribution of the treated and control groups; within each block very little modeling is required, and choice of functional forms is no longer a major issue. This procedure can be summarized in the following general form:

⁹If the model used to estimate the treatment effect, e.g. equation (5) effectively controls for observable characteristics, it must be true that these characteristics cannot predict pre-treatment earnings. But as indicated in figure 1 the purpose of the procedure is precisely to control for pre-treatment characteristics whether by taking the difference in means or by regression adjustment. In some sense the algorithm is just a specification test.

$$\begin{aligned}
E(Y_i | p(X_i), T_i = 1) &\approx \sum_{q=1}^Q h_{q1}(p(X_i)), \\
E(Y_i | p(X_i), T_i = 0) &\approx \sum_{q=1}^Q h_{q0}(p(X_i)),
\end{aligned} \tag{16}$$

where q indexes the blocks (defined over intervals of the propensity score, (p_q, p^q)) and $h_{q1}(p(X_i))$ and $h_{q0}(p(X_i))$ are the functional forms used to model the conditional expectations within each block for treated and control respectively. The simplest model is a step-function function form:

$$h_{q1}(p(X_i)) = \begin{cases} \beta_{q1} & \text{if } p(X_i) \in (p_q, p^q), \\ 0 & \text{otherwise,} \end{cases} \tag{17}$$

and likewise for $h_{q0}(p(X_i))$. Other possibilities include linear within each block, as well as continuous and differentiable specifications such as piece-wise linear regressions. However, as suggested above, since ultimately we use these specifications simply to generate a point estimate of the expectation within each block, further modeling of $h(\cdot)$ at most reduces the residual bias. This approximation is then averaged within each block to obtain a point estimate of $\tau|_{T=1,q}$ and over each block to obtain $\tau|_{T=1}$:

$$\begin{aligned}
E(Y_{1i} | T_i = 1) &= E_{p(X)} \left\{ E[Y_{1i} | p(X_i), T_i = 1] | T_i = 1 \right\} \\
&= \int E(Y_i | p(X_i), T_i = 1) dF(p(X_i) | T_i = 1) \\
&\approx \int \sum_{q=1}^Q h_{q1}(p(X_i)) dF(p(X_i) | T_i = 1),
\end{aligned} \tag{18}$$

and similarly for $E(Y_i|T_i = 0, p(X))$.

The estimates are the sample analogues to the various $h(\cdot)$ functions, suitably weighted. For example, with the step function, estimating $h_{q1}(\cdot)$ amounts to estimating the mean of Y_{i1} for the treated sample within each block and then averaging over blocks, weighting by the number of treated units:

$$\hat{Y}_1 - \hat{Y}_0|_{T_i=1} = \sum_{q=1}^Q \left(\frac{\sum_{i \in I(q)} T_i}{\sum_{\forall i} T_i} \right) \left[\frac{\sum_{i \in I(q), T_i=1} Y_i}{\sum_{i \in I(q)} T_i} - \frac{\sum_{i \in I(q), T_i=0} Y_i}{\sum_{i \in I(q)} (1 - T_i)} \right]. \quad (19)$$

Note that the first term simplifies to the average earnings for treated units which is directly obtained from the data. However, for Y_{i0} this simplification is not possible and the earnings are weighted by the number of treated observations in block q . The step function could easily be replaced by $h(\cdot)$ linear in the propensity score. The sample analogue would be a linear regression of Y_{i1} (likewise Y_{i0}) on the propensity score *within* each block, and the point estimate would be obtained using the predicted values of the regression, suitably weighted. As we will see the benefits of such additional modeling are minimal, conditional on the overlap within a block.

4.3 Matching on the Score

The second estimation strategy that follows also from corollary 2.1 is through pair-wise matching on the score. The conditioning on the propensity score is implemented by matching techniques that pair each treated unit to the single control unit with the closest propensity score. Matching techniques have generally been used when the number of potential comparable units is much larger than the number of treated units. In such cases, one matches the treated units with a smaller number of selected control observation to produce comparable matches and reduce bias. Once each treated unit is matched with a control unit, unmatched control units are discarded and we confine ourselves to the reduced sample.¹⁰ The matched sample will have the property that the distribution of observed covariates for the treated and control groups is approximately the same. Given the assumption of conditional ignorability (Proposition 2), the treatment effect is estimated by taking the difference in means or using least squares adjustment appropriately weighted, to correct for any remaining imbalance. Matching procedures are not straightforward, because rarely do two treated and control units have an *identical* score. An algorithm for determining how to match units within some bands of tolerance on the inexactitude of the match needs to be specified and there are delicate issues regarding the order in which to match the treated units (see Rosenbaum and Rubin [1985b], and Rubin and Thomas [1992]). Results reported in this paper for

¹⁰ In some sense matching is an extreme form of stratification where each treated observation is in a separate stratum, which is sufficiently narrow to include only one control. Note that control units would be thrown out in a matching exercise even if they were previously included in blocks determined by the score.

matching follow a procedure that accounts for the minimal overlap between the treated and control distributions by allowing a given control unit to match with more than one treated unit (see Dehejia and Wahba [1995]).¹¹

4.4 Using the Score as Weights

The previous two estimation strategies use grouping mechanisms to condition on the score and estimate the treatment effect. The score itself does not enter directly in the estimation of the treatment effect but enters indirectly by determining the weights in equation (19). Alternatively, the score could be used directly as weights in estimating the average treatment effect. The following proposition demonstrates that using the score as weights yields a consistent estimator of the treatment effect:

Proposition 4:

$$\frac{1}{N^T} \sum_{i=1}^{N^T+N^C} \left(T_i Y_i - (1-T_i) \frac{p(X_i)}{1-p(X_i)} Y_i \right)$$

is a consistent estimator of $\tau|_{T=1}$, where Y_i is the observed value of outcome for unit i , T_i is an index variable ($=1$ if treated and $=0$ if control), $0 < p(X_i) < 1$, and N^T and N^C are the total number of treated and control observations respectively.

¹¹ Early work on matching revolved around matching on a covariate or a set of covariates. For a theoretical examination of matching on a set of covariates see Rubin (1973, 1979). Applications of matching on the propensity score have generally been limited to the biomedical field (e.g. Reinisch et al. [1993]). The only application we know of in economics is Duggan (1994).

Proof:

In the population consider θ , where:

$$\begin{aligned}
 \theta &\equiv E\left(T_i Y_i - \frac{p(X_i)}{1-p(X_i)}(1-T_i)Y_i\right) \\
 &= E_x\left(E(T_i Y_i | X_i) - \frac{p(X_i)}{1-p(X_i)}E((1-T_i)Y_i | X_i)\right) \\
 &= E_x\left(p(X_i)E(Y_{11} | X_i) - p(X_i)E(Y_{10} | X_i)\right) \\
 &= \int p(X_i)(E(Y_{11} | X_i) - E(Y_{10} | X_i))f(X_i)d(X_i) \\
 &= \int \Pr(T_i = 1)(E(Y_{11} | X_i) - E(Y_{10} | X_i))f(X_i | T_i = 1)d(X_i) \\
 &= \Pr(T_i = 1)E_{x|T_i=1}(E(Y_{11} - Y_{10} | X_i)).
 \end{aligned}$$

Thus, $\tau_{|T_i=1} = \theta / \Pr(T_i = 1)$, and the sample analogue of τ is:

$$\frac{1}{N^T} \sum_{i=1}^{N^T + N^C} \left(T_i Y_i - \frac{p(X_i)}{1-p(X_i)}(1-T_i)Y_i \right).$$

The estimator differs from the first two to the extent that the objective of the causal model adopted in this paper is to be agnostic about which functional form needs to be assumed. In using the score as a weight results will be more sensitive to score estimation. However the power of proposition 4 is that it does not rely on stratification procedures (blocking or matching) to control for observables as in Proposition 2 and its corollary. A very different estimate of the treatment obtained when using weighting, compared to using stratification or matching, would suggest that either the score is misspecified (an issue that can be corrected through additional interaction terms and higher

order terms) or that the treatment effect is not ignorable conditional on the score. In this sense, proposition 4 is an additional self-diagnostic test.

5. The Data

5.1 The National Supported Work Program

The NSW was a federally-funded program with the objective of providing work experience for a period not exceeding twelve months to individuals that had faced economic and social problems prior to enrollment in the program.¹² Four groups were targeted: women on Aid to Families with Dependent Children (AFDC), former addicts, former offenders, and young school dropouts (but for reasons of space, this paper will only use the sample of male participants). Candidates were selected first on the basis of eligibility criteria, then at each site were randomly assigned either to the training program or excluded from it (MDRC [1983,5-9]). Those randomly selected to join the program were assigned to ten demonstration sites across the country and participated in various work ranging from operating a restaurant to building and operating a child care center. Unlike typical clinical trials, the eligible candidates did not all join the NSW program at once, but were screened and subsequently randomized over a period of 51 months between March 1975 when the program started and June 1977 when the last participants were accepted. Information on pre-treatment earnings as well as social characteristics was obtained from initial surveys as well as social security administration records. Finally, both control and treatment groups were given

¹² Several reports document extensively the NSW program. For a general summary of the findings see MDRC (1983).

follow-up interviews at specific intervals, with the last interview 36 months after assignment.

A total of 6,616 individuals were considered eligible for the program and 3,214 were assigned to the program whereas 3,402 served as a control group. However the final number of observations available with complete earnings profile as well as information on background characteristics was 540 observations for the treatment group and 554 for the control group.¹³ A final feature of the NSW design that requires mentioning relates to the month of assignment (moa) to the program, defined as the number of months in training prior to the end of the program (January 1978=0). Because randomization of the eligible individuals across treatment and control was not carried out simultaneously at one point in time (e.g. at the beginning of the program), but rather occurred throughout the entire period between March 1975 and July 1977 as individuals presented themselves to the administrators of the program, this created what has been referred to as the "cohort phenomenon" (MDRC [1983:48]). Individuals joining early in the program may not have had the same characteristics as those entering later. Indeed, the period 1974-1975 was characterized by high unemployment

¹³ Because of cost considerations not all participants were given follow-up interviews, which therefore reduced the number of final observations available and which explains the imbalance between treated and control observations, a smaller percentage of the latter group having been randomly interviewed for the follow-up interviews. As explained in MDRC (1983), the choice of which subset of participants to re-interview was randomized, and so does not affect the experimental design. The number was further reduced because of no-response from some of the interviewees. Although this could potentially affect the effect of randomization through selection bias, Brown (1979) discusses the effect of no-response and finds little evidence that it affects the results of the program. Finally, dropouts from those assigned to the treatment program were not significant (0.008 percent of total treated individuals) to confound the interpretation of the treatment effect.

rates compared to the following period, which suggests that the composition as well as the overall treatment effect may differ across cohorts.

5.2 The Sample

To control for the cohort phenomenon Lalonde (1986) limits his sample to those with a month of assignment (moa) between January 1976 (moa=24) and July 1977 (moa=6). This further reduces the sample to 297 treated observations and 425 control observations. It is important to note that selection of a sub-group based on month of assignment does not affect the properties of the experimentally randomized data set; by virtue of randomization the treated and control groups do not differ systematically along any selected pre-treatment variable, including moa. However this procedure does alter the benchmark training effect obtained by comparing the post-training earnings of the treated with the control observations.

The analysis in Lalonde (1986) uses only one year of pre-treatment earnings. But as Ashenfelter (1978), Ashenfelter and Card (1985), Card and Sullivan (1988) and others indicate, the use of several years of earnings is key in estimating the variability in the training effect. Since the methodology of Section 3 relies on fully exploiting the selection on observables assumption, we obtain additional information on the earnings profile of the participants, and test the sensitivity of the methodology to selection on observables, specifically pre-treatment earnings. From the sample of 297 treated and 425 control, using month-of-assignment we exclude those observations for which earnings in calendar 1974 could not be obtained thus arriving at a reduced sample of

185 treated observations and 260 control observations.¹⁴ Table 1 provides the characteristics of the sample used by Lalonde (297 treated and 425 control) as well as the reduced sample used throughout this paper (185 treated observations and 260 control). First, it demonstrates the effect of randomization on the pretreatment covariates of the treated and control groups. As expected, since both samples are chosen based on objectively defined pre-treatment covariates, the mean of the covariates across the treatment and control is similar and the differences are not statistically significant.¹⁵

5.3 Estimating the Treatment Effect: The Lalonde Results

Table 2 (rows 1 and 2) presents the training effect using the two randomized samples of Table 1 and the training effect for groups selected by sample characteristics for the second sample only. The treatment effect reported in column 3 is estimated by taking the difference-in-means (hereafter referred to as “unadjusted” treatment effect) of 1978 earnings of the treated and control groups. The relatively high standard error of the treatment effect of \$886 (s.e. = 476) suggests the possibility of heterogeneity of the treatment effect among units. The higher treatment effect for the second sample

¹⁴ The variables from the NSW data were originally expressed in experimental time. Lalonde (1984) explains how the data is converted to calendar time so that earnings are comparable to earnings from other data sets such as the PSID or CPS. Using the month of assignment variable (moa=0 in 1978), and the variable on earnings 24 months prior to assignment earnings for calendar 1974 are derived. For more details on how the sample was derived see Dehejia and Wahba (1994).

¹⁵ Table 1 replicates Lalonde Table 3 (1986:608) for the experimental data. The two samples do, however, differ from each other. For example, earnings in 1975 are substantially lower for our sample than for the entire sample (\$ 1,532 against \$ 3,066), as well as for the earnings two years prior to assignment (which is equivalent to earnings in 1974 for the second sample). As expected the average month of assignment increases for the second sample, 18.5 compared to 16.5 for the first sample. These differences simply reflect the “cohort phenomenon” noticed by the designers of the NSW program and do not compromise the validity of a simple comparison of sample means as causal estimates.

(\$1794) is a reflection of the cohort phenomenon as explained previously. Within the second sample, the variation in training effect is indicated in column (4).¹⁶ Those participants, for example, that completed high school or that have more than 11 years of schooling have a treatment effect that is much higher than the average (\$3,085), and significantly different from the treatment effect of their complement (nodegree or less than eleven years of schooling). Unemployment in 1974 is an important covariate that distinguishes participants (treatment effect is \$3,376 for those unemployed in 1974 and \$-685 for those employed in 1974), whereas whether an individual was employed or not in 1975 makes little difference in terms of treatment effect. A question addressed in Section 6.5 is how well are these group-specific treatment effects replicated using artificial control groups. The importance of the earnings profile for 1974 in determining the probability of training participation is discussed in Section 7.2 in the context of the sensitivity to selection on observables assumption.

Non-experimental estimates of the treatment effect are based on the two distinct control groups used by Lalonde (1986), the Panel Study of Income Dynamics (PSID) and Westat's Matched Current Population Survey-Social Security Administration File (CPS-SSA). From these two control groups, several sub-groups are then created following criteria defined in Table 3A (Lalonde [1986: Table 3]).¹⁷ The difference

¹⁶ The variation in the treatment effect is tested by estimating equation (7) using the NSW data for a single characteristic and testing for the null hypothesis on the coefficient of the interaction term. Note that column (4) of Table 2 reports the difference in treatment effect between a group and its complement. In the case of a binary variable such as black, the estimate in column (4) is exactly the estimated parameter of the interaction term in equation (7).

¹⁷ Note that there is a small discrepancy between PSID-1 in Table 3 and in Lalonde's Table 3. This is due to missing observations for three individuals which had to be discarded for regression analysis. The table was not able to replicate exactly CPS-2,3 from the original data files although the results are extremely close.

between the sample characteristics of these various controls and the NSW treated group is quite striking especially pre-treatment earnings, and although the subsequent sub-groups reduce some of these differences, there are still significant differences in variables such as pre-treatment earnings. Following Lalonde, the training effect between treated and control groups is estimated in two ways: first as a difference in means of earnings for treated and control (the unadjusted treatment effect), and second, controlling for sample characteristics as well as pre-treatment earnings using the least squares specification in Lalonde (1986) (equation (5)) by regressing earnings in 1978 on a dummy variable for treatment and a set of covariates (hereafter “adjusted” treatment effect). These two estimators are reported throughout the analysis.

Table 4 presents the complete set of estimates originally reported by Lalonde (1986) with the first row reproducing the experimental treatment effect using the NSW control group.¹⁸ The simple difference in means yields highly negative treatment effects, \$ -15,205 if PSID-1 is used as a control group and \$-8,498 if CPS-1 is used as control. The adjusted treatment effect which controls for pre-treatment earnings is reported in column (10). The treatment effect is \$218 using PSID-1 and \$738 for CPS-1, still \$1,576 to \$1,056 away from the experimental benchmark. The results do not greatly improve when the other control groups are used. With PSID-2, and CPS-2 the unadjusted treatment effects are still negative, and the adjusted treatment effect improves somewhat although the standard errors increase substantially. With PSID-3

¹⁸ This is equivalent to Table 5 in Lalonde (1986) using the smaller sample. As expected from a randomized experiment, the unadjusted estimator is close to the adjusted treatment effect when using the randomized data.

and CPS-3 the estimates are closest to the benchmark treatment effect. Applying one of the specification tests suggested by Lalonde, of regressing pre-treatment earnings in 1975 and in 1974 over the same functional form used to estimate the adjusted treatment effect, the researcher would have to reject all the estimators in column (10) since the difference in pre-treatment earnings (1974 and 1975) of the two groups is significant.¹⁹

The essential insight of Lalonde's study is that adjustment on the composite sample of NSW treated and CPS or PSID controls provides an estimated treatment effect which fails to replicate the experimental treatment effect. The direction in which Lalonde proceeds with this observation is to demonstrate that the results are very sensitive to the specification of the model used. Lalonde also shows that in the absence of clear exclusion restrictions that identify the treatment effect, assumptions on unobserved characteristics that help identify the treatment effect do not fare better than standard methods in estimating the training effect. The choice of control group or subset of control is also an important factor in estimating treatment effects (see also Fraker and Maynard [1987] on this issue). In addition specification tests fail to provide useful guidance to the researcher on which estimate to study further.²⁰ The results in columns (4) and (10) of Table 4 are used as a benchmark to evaluate the approach developed in the previous section.

¹⁹ The usefulness of one additional year of pre-treatment earnings becomes apparent when applying the specification test; both for PSID-1,2, and CPS-1,2 the test fails for earnings in 1974 and 1975. However, for PSID-3, and CPS-3 adjusted pre-treatment earnings in 1975 are statistically the same for treated and control but are significantly different for 1974. Without earnings in 1974 the estimated treatment effect would pass this specification test.

²⁰ The possible role of specification tests in identifying the appropriate model to choose from the array of models used by an econometrician is explored by Heckman and Hotz(1989).

6. Results Using the Propensity Score

6.1 Estimating the Propensity Score

Following the algorithm outlined in Section 4, the propensity score is estimated using equation (15) with the treatment status as the dependent variable and the pre-treatment covariates as independent variables. The final choice of interaction and higher order terms included in the logistic function was determined solely by the need to balance the covariates within blocks as defined in the algorithm. This procedure was repeated for each of the six control groups of Table 3A separately and the resulting logistic functions are presented in the footnote of Table 5. Note that in this procedure the outcome variable (earnings in 1978) plays no role.²¹ Note also, that the procedure embodies specification tests of the type suggested by Lalonde. Within each block unadjusted as well as regression-adjusted differences in all pre-treatment covariates between the two groups, including pre-training earnings, were estimated. If the difference was not significant the blocks were maintained, otherwise the procedure was repeated.

6.2 Treatment Effect Stratifying on the Propensity Score

The first of the estimators discussed in Section 4 uses the stratification resulting from estimating the score and a step-wise functional form for $h(\cdot)$ such as equations (18-19). The equation is estimated by summing over the blocks the difference of the within-block means of the outcome variable for the treated and control observations, where the

²¹ Earnings do enter however in lagged form as pre-treatment earnings in 1975 and 1974.

sum is weighted by the number of treated observations within each block, obtaining the unadjusted treatment effect.²² Alternatively, a treatment effect could be obtained using the same regression specification as column (2) of Table 5 within each block, and again taking a weighted sum over the blocks to obtain the adjusted treatment effect.

Tables 6 and 7 present the disaggregated treatment effect using the stratification resulting from the estimation of the propensity score for PSID-1 and CPS-1 respectively. Note that all control observations with an estimated score lower than the minimum estimated score for treated observations are excluded. For Table 6 this is indicated in the first block where the lowest estimated score is 0.0004. The number of control observations used is determined only by the degree of overlap between the distribution of the score for the treated and control groups, resulting in 1070 control observations from PSID-1. The estimated training effect is \$1,509 and \$1,647 for the unadjusted and adjusted estimates respectively. In Table 7 the minimum estimated score is 0.001, which is the lowest estimated value for the treated observations. The total number of CPS-1 control actually used was 3,992 implying that 12,060 control observations (a full 75 percent of the total number of control observations) had an estimated score less than the minimum estimated score for the treated observations. This illustrates well the weakness of the standard model; in linear models such as those tested by Lalonde (1986) one is extrapolating from a group made up mostly of irrelevant controls. With CPS as a control group, after controlling for observable

²² As explained in footnote 11, in the non-experimental set-up this corresponds to the average treatment effect conditional on having been in the NSW treated group, but because the NSW is a randomized experiment its treated and control groups are drawn from the same population, so the correct benchmark for comparison remains the treatment effect of US \$1,794.

characteristics through the propensity score the unadjusted training effect is \$ 1,713 and \$ 1,774 for the adjusted training effect.

Several other characteristics of the tables should be mentioned. Blocks vary in their score range because though a greater number of observations within a block is desirable, ultimately the block size will depend on the balancing of the covariates as described in the algorithm. The treatment effect varies within each block since it depends on the particular sample characteristics represented in the block. For example, treated observations in block 1 of Table 7 had an average age of 27, 20 percent of them were black participants, and the average earnings was \$ 6,620 in 1975. In contrast, the last block was made up of treated participants with an average age of 26, all of whom were black with an average income of \$ 194 in 1975. The variation in treatment effect is taken up further in Section 6.5. Finally, the similarity of the unadjusted and adjusted treatment effect within each block suggests that by conditioning on the score the effects of randomization are replicated.

The estimated treatment effects from stratification on the score for all six groups are summarized in Table 5 in columns (4) and (5). Columns (1) and (2) repeat the benchmark estimates discussed earlier for convenience. The main feature of these results is that the use of the propensity score has eliminated those observations in the control groups that are not comparable to the treated observations, without resorting to any *ad hoc* assumptions on the characteristics of the control observations used to derive PSID-1 to PSID-3 and similarly CPS-1 to CPS-3. This is not to say that in going from CPS-1 to CPS-3 and PSID-1 to PSID-3 one may not improve the estimate; instead, the

basic point is that ensuring overlap through the score is a more systematic way to eliminate irrelevant controls.

Furthermore, comparing these estimates with the estimated training effect over the entire sample (columns 1 and 2) demonstrates the problem associated with the extrapolation implicit in the least squares training effect when there is minimal overlap in the two distributions. Unlike the estimates in columns (1) and (2) the treatment effect estimated in column (3) is estimated using the same specification as (2) adding the score as a variable and regressing over the overlap sample specified in column (6). The treatment effect for both control groups PSID-1 and CPS-1 is \$542 and \$893 respectively. A constant additive treatment effect estimated in the overlap sample does not result in substantially higher estimates. As results in columns (4) and (5) indicate, ensuring that the distributions overlap *and* relaxing the constant treatment assumption in a more flexible form yield estimates that are considerably closer to the randomized benchmark estimate.

Moving down the various control groups does not significantly alter the estimated treatment effect in columns (1) and (2). The estimated treatment effect in columns (4) and (5) range from a low of \$ 1,335 to a high of \$ 1,713 for the unadjusted estimate with CPS controls and from a low of \$ 1,509 to a high of \$ 1,829 for PSID controls. In the case of adjusted estimates the training effect varies from \$ 1,023 to \$1,774 and \$ 1,647 to \$ 2,538 for the CPS and PSID controls respectively. Note however that under stratification on the propensity score there is no need to construct further control groups since this is achieved by the blocking on the score.

Under this approach a researcher would no longer need to construct somewhat arbitrary control groups such as PSID-2, PSID-3 and CPS-2, CPS-3 and would report only the adjusted training effects that vary between \$1,509 and \$1,774.

6.3 Matching on the Propensity Score

As suggested in Section 4 an alternative to stratifying on the score is pair-wise matching. By matching each treated unit to the control with the nearest propensity score (with replacement), we focus attention on a much smaller subset of the overall control group. For PSID-1 to 3, 52, 31, and 43 controls are used respectively and for CPS-1 to 3, the number of controls matched to the treated observations are 106, 87, and 63 respectively. The characteristics of the matched control samples are reported in Table 3B. Comparing the sample characteristics of the matched sample with those in Table 3A shows precisely the result from matching on the propensity score. Columns (6) and (7) of Table 5 present the unadjusted and adjusted treatment effects.²³ The treatment effect varies from \$ 870 to \$ 2,190 (unadjusted) and \$ 826 and \$ 1,740 (adjusted) with PSID controls. With CPS controls, the treatment effect varies from \$ -466 to \$ 1,445 (unadjusted) and \$ -372 to \$ 1,589 (adjusted). Again, a researcher following our approach would not need to construct control groups other than the original control group so that the adjusted estimated treatment effect under matching methods would vary between \$ 1,174 and \$ 1,690. Although the researcher would

²³ Note however that weights need to be used in matched samples to take into account the matching of more than one treated observations to the same control observation. For more details see Dehejia and Wahba (1995).

miss estimates somewhat closer to the experimental benchmark by not using control groups such as PSID-3 and CPS-2, she would also eschew particularly poor results by not using ad hoc sub-groups such as PSID-2 or CPS-3.

6.4 Using the Propensity Score as Weights

The estimates using the score directly as weights are presented in column (9) of Table 5. The treatment effects for the PSID-1 and CPS-1 samples are \$1,129 and \$1,485 respectively. As we vary the control sample (and accordingly re-run the logistic regression), no noticeable variation in the reported treatment effects are observed. It is difficult to give a proper interpretation to the estimates under the reduced samples; the observations dropped from the control group could be those which are least likely to be treated (low score), or if the chosen criterion for reducing the sample is an inappropriate one, the observations dropped could be those that are most likely to be treated (high score). Either way, by removing them from the sample the information they contribute to estimating the score accurately is lost.

The critical issue concerning sub-groups such as those created by Lalonde (1986) to reduce the bias, is that forming subsets of the control group based on single characteristics such as employment status between PSID-1 and -2, for example, imposes a lexicographic preference in terms of suitability of matches on that characteristic. Instead, by allowing the score to choose from the full data set, one incorporates all observable characteristics weighted by the probability of selection.

6.5 Estimating the Treatment Effect by Sample Characteristics

A notable feature of the results presented in the previous sections is the high standard errors on the treatment effects. Table 2 exposed a significant degree of heterogeneity in the benchmark training effect, suggesting that a model with a constant treatment effect such as equations (5) can be substantially misleading. The potential heterogeneity of treatment effects is made explicit by the non-parametric nature of the estimation strategy followed in this paper.

As Tables 6 and 7 indicate, the treatment effect for the blocks vary from a highly positive to a highly negative effect. These estimated treatment effects are for observations with similar propensity scores and socio-economic characteristics within blocks but different across the blocks. For example, block 6 of Table 7 with a score range of 0.6 to 0.85 is made up of 26 treated observations and 12 control observations all of whom are blacks, with an average age of 26, 10 years of high school, less than a third of them married, with no earnings in 1974, and very low earnings (\$250) in 1975. For this group the average training effect was \$2,364 (unadjusted) and \$3,683 (adjusted) which would compare with the benchmark training effect of \$2,029 for black participants, or \$2,692 for those with no earnings in 1974 reported in Table 2. The blocking, in and of itself, cannot provide as sharp a result as conditioning on specific sample characteristics, but for some of these blocks, where the number of treated and control observations are well balanced, the negative (or positive) treatment effect simply reflects the heterogeneity of the treatment effect in the experimental data set. Also, the minimal overlap between the two distributions of the treated and control

observations implies small sample size as indicated in Tables 6 and 7, resulting in higher standard errors than treatment effects over the entire sample.

Table 8 estimates the treatment effect by sample characteristics using the PSID-1 and CPS-1 controls. We see that for many of the characteristics the estimates closely match the experimental results. This provides substantial added confidence in the accuracy of the non-experimental results, since not only do they track the average treatment effect for the NSW group, but they also track the average treatment effect for sub-sets of the original group. Note however that standard errors are still relatively high for many of the treatment effects controlling for individual characteristics, especially when using PSID-1 as control group as indicated in Table 8.

Thus in summary using both the PSID and CPS we estimate treatment effects which come reasonably close to the experimental benchmark. Lalonde's message from his analysis was that the researcher is presented with an array of estimates which differ dramatically (from \$-15,205 to \$1,326) and with no clear way to choose between them. In contrast, from our array of estimators, the answer which emerges is much more focused. Furthermore, the estimates are based on a simple algorithm for comparing observations as summarized by their propensity score. The flexibility of the approach is also demonstrated in the way it is able to replicate to a large extent non-constant treatment effect embodied in the original experimental data set.

7. Sensitivity Analysis

7.1 Sensitivity to Specification of the Propensity Score

Under the algorithm defined in Section 4, the choice of interaction terms in the logistic function is entirely determined by the need to balance the covariates within blocks. Table 9 presents various point estimates of the treatment effect with CPS-1 and PSID-1 as control groups, starting with the logit function reported in Table 5 and then excluding high order terms (squared and cubic) followed by excluding interaction terms from the logit. Although none of the resulting logistic functions completely balance the covariates for equal values of the score (as did the logit function reported in Table 5), the results indicate that the point estimates stratifying on the score are not highly sensitive to logit specifications. Estimates in column (3) where the score enters linearly in the regression are also not much sensitive to propensity score specification. This points to the crucial characteristic of our approach, namely that the choice of terms in the logit specification is driven only by the need to balance covariates of observations with similar propensity scores. In contrast, the estimation of the treatment effect following equation (7) requires prior information on which term to include. Note also that with CPS controls, standard errors are significantly lower.²⁴ Sensitivity analysis to starting parameters in the logit for score estimation (see the first step in figure 1) were

²⁴The lower standard errors that come with coarser specification of the logistic function suggest a tradeoff between efficiency and unbiasedness. The properties of the algorithm proposed in Section 3 and any other algorithm need to be evaluated fully, a direction for future research.

also conducted and generally produce the same logit specification. Also, results were not sensitive to changes in the initial blocking rule (see the second step in figure 1).²⁵

7.2 Sensitivity to Selection on Observables

The key assumption driving the above analysis is that all the variables generating assignment to treatment (and correlated with potential outcomes Y_{i1} and Y_{i0}) are observed. It is clear that rarely are all the relevant variables observed by the researcher. To this extent, one must examine how far we can go in removing the bias from the results through conditioning on observables. In this section we examine this issue by excluding pre-treatment earnings in 1974 and re-estimating the treatment effect using the estimators described in Section 4. The results of Table 5 are re-computed and presented in Table 10.

The first apparent difference between Tables 5 and 10 is the sensitivity of the PSID sample to pre-treatment earnings in 1974. When 1974 earnings are dropped, estimates of the training effect are negative with very high standard errors. As expected, the use of Lalonde's PSID-1 and PSID-2 samples does not change the results very much. Estimates of the treatment using matching or the score as weights also perform poorly. In contrast, using CPS as a control group results in estimates that are more robust. Stratification on the score produces with CPS-1 an adjusted training effect of \$1,207 (s.e.=880). Pair-wise matching on the score also produces a

²⁵ The results reported in this paper were produced by implementing the iterative aspect of the algorithm summarized in figure 1 manually. A preliminary computerized version of the algorithm has been written and reproduces the general results of the paper.

significant effect of \$ 1,969 (s.e.=808) only \$175 higher than the benchmark case. The reason for this important difference between the two control groups is found by examining the distribution of earnings in 1974 and 1975 across the propensity score blocks. Whereas earnings in 1974 are not balanced for most of the blocks in the case of PSID-1, the opposite is true for most (but not all) of the blocks with CPS-1. The difference in the two samples comes from relatively different pre-treatment earnings profile. In PSID-1, earnings in 1975 do not follow closely earnings in 1974 controlling for the propensity score; for higher propensity score levels, earnings in 1975 do not fall as sharply as earnings in 1974 resulting in a negative correlation between the two years. In contrast, earnings in CPS-1 for 1974 and 1975 follow each other closely, both dropping substantially with higher score levels resulting in positive correlation across all blocks. With the dip in earnings, captured earlier in the CPS sample, dropping earnings in 1974 affects the estimates of the training effect for PSID-1 but not CPS-1 (see Ashenfelter [1974, 1978] and Ashenfelter and Card [1985] on what has been referred to as the “Ashenfelter dip” in earnings prior to enrollment in training programs).

7.3 The Use of More than One Control Group

By comparing the overall results in Table 10 to those in Table 5 the value of using several control groups becomes evident. Whereas a coherent estimate of the treatment effect emerges in Table 5, Table 10 shows that if the researcher did not know that an important covariate was missing, she would report a treatment effect that varies

substantially depending on the control group used. How does one compare the estimates of the treatment effect for two (or more) control groups? In the above analysis we tested sensitivity to the available set of covariates by using our knowledge of the experimental benchmark to see how far we strayed from the true estimate when a key covariate was set aside. In applications, such randomized data sets are not typically available, but though it is more difficult to assess sensitivity to unobservable characteristics, it is not impossible (see Rosenbaum and Rubin [1983b]).

The use of more than one control group provides additional information regarding the sensitivity of the results to unobservable covariates. The practice of utilizing multiple control groups in economics and more specifically in the manpower training literature is not uncommon, but studies generally report only the final control group used in the evaluation.²⁶ There is however a fundamental difference between sensitivity to the choice of control group *within* a specific data set (and sub-groups obtained from it), as was addressed in the previous section, and *between* two distinct control groups. The former is an issue already addressed by making use of the propensity score. Comparing results between two distinct control groups is a more delicate exercise. Some studies (Rosenbaum [1984, 1987]) suggest that the use of a second control group in non-experimental settings can sometimes help detect the presence of important variables not observed in the data. The intuition is simple, and was illustrated by Tables 5 and 10. When a variable that determines assignment to

²⁶ Fraker and Maynard (1987) provide a detailed analysis of the treatment effect for the NSW program using a series of control groups. They conclude that the results are generally sensitive to the choice of control groups.

treatment is not observed there are two possibilities. If the estimated treatment effect across the two samples is quite similar (as in Table 5), this suggests either that all important variables are observed or that the unobserved variable affects the observed covariates of both samples similarly. If instead the estimates differ substantially (as in Table 10), this suggests quite strongly the presence of some unobserved variable which affects each sample differently. Without an experimental data set, the use of multiple control groups can provide a partial test for the presence of unobserved variables.

8. Conclusion

This paper presents a framework for estimating treatment effects in non-experimental settings when assignment to treatment is assumed to be ignorable conditional on observable characteristics. Drawing from the statistics literature on causal inference analysis, the paper defines the role of the propensity score in identifying treatment effect with conditionally ignorable assignment. The paper then proposes an algorithm for estimating the propensity score, and three types of estimators of the treatment effect based on the score.

The estimators are evaluated using Lalonde's seminal re-creation of a non-experimental setting. Results show that the estimates of the training effect are close to the benchmark randomized case, and are robust to specification of the control groups defined by Lalonde. By stratifying observations on the score, a researcher needs only to use the original control groups to estimate the training effect and would report an effect that varies between \$1,509 and \$1,774 compared to the randomized treatment

effect of \$ 1,794. Using estimators based on matching or weighting by the propensity score lead to similar estimates. The paper also evaluates sensitivity to the specification of the propensity score as well as sensitivity to the selection on observables assumption. Results indicate the robustness of the estimated training effect to changes in the benchmark logit specification and to blocking methods. Excluding earnings in 1974 from the analysis affects the estimated training effect when using PSID as control but less so with CPS, a result that underscores the importance of using more than one control group in non-experimental studies.

In most of the estimates the standard errors are high, and although the heterogeneity of the treatment effect as well as the minimal overlap in the distribution of covariates between treated and control go far in explaining the high standard errors, further research is needed to examine the optimality properties of the rule specifying the score, such as the tradeoff between unbiasedness and efficiency (possibly through Monte Carlo studies). While the results obtained in the paper are specific to the data set, further studies based on non-experimental evaluation of randomized studies should provide additional evidence on the merits of this approach and on how general (or specific) are these methods to the data at hand. This does not deny the importance of randomized experiments; indeed, it is thanks to a randomized data set that such an evaluation was made possible.

The Lalonde paper and the ensuing debate may have cast a negative light on standard econometric methods of evaluating social programs. This paper attempts to rehabilitate the assumption of selection on observables with the use of the propensity

score to exploit fully the ignorability assumption. There are however many settings in which the assumption of selection on observables is not sufficient to identify the treatment. The conclusion to draw from this paper is that even when the researcher suspects that important characteristics are unobserved and that exclusion restrictions that identify the treatment may be available, the self-diagnostic nature of the approach reveals valuable information to the researcher by examining the comparability of the distributions of the treated and control units. The techniques exposed in this paper are powerful enough to sort out which observations from a large pool of potential controls are relevant comparisons to treated units under consideration and to help guide the researcher in other possible directions. Our argument would be: before recourse to modeling through assumptions on functional forms and distribution, assumptions on unobservables which by their very nature are difficult to test in the data, there is substantial reward in exploring first the information contained in the variables that *are* observed.

References

- Angrist, J. (1990). "Lifetime Earnings and the Vietnam Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review*, 80, 313-335.
- (1995). "Using Social Security Data on Military Applicants to Estimate the Effect of Voluntary Military Service on Earnings." Massachusetts Institute of Technology, unpublished.
- Angrist, J., G.W. Imbens, and D. Rubin (1996). "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444-454.
- Ashenfelter, O. (1974). "The Effect of Manpower Training on Earnings: Preliminary Results," in J. Stern and B. Dennis (eds.), *Proceedings of the Twenty-Seventh Annual Winter Meetings of the Industrial Relations Research Association*. Madison: Industrial Relations Research Association.
- (1978). "Estimating the Effects of Training Programs on Earnings," *Review of Economics and Statistics*, 60, 47-57.
- and D. Card (1985). "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," *Review of Economics and Statistics*, 67, 648-660.
- Barnow, B., G. Cain, and Arthur Goldberger (1980). "Issues in the Analysis of Selectivity Bias," *Evaluation Studies*, 5, 42-59.
- Burtless, Gary (1995). "The Case for Randomized Field Trials in Economic and Policy Research," *Journal of Economic Perspectives*, 9, 61-84.
- and Orr, L. (1986). "Are Classical Experiments Needed for Manpower Policy?" *Journal of Human Resources*, 21, 606-639.
- Card, David and Daniel Sullivan (1988). "Measuring the Effect of Subsidized Training Programs on Movements in and out of Employment," *Econometrica*, 56, 497-530.
- Cox, D.R. (1992). "Causality: Some Statistical Aspects," *Journal of the Royal Statistical Society, series A*, 155, 291-301.
- Cochran, W. G. (1968). "The Effectiveness of Adjustment by Sub-Classification in Removing Bias in Observational Studies" *Biometrics*, 24, 295-313.

Dehejia, Rajeev H. and Sadek Wahba (1994). "Re-evaluating the Evaluation of Training Programs: On the Methodology of Causal Inference," Harvard University, unpublished.

----- (1995). "An Oversampling Algorithm for Causal Inference in Non-Experimental Studies with Incomplete Matching and Missing Outcome Variables," Harvard University, unpublished.

Duggan, Mark Gregory (1992). *Matched Sampling Methods to Reduce Bias in an Observational Study*. Massachusetts Institute of Technology, unpublished M.Sc. Thesis, Department of Electrical Engineering and Computer Science.

Fisher, R. (1935). *The Design of Experiments*. London: Oliver and Boyd.

Fraker, T. and R. Maynard (1987). "Evaluating Comparison Group Designs with Employment-Related Programs," *Journal of Human Resources*, 22, 194-227.

Goldberger, Arthur (1972a). "Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations," University of Wisconsin, Institute for Research on Poverty, Discussion paper, 123-72.

----- (1972b). "Selection Bias in Evaluating Treatment Effects: The case of Interaction," University of Wisconsin, Institute for Research on Poverty, Discussion paper, 129-72.

Hardle, Wolfgang (1990). *Applied Nonparametric Regression*. Econometric Society Monographs, Cambridge: Cambridge University Press.

Heckman, J. (1979). "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 931-961.

----- (1990). "Varieties of Selection Bias," *American Economic Review*, 80, 313-318.

----- and J. Hotz (1989). "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training," *Journal of the American Statistical Association*, 84, 862-880.

----- and Richard Robb (1985). "Alternative Methods for Evaluating the Impact of Interventions," in James Heckman and Burton Singer (eds.), *Longitudinal Analysis of Labor Market Data*. Econometric Society Monograph No. 10, Cambridge: Cambridge University Press.

----- (1986). "Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatment on Outcomes," in Howard Rainer (ed.), *Drawing Inferences from Self-Selected Samples*. New York: Springer-Verlag.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith and Petra Todd (1995). "Non-Parametric Characterization of Selection Bias Using Experimental Data: A Study of Adult Males in JTPA," University of Chicago, unpublished.

Holland, Paul W (1986). "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81, 945-960.

Imbens, Guido W. and J. Angrist (1994). "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467-475.

Katz, Lawrence (1992). "Recent Developments in Labor Economics," American Economic Association Meetings, January 1992.

Lalonde, Robert (1984). "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," Princeton University, Industrial Relations Section, Working Paper No 183.

----- (1986). "Evaluating the Econometric Evaluations of Training Programs," *American Economic Review*, 76, 604-620.

Leamer, Edward (1983). "Let's Take the Con Out of Econometrics," *American Economic Review*, 73, 31-43.

Maddala, G.S. (1983). *Qualitative and Limited Dependent Variable Models in Econometrics*. Econometric Monograph No. 3, Cambridge: Cambridge University Press.

Manpower Demonstration Research Corporation (1983). *Summary and Findings of the National Supported Work Demonstration*. Cambridge: Ballinger.

Manski, Charles F. (1995). *Identification Problems in the Social Sciences*. Cambridge: Harvard University Press.

-----, G. Sandefur, S. McLanahan, and D. Powers (1992). "Alternative Estimates of the Effect of Family Structure During Adolescence on High School Graduation," *Journal of the American Statistical Association*, 87, 25-37.

Mroz, T.A. (1987). "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions," *Econometrica*, 55, 765-799.

Neyman, J. (1935). "Statistical Problems in Agricultural Experimentation," *Journal of the Royal Statistical Society*, supplement, II, 107-180.

Pratt, John and Robert Schlaifer (1988). "On the Interpretation and Observations of Laws," *Journal of Econometrics*, 39, 23-52.

Reinisch, June, Stephanie Sanders, E. Mortensen, and Donald Rubin (1993). "Prenatal Exposure to Phenobarbital and Intelligence Deficits in Adult Human Males," The Kinsey Institute for Research in Sex, Gender and Reproduction, unpublished.

Rosenbaum, P. (1987). "The Role of a Second Control Group in an Observational Study," *Statistical Science*, 2(3).

Rosenbaum, P., and D. Rubin (1983a). "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41-55.

----- (1983b). "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome," *Journal of the Royal Statistical Society*, Series B, vol. 45.

----- (1985a). "Reducing Bias in Observational Studies Using the Subclassification on the Propensity Score," *Journal American Statistical Association*, 79, 516-524.

----- (1985b). "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity," *American Statistician*, 39, 33-38.

Rubin, D. (1973). "Matching to Remove Bias in Observational Studies," *Biometrics*, 29, 159-183.

----- (1974). "Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies," *Journal of Educational Psychology*, 66, 688-701.

----- (1977). "Assignment to a Treatment Group on the Basis of a Covariate," *Journal of Educational Statistics*, 2, 1-26.

----- (1978). "Bayesian Inference for Causal Effects: The Role of Randomization," *The Annals of Statistics*, 6, 34-58.

----- (1979). "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observation Studies," *Journal of the American Statistical Association*, 74, 318-328.

Rubin, Donald B. and Neal Thomas (1992). "Characterizing the Effect of Matching Using Linear Propensity Score Methods with Normal Distributions," *Biometrika*, 79, 797-809.

Table 1: Sample Means and Standard Errors of Covariates For Male NSW Participants

Variable	National Supported Work Sample (Treatment and Control)		Dehejia-Wahba Sample	
	Lalonde Sample Treatment	Lalonde Sample Control	Treatment	Control
Age	24.63 (0.39)	24.45 (0.32)	25.81 (0.52)	25.05 (0.45)
Years of schooling	10.38 (0.11)	10.19 (0.08)	10.35 (0.15)	10.09 (0.1)
Proportion of School Dropouts	0.73 (0.03)	0.81 (0.02)	0.71 (0.03)	0.83 (0.02)
Proportion of Blacks	0.80 (0.02)	0.80 (0.02)	0.84 (0.03)	0.83 (0.02)
Proportion of Hispanic	0.09 (0.02)	0.11 (0.02)	0.059 (0.017)	0.1 (0.019)
Proportion Married	0.17 (0.02)	0.16 (0.02)	0.19 (0.03)	0.15 (0.02)
No. of Children	0.38 (0.05)	0.36 (0.05)	0.41 (0.07)	0.37 (0.06)
No show variable	0 (0)	n/a	0 (0)	n/a
Month of Assignment (Jan. 1978=0)	16.47 (0.28)	15.73 (0.27)	18.49 (0.36)	17.86 (0.35)
Real Earnings 12 months before training	1,771 (175)	1,700 (161)	1,689 (235)	1,425 (182)
Real Earnings 24 months before training	3,571 (335)	3,672 (316)	2,096 (359)	2,107 (353)
Hours Worked 1 year Before Training	322 (29)	293 (24)	294 (36)	243 (27)
Hours Worked 2 Years Before Training	543 (43)	516 (36)	306 (46)	267 (37)
Sample Size	297	425	185	260

Table 2: Training Effect and Sample Means of Covariates Blocking on Selected Sample Characteristics for NSW Male Group

	No. of Obs.		Treatment Effect US\$ (s.e.)		Sample Means of Covariates (s.e.)										
	(1) Treat- ment	(2) Con- trol	(3) Unadju- st- ed	(4) Diff. with Comp. ^b	Age	Educ	Black	Hisp	Ndegree	Married	RE74 ^a US\$	RE75 US\$	MOA	U74 ^a	U75
Sample-1	297	425	886 (476)		24.63	10.38	0.8	0.09	0.73 [*]	0.17	3,571	3,066	16.47	0.44	0.37
Sample-2	185	260	1,794 (633)		25.81	10.35	0.84	0.059	0.71 [*]	0.19	2,096	1,532	18.49	0.71	0.60
Black	156	215	2,029 (706)		25.98	10.31	1	0	0.72 [*]	0.19	2,155	1,491	18.38	0.71	0.62
Non- Black	29	45	803 (1331)	1,226 (1703)	24.93	10.52	0	0.38	0.62	0.21	1,776	1,755	19.06	0.69	0.52
Ndegree =1	131	217	1,154 (696)		25.37	9.47	0.86	0.07	1	0.20	1,545	1,603	18.50	0.74	0.61
Ndegree =0	54	43	3,192 (1517)	-2,037 (1523)	26.89	12.46	0.80	0.04	0	0.17	4,431	1,360	18.48	0.63	0.57
Educ >=11	98	113	3,085 (1033)		26.59	11.81	0.84	0.03	0.45 [*]	0.21	2,419	1,401	18.52	0.69	0.61
Educ <11	87	147	402 (753)	2,683 (1265)	24.94	8.70	0.85	0.09	1	0.16	1,731	1,680	18.46	0.72	0.58

Table 2: Training Effect and Sample Means of Covariates Blocking on Selected Sample Characteristics for NSW Male Group (Cont)

Cov.	No. of Obs.		Treatment Effect US\$ (s.e.)		Sample Means of Covariates (s.e.)										
	Treat ment	Cont rol	Diff. in mean	Diff. with Comp.	Age	Educ	Black	Hisp	Ndegre	Married	RE74 US\$	RE75 US\$	MOA	U74	U75
MOA >=18	108	141	2,717 (956)		25.37	10.49	0.82	0.07	0.71	0.20	3,590	2,596	22.32	0.5	0.32*
MOA <18	77	119	482 (748)	2,235 (1272)	26.44	10.14	0.87	0.04	0.70	0.17	0	40	13.12	1	0.99
U74=1	131	195	2,692 (722)		26.52	10.25	0.85	0.05	0.74*	0.16	0	434	16.95	1	0.83
U74=0	54	65	-685 (1278)	3,376 (1414)	24.11	10.57	0.83	0.09	0.63*	0.26	7,179	4,195	22.22*	0	0.04
U75=1	111	178	1,711 (681)		26.64	10.25	0.86	0.05	0.72	0.14	73.08	0	16.08	0.98	1
U75=0	74	82	1,691 (1289)	20 (1320)	24.58	10.49	0.81	0.08	0.69	0.25	5,129	3,830	22	0.30	0

Notes:

*=difference between treated and control is significant at 5%. Age=age of participant. Educ= No. high school years. Black: proportion of black in sample. Hisp: proportion of hispanics in sample. Ndegree: indicator of participants with no school degrees. Married: proportion of married participants. RE74: real earnings (1982 US\$) in 1974. RE74: real earnings (1982US\$) in 1974. RE75: real earnings (US\$ 1982) in 1975. MOA: month of assignment to experiment (Jan. 1978=0). U74: unemployed in 1974. U75: unemployed in 1975.

a: For Sample 1, RE74 and U74 refer to earnings two years prior to assignment and unemployed two years prior to assignment respectively.

b: Difference in average training effect between a group and its complement.

Table 3a: Sample Means of Characteristics for NSW and Control Samples

Control Sample	No of Obs.	Sample Characteristics										
		Age	School	Black	Hisp	Ndegree	Married	RE74 US\$	RE75 US\$	RE78 US\$	U74	U75
NSW	185	25.81	10.35	0.84	0.059	0.71	0.19	2,096	1,532	6,349	0.71	0.60
PSID-1	2,490	34.85	12.11	0.25	0.032	0.31	0.87	19,429	19,063	21,542	0.09	0.10
PSID-2	253	36.10	10.77	0.39	0.067	0.49	0.74	11,027	7,569	9,996	0.34	0.23
PSID-3	128	38.25	10.30	0.45	0.18	0.51	0.70	5,566	2,611	5,279	0.41	0.61
CPS-1	15,992	33.22	12.02	0.07	0.07	0.29	0.71	14,016	13,650	14,847	0.10	0.12
CPS-2	2,369	28.25	11.24	0.11	0.08	0.45	0.46	8,728	7,397	10,171	0.18	0.21
CPS-3	429	28.03	10.23	0.21	0.14	0.60	0.51	5,619	2,467	6,984	0.31	0.26

Notes:

Definition of Control Groups (Lalonde 1986): PSID-1: All male household heads continuously from 1975 through 1978, who were less than 55 years old and did not classify themselves as retired in 1975. PSID-2: Selects from PSID1 group all men who were not working when surveyed in the spring of 1976. PSID-3: Selects from PSID2 all men who were not working in 1975. CPS-2: Selects from CPS-1 all males who were not working when surveyed in March 1976. CPS-3: Selects from CPS-2 all the unemployed males in 1976 whose income in 1975 was below the poverty level.

Table 3b: Sample Means of Characteristics for Matched Control Samples

MATCHED SAMPLES	No. of Obs.	Sample Characteristics										
		Age	School	Black	Hisp	Ndegree	Married	RE74 US\$	RE75 US\$	RE78 US\$	U74	U75
NSW	185	25.81	10.35	0.84	0.059	0.71	0.19	2,096	1,532	6,349	0.71	0.60
MPSID-1	52	25.64	10.57	0.85	0.01	0.65	0.16	1,845	1,385	4,159	0.65	0.57
MPSID-2	31	26.03	11.03	0.88	0.01	0.57	0.16	2,190	1,506	4,815	0.69	0.59
MPSID-3	43	25.56	11.03	0.91	0.01	0.59	0.19	2,227	2,365	5,479	0.68	0.57
MCPS-1	106	25.58	10.80	0.88	0.03	0.60	0.18	1,792	1,435	5,278	0.77	0.60
MCPS-2	87	26.17	10.43	0.84	0.08	0.62	0.16	2,487	1,699	4,904	0.70	0.50
MCPS-3	63	25.49	10.50	0.89	0.05	0.68	0.16	1,771	1,666	6,816	0.67	0.71

Table 4: Lalonde's Earnings Comparisons and Estimated Training Effects for the NSW Male Participants Using Comparisons Groups from the PSID and the CPS-SSA^{a,b}

Comparison Group	Comparison Group Earnings Growth 1975-1978	NSW Treatment Earnings Less Comparisons Group Earnings				Difference in Differences: Difference in Earnings Growth 1975-1978 Treatment Less Comparisons		Unrestricted Difference in Differences: Quasi Difference in Earnings Growth 1975-1978		Controlling for All Observed Variables and Pre-Training Earnings
		Pre-Training Year 1975		Post-Training Year, 1978		Without Age	With Age	Unadjusted	Adjusted ^c	
		Unadjusted	Adjusted ^c	Unadjusted	Adjusted ^c					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
NSW	3,288 (375)	265 (303)	232 (306)	1,794 (632)	1,672 (637)	1,529 (679)	1,510 (681)	1,750 (632)	1,631 (637)	1,672 (638)
PSID-1	2,490 (217)	-17,531 (1002)	-10,298 (1008)	-15,205 (1154)	-7,741 (1175)	2,327 (813)	1,139 (827)	-582 (841)	-265 (880)	218 (866)
PSID-2	2,427 (152)	-6,037 (695)	-5,244 (825)	-3,647 (959)	-2,810 (1081)	2,390 (839)	1,200 (928)	720 (886)	297 (1004)	907 (1004)
PSID-3	2,669 (745)	-1,079 (498)	-1,081 (639)	1,069 (899)	35 (1100)	2,148 (958)	864 (1121)	1,370 (896)	243 (1100)	822 (1101)
CPS-1	1,196 (161)	-12,119 (682)	-6,908 (628)	-8,498 (712)	-4,417 (713)	3,621 (570)	2,452 (557)	-78 (536)	525 (556)	738 (547)
CPS-2	2,774 (164)	-5,866 (600)	-4,325 (596)	-3,822 (670)	-2,208 (745)	2,043 (610)	1,629 (593)	-263 (573)	371 (662)	879 (654)
CPS-3	4,517 (334)	-934 (287)	-944 (354)	-635 (657)	375 (821)	299 (647)	378 (649)	-91 (641)	844 (807)	1,326 (796)

Notes:

^a The columns above present the estimated training effect for each econometric model and comparisons group using the sample defined in Table 1. The dependent variable is earnings in 1978. Based on the experimental data an unbiased estimate of the impact of training presented in col. 4 is \$ 1,794. The first three columns present the difference between each comparison group's 1975 and 1978 earnings and the differences between the pre-training earnings of each comparison group and the NSW treatments. The estimates are in 1982 dollars. The number in parentheses are the standard errors.

^b For a definition of the comparisons groups see Table 3A.

^c The exogenous variables used in the regressions adjusted equations are age, age squared, years of schooling, high school dropout status, and race.

Table 5: Estimated Training Effects for the NSW Male Participants Using Comparison Groups from PSID and CPS-SSA

	NSW EARNINGS LESS COMPARISON GROUP EARNINGS		NSW COMPARISON TREATMENT EARNINGS LESS COMPARISON GROUP EARNINGS CONDITIONAL ON THE ESTIMATED PROPENSITY SCORE						
	(1) Unadjusted	(2) Adjusted ^a	LINEAR WITH SCORE ^a	STRATIFYING ON THE SCORE		Obs. ^h (6)	MATCHING ON THE SCORE		WEIGHTING WITH THE SCORE ^f
			(3)	(4) Unadjusted	(5) Adjusted ^a		(7) Unadjusted	(8) Adjusted ^a	(9)
NSW	1,794 (633)	1,672 (638)							
PSID-1^b	-15,205 (1154)	218 (866)	542 (1197)	1,509 (1823)	1,647 (1862)	1,255	2,190 (761)	1,690 (1020)	1,129 (2927)
PSID-2^c	-3,647 (959)	907 (1004)	434 (1193)	1,648 (2010)	2,538 (2063)	389	870 (977)	826 (962)	1,951 (1178)
PSID-3^c	1,069 (899)	822 (1101)	862 (1334)	1,829 (2250)	2,308 (2468)	247	1,534 (1223)	1,740 (1063)	1,618 (1231)
CPS-1^d	-8,498 (712)	738 (547)	893 (642)	1,713 (1115)	1,774 (1152)	4,117	1,253 (988)	1,174 (798)	1,485 (3148)
CPS-2^d	-3,822 (670)	879 (654)	399 (765)	1,358 (1432)	1,378 (1582)	1493	1,445 (962)	1,589 (946)	862 (4059)
CPS-3^d	-635 (657)	1,326 (796)	526 (891)	1,335 (1765)	1,023 (1956)	514	-466 (951)	-372 (945)	379 (2308)

Notes:

II. Regressions;

(a) Least Squares Regression: $Re78$ over, c , $expstat$, age , $age2$, $educ$, $nodegree$, $black$, $hisp$, $re74$, $re75$.

(b) Logit: $Prob(expstat = 1) = F(age, age2, educ, educ2, married, nodegree, black, hisp, re74, re75, u74, u75, re74^2, re75^2, u74 * hisp)$

(c) Logit: $Prob(expstat = 1) = F(age, age2, educ, educ2, nodegree, married, black, hisp, re74, re74^2, re75, re75^2, u74, u75)$

(d) Logit: $Prob(expstat = 1) = F(age, age2, educ, educ2, married, nodegree, black, hisp, re74, re75, u74, u75, educ * re74, age3)$

(e) Weighted Least Squares (same as [a]) using number of matched control observations to treated observations as weights.

(f) Weights for treated $W_{it} = 1$; Weights for control $W_{ic} = p_i / (1 - p_i)$ averaged over the number of treated observations.

(g) Least squares regression (a) including score.

III. Sample Size;

(h): Number of observations refer to actual number of control and treated used for (3) to (5).

**Table 6: Estimating Treatment Effect for NSW Male Group
with PSID-1 Control Group Stratifying on the Score^a**

Score Range	Number of Observations		Average Score ^b	NSW Treatment Earnings Less Comparison group Earnings	
	NSW	PSID		Unadjusted	Adjusted ^c
0.0004-0.1	9	942	0.022	-2,693 (3660)	-4,896 (2830)
0.1-0.2	10	57	0.13	-2,759 (5462)	1,078 (5355)
0.2-0.45	19	39	0.34	-1,171 (1410)	-1,070 (1424)
0.45-0.6	20	15	0.53 [*]	-3,044 (2354)	-4,244 (2700)
0.6-0.85	35	10	0.70	-837 (2603)	-265 (2822)
0.85-1	92	7	0.94	4,819 (3412)	4,918 (3469)
Weighted Average	185	1,070	0.70	1,509 (1823)	1,647 (1862)

Notes:

a: Logit used is reported in Table 5.

b: Average estimated score for treated observations

c: OLS as reported in Table 5.

*: Difference in score between treated and control observations significant at 5 %.

Table 7: Estimating Treatment Effect for NSW Male Group with CPS-1 Control Group Stratifying on the Score^a

Score Range	Number of Observations		Average Score ^b	NSW Treatment Earnings Less Comparison group Earnings	
	NSW	CPS		Unadjusted	Adjusted ^c
0.001-0.01	10	2767	0.003	-1,740 (2596)	-2,242 (2134)
0.01-0.1	30	935	0.078*	-139 (1354)	547 (1175)
0.1-0.3	32	150	0.187	500 (1109)	493 (1038)
0.3-0.5	35	46	0.39	200 (1397)	456 (1462)
0.5-0.6	22	17	0.56	6,445 (2604)	4,895 (3116)
0.6-0.85	26	12	0.7	2,364 (2213)	3,683 (2521)
0.85-1	30	5	0.90	3,740 (5772)	3,299 (5869)
Weighted Average	185	3,932	0.43	1,713 (1115)	1,774 (1152)

Notes:

a: Logit as reported in Table 5.

b: Average estimated score for treated observations

c: OLS as reported in Table 5.

*: Difference in score between treated and control observations significant at 5 %.

Table 8a: Treatment Effect for NSW Male and CPS Control Sample

Sample	Blocking on Selected Sample Characteristics (s.e.)						
	All	Black	Non-Black	Ndgree=1	Ndgree=0	Educ >=11	Educ <11
NSW	1,794 (633)	2,029 (706)	803 (1331)	1,154 (696)	3,192 (1517)	3,085 (1033)	402 (753)
CPS-1: Unadjusted	-8,498 (712)	-5,870 (785)	-7,578 (1789)	-6,936 (793)	-7,750 (1329)	-8,040 (987)	-7,541 (967)
Adjusted	738 (547)	1,487 (617)	1,087 (1311)	511 (635)	1,946 (1004)	1,628 (760)	223 (749)
Stratifying on the Score: -Unadjusted	1,713 (1115)	1,738 (1191)	1,367 (1397)	1,439 (1345)	2,475 (1420)	2,501 (985)	1,012 (883)
-Adjusted	1,774 (1152)	1,905 (1331)	1,799 (1847)	1,144 (1536)	2,683 (1256)	2,377 (1913)	738 (1149)

Table 8a: Treatment Effect For NSW Male And CPS Control Sample (cont.)

Blocking on Selected Sample Characteristics (s.e.)					
Sample	All	U74=1	U74=0	U75=1	U75=0
NSW	1,794 (633)	2,692 (722)	-685 (1278)	1,711 (681)	1,691 (1289)
CPS-1:					
Unadjusted	-8,498 (712)	1,710 (647)	-10,735 (1248)	1,949 (665)	-9,198 (1058)
Adjusted	738 (547)	3,189 (699)	-3,275 (982)	2,812 (728)	-1,132 (844)
Stratifying on the Score:					
-Unadjusted	1,713 (1115)	3,334 (1398)	-1,912 (1085)	2,582 (1070)	214 (1334)
-Adjusted	1,774 (1152)	3,445 (1578)	-1,064 (1340)	2,523 (1278)	-153 (1061)

Note: Adjusted training effect uses least squares regressions of Table 5.

Table 8b: Treatment Effect for NSW Male and PSID Control Sample**Blocking on Selected Sample Characteristics (s.e.)**

Sample	All	Black	Non-Black	Ndgree=1	Ndgree=0	Educ > =11	Educ <11
NSW	1,794 (633)	2,029 (706)	803 (1331)	1,154 (696)	3,192 (1517)	3,085 (1033)	402 (753)
PSID-1: Unadjusted	-15,205 (1154)	-9,733 (1002)	-15,961 (3008)	-9,701 (1148)	-16,233 (2172)	-16,117 (1623)	-10,043 (1335)
Adjusted	218 (866)	1,091 (916)	-632 (2078)	1,695 (993)	179 (1569)	1,071 (1222)	474 (1165)
Stratifying on the Score: -Unadjusted	1,509 (1823)	1,486 (2067)	2,880 (2366)	1,667 (2298)	1,137 (2907)	1,806 (2522)	1,381 (2163)
-Adjusted	1,647 (1862)	1,936 (2146)	..	1,826 (2435)	1,435 (3425)	1,001 (2725)	1,694 (2305)

Note: Adjusted training effect uses least squares regressions of Table 5.

Table 8b: Treatment Effect for NSW Male and PSID Control Sample (cont.)

Sample	Blocking on Selected Sample Characteristics (s.e.)				
	All	U74=1	U74=0	U75=1	U75=0
NSW	1,794 (633)	2,692 (722)	-685 (1278)	1,711 (681)	1,691 (1289)
PSID-1: Unadjusted	-15,205 (1154)	-446 (1499)	-17,465 (2022)	501 (1364)	-16,364 (1720)
Adjusted	218 (866)	4,534 (1702)	-4,428 (1431)	1,823 (1813)	-1527 (1241)
Stratifying on the Score: -Unadjusted	1,509 (1823)	4,444 (25000)	-4,681 (1576)	4,160 (2449)	-931 (3518)
-Adjusted	1,647 (1862)	4,408 (2458)	-3,285 (1939)	4,935 (2514)	-2,854 (4883)

Note: Adjusted training effect uses least squares regressions of Table 5.

Table 9: Sensitivity of Estimated Training Effects for the NSW Male Participants to Specification of the Propensity Score

LOGIT BY SAMPLE	NSW EARNINGS LESS COMPARISON GROUP EARNINGS		NSW COMPARISON TREATMENT EARNINGS LESS COMPARISON GROUP EARNINGS CONDITIONAL ON THE ESTIMATED PROPENSITY SCORE						
	(1) Unadjusted	(2) Adjusted ^a	LINEAR WITH SCORE ^d (3)	STRATIFYING ON THE SCORE (4) Unadjusted	(5) Adjusted ^a	No. Obs. ^c (6)	MATCHING ON THE SCORE (7) Unadjusted	(8) Adjusted ^b	WEIGHTING WITH THE SCORE ^c (9)
NSW	1,794 (633)	1,674 (638)				445			
PSID-1:									
1	-15,205 (1154)	218 (866)	542 (1197)	1,509 (1823)	1,647 (1862)	1,255	2,190 (761)	1,690 (1020)	1,129 (2927)
2	-15,205 (1154)	105 (863)	-225 (1217)	1,348 (1558)	2,128 (1699)	1,465	871 (988)	795 (937)	2,017 (2673)
3	-15,205 (1154)	105 (863)	463 (1080)	1,044 (1087)	136 (1226)	1,373	2,124 (869)	2,338 (842)	2,125 (1570)
CPS-1:									
4	-8,498 (712)	738 (547)	893 (642)	1,713 (1115)	1,774 (1152)	4,117	1,253 (988)	1,174 (798)	1,485 (3148)
5	-8,498 (712)	684 (546)	1,103 (614)	1,485 (653)	1,636 (682)	6,365	1,179 (821)	1,258 (897)	1,414 (2221)
6	-8,498 (712)	684 (546)	1,147 (582)	1,456 (595)	1,728 (610)	6,017			1,236 (1824)

Notes:

I. Regressions;

a: Least Squares Regression: Re78 over, c, expstat, age, educ, nodegree, black, hisp, re74, re75.

b: Weighted Least Squares (same as [a]) using number of matched control observations to treated observations as weights.

Logit 1 = Same as Table 5; Prob (expstat = 1) = F(age, age2, educ, educ2, married, nodegree, black, hisp, re74, re75, u74, u75, re 74², re75², u74*hisp)

Logit 2 = Logit 1 - high order terms: Prob (expstat = 1) = F(age, educ, nodegree, married, black, hisp, re74, re75, u74, u75, u74*hisp)

Logit 3 = Logit 2 - interaction terms: Prob (expstat = 1) = F(age, educ, nodegree, married, black, hisp, re74, re75)

Logit 4 = Same as Table 5: Prob (expstat = 1) = F(age, age2, educ, educ2, married, nodegree, black, hisp, re74, re75, u74, u75, educ*re74, age3)

Logit 5 = Logit 4 - high order terms: Prob (expstat = 1) = F(age, educ, married, nodegree, black, hisp, re74, re75, u74, u75, educ*re74)

Logit 6 = Logit 5 - interaction terms: Prob (expstat = 1) = F(age, educ, married, nodegree, black, hisp, re74, re75)

c: Weights for treated $W_{it} = 1$; Weights for control $W_{ic} = p_i / (1 - p_i)$ averaged over the number of treated observations.

d: Least squares regression as in (a) and including score.

II. Sample Size;

e: Number of observations refer to actual number of control and treated used under stratification used in (3)-(5).

Table 10: Sensitivity of Estimated Training Effects for the NSW Male Participants When Dropping Pre-Treatment Earnings in 1974
NSW EARNINGS LESS COMPARISON GROUP EARNINGS
NSW COMPARISON TREATMENT EARNINGS LESS COMPARISON GROUP EARNINGS
CONDITIONAL ON THE ESTIMATED PROPENSITY SCORE

	(1)		(2)		(3)		(4)		(5)		(6)		(7)		(8)		(9)		
	Unadjusted	Adjusted ^a	Unadjusted	Adjusted ^a	Linear with score ¹	Unadjusted	Adjusted ^a	Unadjusted	Adjusted ^a	Obs ¹	Unadjusted	Adjusted ^a	Unadjusted	Adjusted ^a	Unadjusted	Adjusted ^a	Unadjusted	Adjusted ^a	
NSW	1,794 (633)	1,631 (637)																	
PSID-1 ^b	-15,205 (1154)	-265 (880)	-1,085 (1110)	-47 (1442)	635 (1455)	1284	547 (1419)	1,597 (1308)	-251 (3342)										
PSID-2 ^c	-3,647 (959)	297 (1004)	-544 (1149)	223 (1619)	-592 (1615)	356	647 (1232)	1,298 (1033)	1,067 (1915)										
PSID-3 ^d	1,069 (899)	243 (1100)	1,366 (1246)	586 (1551)	440 (1620)	252	1,558 (1059)	2,093 (940)	-355 (1478)										
CPS-1 ^e	-8,498 (712)	525 (557)	949 (632)	1,294 (765)	1,207 (880)	4,558	1,827 (581)	1,969 (808)	1,198 (2748)										
CPS-2 ^f	-3,822 (670)	371 (662)	656 (706)	1,475 (1054)	1,661 (1101)	1,222	924 (777)	857 (813)	2,238 (1528)										
CPS-3 ^f	-635 (657)	844 (807)	988 (860)	1,044 (1417)	1,129 (1509)	504	1,325 (928)	1,074 (942)	1,241 (1494)										

Notes:

I. Regressions;

- (a) Least Squares Regression: Re78 over, c, expstat, age, age2, educ, nodegree, black, hisp, re75.
- (b) Logit: Prob (expstat = 1) = F(age, age2, educ, educ2, married, nodegree, black, hisp, re75, u75, re75², u75*hisp)
- (c) Logit: Prob (expstat = 1) = F(age, age2, educ, educ2, married, nodegree, black, hisp, re75, u75, re75², re75³)
- (d) Logit: Prob (expstat = 1) = F(age, age2, educ, educ2, nodegree, married, black, hisp, re75, re75², u75)
- (e) Logit: Prob (expstat = 1) = F(age, age2, educ, educ2, married, nodegree, black, hisp, re75, u75, educ*re75, age3)
- (f) Logit: Prob (expstat = 1) = F(age, age2, educ, educ2, married, nodegree, black, hisp, re75, u75, educ*re75)
- (g) Weighted Least Squares (same as [a]) using number of matched control observations to treated observations as weights.
- (h) Weights for treated $W_{it} = 1$; Weights for control $W_{ic} = p_i/1-p_i$ averaged over the number of treated observations.
- (i) Least squares regression as (a) including score.

II. Sample Size;

- (j): Number of observations refer to actual number of control and treated (185) used (3)to (5).

A DECISION-THEORETIC APPROACH TO PROGRAM EVALUATION

Abstract

This paper develops a decision-theoretic approach to program evaluation, and applies it to data from California's GAIN experiment (a randomized trial of a welfare-to-work alternative to AFDC). I use a mixture-of-normals specification to model individual earnings in a flexible way, and then make use of individual-specific predictive distributions of earnings to examine a range of decision problems relevant in the context of program evaluation. I show that thinking of program evaluation as a decision problem leads both to much sharper and richer rankings of the treatment and control policies than is available through the alternative methodology.

“Clinton Signs Bill to Cut Welfare and Change State Role”

The New York Times, August 23, 1996

In a sweeping reversal of federal policy, President Clinton on Thursday ended six decades of guaranteed help to the nation's poorest children by signing into law a vast new welfare overhaul requiring the 50 states to deal more directly with the social burdens and the budget expense of poverty.

“Today we are taking a historic chance to make welfare what it was meant to be: a second chance, not a way of life,” Clinton declared in signing the measure, which will affect tens of millions of poor Americans, largely by mandating work requirements and imposing a five-year lifetime limit on welfare help to needy families.

* * *

The heart of the complex new law abolishes Aid to Families With Dependent Children, the government's welfare bulwark, which provides monthly cash benefits to 12.8 million people, including more than 8 million children.

This is to be replaced by a system of block grants and vast new authority for the states, in the hope that they can fashion new work and welfare programs to solve the long-intractable problem of dependence on government.

Job creation will be a particular state burden, since the law requires most poor adults to find a job within two years of first receiving aid.

1. Introduction

Greater Avenues for Independence (GAIN) is a welfare-to-work program initiated in California in 1986 as an alternative to Aid to Families with Dependent Children (AFDC). It is one of many programs that seek to supplement AFDC with additional services and requirements (see Greenberg and Wiseman [1992] who survey 24 such programs), and more generally is one in a long line of social experiments (see Burtless [1995] for a recent survey).

In the context of social experiments, program evaluation is typically carried out by comparing the values of a range of outcomes of interest between groups consisting of individuals assigned into one or the other of the programs being considered (called the treatment and control groups). It is well-known that in randomized trials such comparisons give unbiased estimates of the treatment effect (see Fisher [1935] and Neyman [1935]).^{1,2} Table 1 offers an example in the context of the GAIN experiment, using data drawn from Alameda county. Six different outcomes are compared across the groups participating in GAIN and AFDC (the difference between these two programs will be described in greater detail below).³ The GAIN program is seen to increase earnings and the probability of employment, though not by statistically significant amounts. However, it does increase the total cost to the government by a statistically

¹ There are many relevant issues other than obtaining unbiased estimates of the treatment effect, such as the interaction of local conditions and implementation-related issues on the outcome. See Heckman (1992) and Heckman and Smith (1995).

² The debate on the non-experimental evaluation side is more controversial (see Dehejia and Wahba [1996], Heckman [1989, 1990], Heckman and Robb [1985, 1986], Lalonde [1986], Manski [1989, 1993], and Manski and Garfinkel [1992], *inter alia*).

³ I will use the term GAIN to denote the GAIN package of services (including AFDC), and AFDC to denote the basic AFDC package.

significant amount. Typically such differences are considered for a wide array of outcome variables and for many different subsets of the sample.

The methodology adopted in this paper differs from the typical experimental evaluation approach in two respects. First, modeling earnings at the individual level allows me to introduce individual choice into the evaluation problem. I use a Bayesian model to compare the entire (predictive) distribution of earnings for each individual under each program. Based on this, I imagine allowing the individual to choose which of GAIN or standard AFDC she enters, and am able to ascertain how much the individual values her opportunity to choose. In addition, since this approach bases the individual's decision on her predictive distribution of earnings, both parameter uncertainty from estimation and individual heterogeneity are consistently being accounted for in the evaluation. Second, I model the social choice problem of choosing between programs, building on the individual-level choice. As a result, I am able to consider relevant policy alternatives not addressed in the original experiment or other program evaluations, policies such as allowing individuals to choose which of GAIN or AFDC to join or allowing a career counselor to make this choice.⁴

I apply this methodology to data from the Alameda county portion of the GAIN experiment. There are four findings. First, giving individuals the choice of which program to join leads about 60 percent of the sample to opt for GAIN and 40 percent to opt for standard AFDC. One measure of the value of this choice is that, when individuals are allowed to choose, average earnings are about \$100 to \$200 per quarter

⁴ An exception is Manski (1995), whose motivation is similar but follows a very different approach. See also Manski (1996).

higher than earnings when individuals are not allowed to choose, a large amount for a group whose average earnings are on the order of \$300 per quarter. This stems from the fact that a subset of individuals benefits from treatment on the order of \$300 per quarter, while the complement of this group is, in fact, made worse off by the treatment.⁵ Second, for 80 percent of the sample the choice between programs is clear-cut: the distribution of earnings under one program first-order stochastically dominates the other. Third, the policy of allowing individuals to choose between programs dominates the policy of assigning individuals into GAIN; it achieves higher earnings at a lower cost to the government (even though individuals are not explicitly taking cost into account). Fourth, a policymaker whose concern is individual-level utility will prefer allowing individuals to choose over requiring participation in either GAIN or AFDC.⁶

This suggests that other program evaluations are potentially excluding some of the key issues from consideration. In particular, by dwelling on the statistical significance of the treatment effect, they would fail to detect the fact that for certain individuals and for certain social welfare functions the treatment first-order stochastically dominates the control.⁷ Second, by not considering individual decisions, they are un-

⁵ As will become clear in Section 6, "choice" essentially amounts to offering individuals advice based on observable pre-treatment characteristics. To the extent that individuals have access to unobserved information that could be correlated with the treatment effect, these numbers are lower bounds on the value of choice.

⁶ A highly relevant issue which I do not discuss here is: to what extent can one extrapolate the result to other populations of interest and to other time periods? When treatment effects are estimated at the individual level, one can, in principle, extrapolate to other populations to the extent that they have the same support in the space of pre-treatment variables as the original sample (assuming ignorable assignment). If the model is suitably specified, one can also extrapolate through time. These issues are the subject of future research.

⁷ The finance literature has made a similar point in a very different context. See Kandel and Stambaugh (1996). In addressing the questions, "Are stock market returns predictable and does it matter?" they argue that rather than formulating the question in terms of the statistical significance of the relevant parameters in an econometric model, one should look at the impact of such predictability on the portfolio

able to incorporate policies which allow some degree of individual choice, and certainly in the application I consider these policies turn out to be of central interest.

The paper is organized as follows. Section 2 briefly describes the GAIN program and experiment. Section 3 presents a snapshot of the findings from GAIN. Section 4 lays the foundation for the analysis by setting up the individual decision problem. Section 5 describes the econometric model which I use to operationalize the individual decision problem. In Section 6, I examine the decision problems faced by typical individuals, and determine the value they place on being allowed to choose which program they join. Section 7 discusses the social decision problem and the choice of social welfare functions. Section 8 examines the results of the model at the social level, and Section 9 concludes the paper.

2. The GAIN Program and the GAIN Experiment

The GAIN program began operating in California in 1986, with the aim of “increasing employment and fostering self-sufficiency” among AFDC recipients (see Riccio, *et al.*, [1994]). In 1988, six counties -- Alameda, Butte, Los Angeles, Riverside, San Diego, and Tulare -- were chosen for an experimental evaluation of the benefits of GAIN. A subset of AFDC recipients (single parents with children aged six or older and unemployed heads of two-parent households) were required to participate in the GAIN experiment (see Table 2). For its evaluation Alameda confined itself to the further subset

decision of interest. See also Barberis (1996), Chamberlain and Imbens (1996), Geweke and Keane (1996), and Rossi, McCulloch, and Allenby (1995). In a different context, see Clements, Smith, and Heckman (1994).

of long-term welfare recipients (individuals having already received welfare for two years or more).

Potential participants from the mandatory group were referred to a GAIN orientation session when they visited an Income Maintenance office (either to sign up for welfare or to qualify for continued benefits).⁸ As a result, the chronology of the data and subsequent results is in experimental time, rather than calendar time. No sanctions were used if individuals failed to attend the orientation sessions. However, once individuals started in the GAIN program, sanctions were used to ensure their ongoing participation. At the time of enrollment into the program, a variety of background characteristics were recorded for both treatment and control units including: demographic characteristics; results of a reading and mathematics proficiency test; and data on ten quarters of pre-treatment earnings, AFDC, and food stamp receipts.⁹

Of those who attended the orientation session, a fraction were randomly assigned to the GAIN program,¹⁰ and the others prohibited from participating in GAIN.¹¹ Each of the counties randomized a different proportion of its participants into treatment, ranging from a 50-50 split in Alameda to an 85-15 split in San Diego (see Table

⁸ In some counties AFDC recipients were allowed to volunteer into the GAIN program, but these units are not included in the public use sample.

⁹ Data on AFDC and Food Stamp receipts were taken from each county's welfare records. Data on earnings were taken from the California State Unemployment Insurance Earnings and Benefits Records. Other background characteristics were taken from California's client information ("GAIN-26") form. See Riccio, *et al.*, (1994).

¹⁰ The randomization was (as far as we know) independent of pre-treatment covariates. A different fraction was randomized into treatment in each county. See Table 2.

¹¹ Of course, these individuals could participate in non-GAIN employment-creating activities. The existence of non-GAIN activities is important in interpreting the treatment effect from GAIN. The treatment effect measures the increase in earnings, employment, etc., from the availability of and encouragement (or requirement) to use GAIN-related activities compared to pre-existing employment services. To the extent that AFDC benefits are received by both groups, the real comparison at one level is between the two differing packages of supplementary services and requirements.

2). Because assignment to treatment was random, the distribution of pre-treatment covariates is balanced across the treatment and control groups; the data legend lists each of the covariates.¹² In terms of the chronology of data gathering, “experimental” time (which I also refer to as “post-experimental” or “post-treatment” time) begins when individuals attend the GAIN orientation session. The early stages of experimental time thus coincide with the education and training program of GAIN participants.¹³

In the GAIN experiment, the treatment is participating in the GAIN program; the control is receiving standard AFDC benefits. The GAIN program works as follows: based on test results and an interview with a case manager, participants are assigned to one of two activities. Those deemed not to be in need of basic education are referred to a job search activity (which lasts about three weeks); those who do not find work are placed in job training (which includes vocational or on-the-job training and paid or unpaid work experience, lasting about three to four months). Those deemed to be in need of basic education may choose to enter job search immediately, but if they

¹² Since participants were randomized after registration, we expect to find that the distribution of background characteristics is (up to sampling variation) the same across the treatment and control groups within counties; this is partly true. Let b be a vector of pre-treatment characteristics. Assuming that $\bar{b}_T \sim N(\mu_T, \frac{1}{n_T} \Sigma)$ and $\bar{b}_C \sim N(\mu_C, \frac{1}{n_C} \Sigma)$, where T and C represent the treatment and control groups, n_T and n_C represent the number of treated and control units, and the μ 's are k -vectors, we have

$(\bar{b}_T - \bar{b}_C) \left(\left(\frac{1}{n_T} + \frac{1}{n_C} \right) \hat{\Sigma} \right)^{-1} (\bar{b}_T - \bar{b}_C) \sim \chi^2(k)$. The p-values are: Alameda 0.3364, Butte 0.5346, Los Angeles 0, Riverside 0, San Diego 0, and Tulare 0.0784. There are many possible reasons for the seeming lack of randomization in the three larger counties, including the possibility that randomization proportions may have varied across administrative offices within each county.

¹³ More precisely, individuals were registered in the first quarter of experimental time. This means that in some cases the first quarter of experimental time in fact includes information one or two months prior to the commencement of the experiment. So for example, for an individual who attended an orientation session in February 1989, the first quarter of experimental time is from January to March 1989. Of course, some part of the first and second quarters could be spent participating in treatment activities. Pre-treatment data would cover the ten quarters from July 1986 to December 1988.

fail to find a job they must register for preparation toward the General Educational Development certificate, Adult Basic Education, or English as a Second Language programs (lasting three to four months).¹⁴ Participants were exempted from the requirement to participate in GAIN activities if they found work on their own.¹⁵

3. The GAIN Data in Alameda

From the six counties available in the GAIN experiment, I focus in this paper on Alameda county. This section reviews some features of the data from Alameda. I briefly review the pre-treatment characteristics of the sample, and analyze the treatment effect on earnings and employment status. Finally, I touch on some examples of heterogeneity in the treatment effect.

3.1 Pre-treatment Covariates

Table 3 presents a profile of the participants in Alameda county: 85 per cent of the participants are women, who have on average more than two children; the mean level of education is grade 10; and a quarter of the participants have previously participated in training programs. The average level of pre-treatment earnings is very low, ranging from \$150 to \$190 per quarter, but because 87 percent of pre-treatment earnings are

¹⁴ The public use data do not contain information on each individual's participation in the various components of the program.

¹⁵ Note that only about eight-five per cent of the treated units actively participated in any GAIN activities (though by virtue of being in the GAIN sample they did attend an orientation meeting); the balance satisfied the requirements of the GAIN program on their own (in most cases finding employment within the first two or three quarters of experimental time). Thus, as observed earlier, this is important in interpreting the treatment effect as a comparison between earnings, employment, etc., when individuals are required to find a job or to participate in GAIN-related activities and when they are not obliged to find jobs and only pre-existing employment-related services are available.

zero, the average of non-zero pre-treatment earnings is higher, on the order of \$1,110 per quarter.

3.2 The Treatment Effect

To get a sense of the pattern of the zeros in post-treatment earnings, Table 4 documents the proportion of individuals with zero earnings for the treatment and control groups for a given (one-period) employment history. Between 70 percent and 82 percent of units with zero earnings in a given period have zero earnings in the next period; though this proportion declines for both treated and control units over the 13 quarters of experimental time, it declines more for treated units. On the other hand, only 3 percent to 4 percent of (both treated and control) units with positive earnings in a given period have zero earnings the following period. Thus, the treatment effect in terms of employment status is given by the increased proportion of individuals who move from unemployment to employment.¹⁶

Table 5 explicitly shows the impact of GAIN on the probability of unemployment for the 13 post-treatment quarters. Consistent with the results of Table 4, in the first post-treatment quarter GAIN's impact is negative; this is not surprising since treatment units are participating in training activities in the first quarter. After a few quarters the treatment effect increases to the range of 4 percent to 6 percent (and is statistically significant).

¹⁶ For more detailed analysis of this issue, see Ashenfelter and Card (1985).

Figure 1 displays the average treatment effect on earnings for 13 quarters for each of the six counties. The figures illustrate the basic conclusion that GAIN participants in most counties enjoyed higher post-treatment earnings than their control counterparts. The benefits that GAIN participants in Alameda enjoyed increased over time.

Table 6 explores the impact of participation in GAIN on earnings through the use of OLS and Tobit regression analysis, both of which are implemented with and without covariates. In column 2 we see a simple period-by-period difference in means. This reveals a pattern similar to that shown in Table 5: the treatment effect starts out negative, and then increases in most of the subsequent periods (with a number of statistically significant effects). As expected, these results change only slightly with the addition of covariates (column 4). Columns 1 and 3 use Tobit regressions to estimate average treatment effects, which might provide a better fit to the data, because of the mass point at zero earnings. The results shown in columns 1 and 3 are very similar to those shown in columns 2 and 4: well within one standard error, often much closer. The fact that the Tobit does not produce a dramatically different set of results is useful to bear in mind in subsequent sections where it forms the basis of the statistical model.

3.3 A Heterogeneous Treatment Effect

The assumption of a constant treatment effect across all individuals is very restrictive and unrealistic. The average treatment effects considered in the previous section potentially embody an array of heterogeneous treatment effects. Two examples illustrate the point. Figure 2 explores the interaction between the treatment effect and the score on

the reading test: individuals with a score of 200 or more on the test enjoy a higher treatment effect (although the standard error is quite large). In Figure 3, we see that individuals who have previously participated in job training programs enjoy a higher treatment effect as well. Many such interactions potentially exist in the data. Thus, in subsequent sections the econometric model will allow for a heterogeneous treatment effect.

4. The Individual Choice Problem

As indicated in Section 1, the decision-theoretic analysis of GAIN rests on two foundations. The first is a model of the individual choice between participating in GAIN and in AFDC; this section lays out the individual choice problem. The second is an econometric model which predicts individuals' distributions of earnings; this is presented in Section 5. The two are combined in Section 6.

In the GAIN experiment, participants were randomized into treatment and control, so that no individual choice was exercised. The question which I focus on is how an individual similar to those who participated in the experiment would decide which program to join if she were offered the choice between the available programs, which we take to be GAIN and AFDC. This thought experiment includes those who in fact participated in the experiment (and were not offered the choice).

Of course, after individuals have chosen either GAIN or AFDC, they face a multitude of other choices, such as what kind of training to seek, where to look for a job, and how hard to try in these activities. My formulation does not address these

subsequent choices. Once committed to GAIN or AFDC, I take an individual's distribution of earnings (and other outcomes) to be a function of observed pre-treatment characteristics and, of course, treatment status. I will further simplify the analysis by assuming individuals choose between programs on the basis of the distribution of their earnings under each program (and, in a few cases in Section 8, AFDC and food stamps as well).

An individual is identified by: a vector of pre-treatment covariates, Z_i , which includes pre-treatment income history; a vector of parameters defining the budget set, W_i ; and by preferences. I assume that individual preferences are characterized by a vector of preference parameters, p_i . Given some $d_i \equiv (Z_i, W_i, p_i, j)$, where $j=1$ (GAIN) or 0 (AFDC), the individual makes optimal choices for job search, effort, etc. This results in a sequence of outcomes such as earnings and welfare receipts, captured in the reduced form expression $\{Y'_i(d_i)\}_{t=1}^{\tau}$, where $Y'_i(d_i)$ is the vector of outcome variables (e.g. earnings, AFDC benefits, and food stamps) and where $i=1, \dots, I$ indexes individuals and $t=1, \dots, \tau$ indexes time. The formulation so far ignores uncertainty. For a given d_i and a given vector of choices for job search, effort, etc., there is a distribution of each outcome for each time period. For each d_i when the control variables are chosen optimally we denote the resulting distribution of outcomes over each of the τ periods by $\Lambda_i(d_i, t)$, $t=1, \dots, \tau$.

To simplify the analysis I will assume that an agent's utility in a given period is based on her total income in that period. (Operationally I will focus on earnings, but at

a few points in Section 8, I add in AFDC receipts and food stamps, ignoring the fact that these three sources of income are not genuinely fungible.¹⁷ For notational consistency they are included in the vector of outcomes.) Assuming time-separable Von Neumann-Morgenstern preferences, the agent's utility is given by:

$$U(d_i) = \sum_{t=1}^T \alpha^t \int u(Y_t^i(d_i)) d(\Lambda(d_i, t)),$$

where u is the individual's (Bernoulli) utility function and α is the discount factor. We are integrating utility, which is a function of $Y_t^i(d_i)$, over the distribution Λ .

Imagine that individuals are offered the choice between participating in the GAIN program and receiving the standard AFDC benefits. Then we present the individual with the following choice:

$$\max_{j \in \{1,0\}} \{U(Z_i, W_i, p_i, j)\},$$

where $j=1$ corresponds to GAIN and $j=0$ corresponds to AFDC.

In order to implement this choice framework empirically, we will need a model of $\Lambda(d_i, t)$ for each individual.

5. A Model of the Data

5.1 The Statistical Model

In my implementation I model only earnings, given that AFDC and food stamps are in principle determined by non-stochastic rules. I denote the earnings component of the

¹⁷ Food stamp aid in general is tied to a specific use. However, in the case of California, food stamp benefits have been converted to cash that is included in the state supplementary payments to SSI. Of course, stigma may attach to welfare receipts more generally; and work effort must be expended for earnings.

vector of outcome variables, $Y_i^t(d_i)$, as Y_{ij}^t , where $j=1$ (GAIN) or 0 (AFDC). Y_{i1}^t is interpreted as individual i 's earnings in period t if she was in GAIN, and Y_{i0}^t as her earnings if she was in AFDC; obviously one of these is counter-factual. Thus, observed earnings are defined as:

$$Y_{it} = T_i Y_{i1}^t + (1 - T_i) Y_{i0}^t,$$

where T_i is a treatment indicator ($=1$, if individual i was assigned to GAIN, and $=0$, if she was assigned to AFDC). Realizations of the random variable are denoted in lower case, y_{it} .

A key feature of the distribution of earnings, which influences the model choice and was highlighted in Section 3, is the presence of a mass point in the distribution of earnings at zero. The strategy adopted is to model the probability of positive earnings and the distribution of positive earnings separately. For the former, a probit model is used for an indicator for positive earnings, and for the latter log earnings are modeled using a mixture of normals.

Define an indicator variable, y_{it}^* , for when earnings are positive:

$$y_{it}^* = \begin{cases} 1 & \text{if } y_{it} > 0 \\ 0 & \text{if } y_{it} = 0, \end{cases}$$

and consider a latent variable, Y_{it}^{**} , determining which value Y_{it}^* takes on:

$$y_{it}^* = \begin{cases} 1 & \text{if } Y_{it}^{**} > 0 \\ 0 & \text{if } Y_{it}^{**} < 0. \end{cases} \quad (1)$$

Assume

$$Y_u^* \left| \left\{ Y_u^* = y_u^* \right\}_{t=1}^{\tau-1}, \left\{ X_u = x_u \right\}_{t=1}^{\tau}, \beta \sim N(x_u \beta, 1), \quad (P)$$

for $i=1, \dots, I$ and $t=1, \dots, \tau$ (the probit model). The vector of explanatory variables is given by $x_{it} = \{1_{1it}, \dots, 1_{\tau it}, [1_{1it} \dots 1_{\tau it}] \cdot T_i, Z_i \cdot T_i, 1(Y_{i,t-1} = 0)\}$, a $(2\tau + k + 1)$ -vector of regressor variables. $[1_1 \dots 1_\tau]$ is a set of indicator variables for each quarter of post-experimental time ($1_{jit} = 1$ if $t=j$, $=0$ otherwise), giving each period its own intercept. The treatment indicator is interacted with $[1_1 \dots 1_\tau]$. Since each period corresponds to experimental, rather than calendar, time, the treatment dummies produce a profile of the treatment effects over thirteen quarters. Interactions between k exogenous regressors, Z_i , and the treatment indicator are also included, which allow treatment effect to vary with observable pre-treatment characteristics; these include: indicators for the age and number of children, ethnicity, educational attainment, score on the reading and mathematics tests, sex, an indicator for previous participation in other training programs, ten periods of pre-treatment earnings history, and a calendar time trend. As well, the model allows for persistence in a very simple form, through lagged earnings, in particular an indicator for zero lagged earnings.

For positive earnings, a mixture-of-normals likelihood is adopted for log earnings:

$$\begin{aligned} \tilde{Y}_u \left| \left\{ Y_u^* = y_u^* \right\}_{t=1}^{\tau-1}, \left\{ X_u = x_u \right\}_{t=1}^{\tau}, \gamma, \theta, \sigma, \tau, p \\ \sim pN(x_u \gamma, \sigma^2) + (1-p)N(x_u \gamma + \theta, \tau^2), \end{aligned} \quad (M)$$

where $\{\tilde{Y}_i\}$ are the log of the positive elements of $\{Y_i\}$. Note that (M) conditions only on the indicator for positive lagged earnings (rather than the level of lagged earnings). Priors are discussed in the next section.

5.2 The Estimation Procedure

The posterior distribution of the parameters of the two models (P) and (M) are obtained through Gibbs sampling procedures. The Gibbs sampler is a Markov chain Monte Carlo simulation technique which simulates the joint posterior of the parameters of the model. Instead of drawing directly from the joint posterior (often intractable), it draws successively from the posterior of each parameter (or block of parameters) conditional on all of the other parameters. For any starting values (given certain conditions), these draws will eventually converge to draws from the true posterior (see Geman and Geman [1984], Gelfand and Smith [1990], Tanner and Wong [1987], as well as Chamberlain and Imbens [1996], Gelman, Carlin, Stern, and Rubin [1996], and Rossi, McCulloch, and Allenby [1995]).

In many cases, such as the probit and mixture models, the task of drawing from the joint posterior is simplified by augmenting the parameter space of the model. For the probit model, the parameter space is expanded to include the latent variables y_{it}^{**} ; conditional on these, the probit model reduces to a standard regression model, and conditional on all other parameters, it is easy to draw from the posterior distribution of y_{it}^{**} . For the mixture model, the parameter space is expanded to include indicators for

which component of the mixture each observation is drawn from; again, conditional on these indicators, it is easy to update the other parameters, and *vice versa*.

5.2.1 The Probit Model

The probit model is closely related to the Tobit model, for which the Gibbs sampling algorithm has been worked out by Chib (1992). The modification for the probit context is immediate (see also Albert and Chib [1993] and Chib and Greenberg [1994]).

We stack the observations in the form:

$$y_{it}^{**} = x_{it}\beta + \varepsilon_{it}$$

for $y_{it}^{**} \in y_i^{**} = (y_{i1}^{**}, y_{i2}^{**}, \dots, y_{i\tau}^{**})'$, for $i=1, \dots, I$, $I=1360$, and $\tau=13$. Of course, y_{it}^{**} is not observed. The key to the Gibbs sampling procedure is that conditional on β , it is easy to draw from the posterior distribution of y_{it}^{**} , and then using these to draw from the posterior of β . Given diffuse priors for β and an arbitrary starting value $\beta^{(0)}$, the Gibbs sampling scheme is then:

(1) Conditional on $\beta^{(j)}$, draw values for y_{it}^{**} : for $\{it : y_{it}^* = 0\}$, from the negative portion of a normal distribution with mean $x_{it}\beta^{(j)}$ and variance 1, and for $\{it : y_{it}^* = 1\}$, from the positive portion of the same distribution. Denote the filled-in dependent variable $y_{i,z}^{(j+1)}$, so that $y_{i,z}^{(j+1)} = (y_{i1,z}^{(j+1)} \dots y_{i\tau,z}^{(j+1)})'$.

(2) Conditional on $Y_z^{(j+1)} = (Y_{1z}^{(j+1)}, \dots, Y_{Iz}^{(j+1)})'$, draw for $\beta^{(j+1)}$ from

$$N(\hat{\beta}^{(j+1)}, (X'X)^{-1}),$$

where $\hat{\beta}^{(j+1)} = (X'X)^{-1}X'Y_z^{(j+1)}$, with $X_i = (x_{i1} \dots x_{it})'$, and $X = (X_1', \dots, X_t)'$.

From an arbitrary starting value, this is iterated 2000 times, producing $(Y_z^{(j)}, \beta^{(j)})$. The first 500 iterations are discarded, leaving 1500 draws from posterior distribution of the parameters, which will be indexed $j=1, \dots, 1500$.¹⁸

5.2.2. The Mixture Model

When proper parametric priors are used, the Gibbs sampler for a mixture model is straightforward. It is, however, essential that priors be informative to some extent. Even though with parametric priors it is possible to identify the mixture, without sufficient prior information the computational algorithm can break down (see Robert [1996], and also Geweke and Keane [1996]). In setting up the Gibbs sampler, prior information is incorporated in two ways. First, the prior requires that $\theta > 0$; a normal prior (with mean zero and large variance) is used, and truncated suitably. Second, prior information is provided about the variance of each component, σ^2 and τ^2 , in the form of the number of prior observations and the specified prior variance for each component. For σ^2 , the prior is 30 observations with a sample variance of 1.5, similar to the variance arising in a single component normal model for this data. The 5th and 95th percentiles of the prior are 1.03 and 2.41. Though this is reasonably tight, in the updating procedure the weight on the prior will be very low (see the discussion below). For

¹⁸ Several diagnostics suggest that throwing out the first 500 runs is sufficient to converge to draws from the posterior. These include considering a wide variety of starting points, running the sampler for more iterations, and comparing the mean of the posterior of the parameters with maximum likelihood estimates of the same parameters.

τ^2 , the prior is 30 observations with a sample variance of 0.25, with the aim of picking up peaked segments observed in the empirical distribution. (The 5th and 95th percentiles of this prior are 0.17 and 0.41.)

A flat prior is used for γ . The prior on p is expressed in terms of the number of prior observations seen from each component of the mixture; these are set to 1 for each component. The priors can be summarized as: $p(\gamma) \propto c$, $\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2)^{19}$, $\nu_0 = 30$, $\sigma_0^2 = 1.5$, $\tau^2 \sim \text{Inv-}\chi^2(n_0, \tau_0^2)$, $n_0 = 30$, $\tau_0^2 = 0.25$, $\theta \sim N(\theta_M, \theta_V) \cdot I(\theta > 0) / (1 - \Phi(-\theta_M / \theta_V^{1/2}))$, $\theta_M = 0$, $\theta_V = 100$, $p \sim \text{Beta}(\alpha_1, \alpha_2)$, $\alpha_1 = \alpha_2 = 1$.

The Gibbs chain generates a latent variable, z_i , which is an indicator for which of the two components a given observation originates from (1 if it is from the first component, 0 for the second). Conditional on z_i , the chain is straightforward, and conditional on all other parameters it is easy to update the z 's (see Robert [1996]). Take starting values $z^{(j)} = \{z_i^{(j)}\}$, $\sigma^{(j)}$, $\tau^{(j)}$, $p^{(j)}$, $\theta^{(j)}$, and $n_1 = \sum_i z_i^{(j)}$, $n_2 = \sum_i (1 - z_i^{(j)})$:

(1) Conditional on $z^{(j)}$, $\sigma^{(j)}$, $\tau^{(j)}$, $p^{(j)}$, $\theta^{(j)}$, draw for $\gamma^{(j+1)}$ from

$$N(\tilde{\gamma}_{GLS}, (\tilde{X}^T \Lambda^{-1} \tilde{X})^{-1}), \text{ where } \Lambda = \begin{bmatrix} (\sigma^2)^{(j)} I_{n_1} & 0 \\ 0 & (\tau^2)^{(j)} I_{n_2} \end{bmatrix}, \tilde{\gamma}_{GLS} = (\tilde{X}^T \Lambda^{-1} \tilde{X})^{-1} \tilde{X}^T \Lambda^{-1} Y^{(j)},$$

$$Y^{(j)} = [Y_1^{(j)}, Y_2^{(j)}], \quad Y_1^{(j)} = \{\tilde{Y}_i : z_i^{(j)} = 1\}, \text{ and } Y_2^{(j)} = \{\tilde{Y}_i - \theta^{(j)} : z_i^{(j)} = 0\}, \text{ with } \tilde{Y}_i \text{ formed}$$

from the log of the stacked, positive elements of y_{it} , and \tilde{X} formed from the elements of X corresponding to $Y^{(j)}$.

¹⁹ If $\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2)$, then $(1/\sigma^2) \sim \chi_{\nu_0}^2 / \nu_0 \sigma_0^2$.

(2) Conditional on $z^{(j)}$, $\gamma^{(j+1)}$, $\tau^{(j)}$, $p^{(j)}$, $\theta^{(j)}$, draw for $(\sigma^2)^{(j+1)}$ from $\text{Inv} - \chi^2\left(v_0 + n_1, \frac{v_0\sigma_0^2 + n_1s_1^2}{v_0 + n_1}\right)$, where $s_1^2 = \sum_{i=1}^{n_1} \left(\left(Y_1^{(j)}\right)_i - X_i^{(1)}\gamma\right)^2 / n_1$ and $X^{(1)}$ are the elements of \tilde{X} corresponding to $z^{(j)}=1$. We see that the posterior weights the prior and the sample information by the number of prior and sample observations; with over 3000 observations, clearly most of the weight will be on sample information.

(3) Conditional on $z^{(j)}$, $\gamma^{(j+1)}$, $\sigma^{(j+1)}$, $p^{(j)}$, $\theta^{(j)}$, draw for $(\tau^2)^{(j+1)}$ from $\text{Inv} - \chi^2\left(n_0 + n_2, \frac{n_0\tau_0^2 + n_2s_2^2}{n_0 + n_2}\right)$, where $s_2^2 = \sum_{i=1}^{n_2} \left(\left(Y_2^{(j)}\right)_i - X_i^{(2)}\gamma^{(j+1)} - \theta^{(j)}\right)^2 / n_2$ and $X^{(2)}$ are the elements of \tilde{X} corresponding to $z^{(j)}=0$.

(4) Conditional on $z^{(j)}$, $\gamma^{(j+1)}$, $\sigma^{(j+1)}$, $\tau^{(j+1)}$, $p^{(j)}$, draw for $\theta^{(j+1)}$ from $N\left(\frac{\frac{1}{\theta_v}\theta_M + \frac{n_2}{(\tau^2)^{(j+1)}}\bar{Y}_2}{\frac{1}{\theta_v} + \frac{n_2}{(\tau^2)^{(j+1)}}}, \left(\frac{1}{\theta_v} + \frac{n_2}{(\tau^2)^{(j+1)}}\right)^{-1}\right)$, where $\bar{Y}_2 = \sum_{i=1}^{n_2} \left(\left(Y_2^{(j)}\right)_i - X_i^{(2)}\gamma^{(j+1)}\right) / n_2$.

(5) Conditional on $z^{(j)}$, $\gamma^{(j+1)}$, $\sigma^{(j+1)}$, $\tau^{(j+1)}$, $\theta^{(j+1)}$, draw for $p^{(j+1)}$ from $\text{Beta}(\alpha_1 + n_1, \alpha_2 + n_2)$. With $\alpha_1 = \alpha_2 = 1$, little weight will be placed on the prior.

(6) Conditional on $\gamma^{(j+1)}$, $\sigma^{(j+1)}$, $\tau^{(j+1)}$, $\theta^{(j+1)}$, $p^{(j+1)}$, draw for $z^{(j+1)}$ as follows.

Create $p_{u,k} = \frac{p_k^{(j+1)} f_k(\tilde{Y}_u | X_u, \gamma^{(j+1)}, \sigma^{(j+1)}, \tau^{(j+1)}, \theta^{(j+1)})}{\sum_{k=1,2} p_k^{(j+1)} f_k(\tilde{Y}_u | X_u, \gamma^{(j+1)}, \sigma^{(j+1)}, \tau^{(j+1)}, \theta^{(j+1)})}$, where $f_1(\cdot)$ and $f_2(\cdot)$ are

the densities for each observation of the first and second components of the mixture as

given above in (M) and $p_k^{(j+1)} = p^{(j+1)}$ ($= 1 - p^{(j+1)}$) if $k=1$ (if $k=2$). For each it , draw $u \sim \text{Uniform}$, and set $z_u^{(j+1)} = 1$ if $u \leq p_{u,1}$, and 0 otherwise.

From an arbitrary starting value, this is iterated 2000 times, producing $(z^{(j)}, \gamma^{(j)}, \sigma^{(j)}, \tau^{(j)}, \theta^{(j)}, p^{(j)})$, $j=1,2,\dots,2000$. The first 500 iterations are discarded, leaving 1500 draws from posterior distribution of the parameters.

5.3 The Predictive Distribution

For each model, we can use the draws from the posterior distribution of the parameters to simulate the (posterior) predictive distribution of the outcomes, i.e., from (P) a distribution for the probability of positive earnings and from (M) a distribution for the level of positive earnings, in which we integrate out for the uncertain parameters.

Consider the $(I+1)$ st individual, who is exchangeable with (and may, or may not, have been one of) the individuals in the original sample. Denote by X_{I+1} the exogenous covariates corresponding to $X_{I+1,1}, \dots, X_{I+1,13}$, so this includes pre-treatment covariates and earnings and the entire set of period indicators for periods 1 to 13 defined in Section 5.1, but does not include earnings information in post-treatment periods. As well, X_{I+1} includes a treatment indicator, which we take to be 1 (or 0) to imagine $I+1$ as part of the treatment group (control group).

Taking both the probit and mixture model together, we wish to compute to the joint predictive density of earnings across 13 periods of post-experimental time. Thus, we wish to compute:

$$\begin{aligned}
p(Y_{I+1,1}, Y_{I+1,2}, \dots, Y_{I+1,13} | Data, X_{I+1}) = \\
\int \{ p(Y_{I+1,1} | X_{I+1}, \Theta) p(Y_{I+1,2} | Y_{I+1,1}, X_{I+1}, \Theta) \\
\cdots p(Y_{I+1,13} | Y_{I+1,12}, X_{I+1}, \Theta) \} p(\Theta | Data) d\Theta,
\end{aligned}$$

where $\Theta = \{\beta, \gamma, \theta, \sigma, \tau, \rho\}$. Conditional on parameters, we use the likelihoods defined in (P) and (M):

$$p(Y_{I+1,t} | Y_{I+1,t-1}, X_{I+1}, \Theta) = \begin{cases} \Pr(Y_{I+1,t} = 0 | Y_{I+1,t-1}^*, X_{I+1}, \Theta) & \text{for } Y_{I+1,t} = 0 \\ \Pr(Y_{I+1,t} > 0 | Y_{I+1,t-1}^*, X_{I+1}, \Theta) \\ \quad \times p(Y_{I+1,t} | Y_{I+1,t-1}^*, X_{I+1}, \Theta) & \text{for } Y_{I+1,t} > 0, \end{cases}$$

where the probability of positive earnings corresponds to the probit model and density for positive earnings to the mixture model. Finally, we integrate out for the unknown parameters using their posterior distribution.

Conditional on parameters, we can simulate the distribution by drawing for $Y_{I+1,1}$, and substituting forward for earnings in subsequent periods. To obtain the predictive distribution, we must account for parameter uncertainty, and, thus, we use draws from the posterior distribution of Θ (obtained from the two Gibbs samplers outlined in the previous section). For each draw from $p(\Theta | Data)$, we can simulate the distribution from the likelihoods. These draws are then weighted by $p(\Theta | Data)$ to obtain the predictive distribution.

Thus, using this procedure, we have the entire joint predictive distribution of earnings for individual $I+1$, from periods 1, ... 13. Given the values of covariates, this predictive distribution represents the distribution of possible values for earnings incorporating parameter uncertainty.

5.4 The Choice and Fit of the Model

One issue in model choice, given the use of probit and mixture-of-normals likelihoods, is the use of the indicator for positive lagged earnings, rather than simply lagged earnings. For the GAIN data, a standard AR(1) does not capture the dynamics of the earnings process well. In particular, for individuals with positive earnings in a given period, in the following period it tends to over-estimate the persistence in their level of earnings. When extrapolating 10 or 12 periods forward, the process dramatically overstates the level of earnings. Hence, I use an alternative specification, with an indicator for when lagged earnings are positive.

Another major issue is the choice of likelihood. Figures 4 and 5 give a sense of the fit of the model. These figures show the density of the empirical distribution of earnings for treated and control units (estimated through a histogram), averaged over the 13 post-treatment quarters, and plot the density of the predictive distribution of earnings. The joint predictive distribution of earnings across 13 periods is averaged over the individuals in the Alameda sample, and is then averaged over 13 periods, to produce the average predictive distribution which is plotted.

The use of a probit to model the probability of positive earnings is not controversial (other choices such as a logit yield similar results, and the probit is computationally convenient). As we see from the figures, the model fits the mass point at zero with substantial accuracy. For positive earnings, the figures illustrate the value of using a mixture of normals. The empirical distributions of log earnings for treated and

control units are well approximate by the mixture, because, for both distributions, the mixture is able to reflect the skewness of the empirical distribution. In fact, with a single normal, the predictive distribution of earnings would have a much thicker upper tail than the empirical distribution.

6. The Individual Decision Problem

Imagine that an individual in the sample is offered the choice between GAIN and AFDC. Of course, because this individual is drawn from the data, she was in fact assigned to one of the two programs. But as a thought experiment, either imagine setting her back at time zero and letting her choose, or offering the choice to an individual with the same observable characteristics.²⁰ Having estimated the model in Section 5, suppose that a counselor uses this framework to advise her regarding the best choice to make. The relevant distribution for this purpose is the predictive distribution described in Section 5.3.

6.1 The Predictive Distribution

Based on the model outlined in Section 5.1, I implement the Gibbs sampling procedure to sample from the posterior distribution of the parameters of the model. With the pos-

²⁰ In the individual choice problems that I examine, I consider the predictive distributions of earnings under both treatment and control, ignoring the fact that if the individual under consideration did in fact participate in the treatment one “knows” her earnings under the treatment. This is reasonable because, given the observed data, the predictive distribution captures the uncertainty regarding the value of earnings if an individual with the same covariates were put through the program.

terior of the parameters in hand, I can plug in any given value of pre-treatment covariates to obtain the predictive distribution of earnings, denoted $\Lambda|_{X,T=1}$ and $\Lambda|_{X,T=0}$.

Of course the data available from the GAIN experiment does not include important variables such as wealth. In order to identify draws from the predictive distribution with draws from $\Lambda(d_i, t)$, the key step is the assumption of exchangeability conditional on covariates; i.e., that the joint probability density $p(\{(x_i, y_i)\}_{i=1, \dots, I, I+1})$ is invariant to permutations of the indices. If for the I individuals in the sample Y_{it} follows the distributions in (P) and (M), then we assume that, for the individual under consideration, $I+1$, $Y_{I+1, t}$ is drawn from (P) and (M). Intuitively, although $(I+1)$'s wealth and preference parameters may differ from that of individuals in the sample with the same vector of observable covariates, I assume that it is still reasonable to consider them as drawn from the same distribution.²¹

6.2 Two Typical Examples

Tables 7a and 7b list the pre-treatment covariates of two individuals from the Alameda county sample for whom we see typical patterns in the distributions of earnings under treatment and control (one being a clear winner from GAIN and the other a clear loser). The first is a woman ("Ms. Thirteen Fifty-Three"), aged 42. She is the head of a single-parent household; has one child between the age of 12 and 18; and has com-

²¹ The assumption of exchangeability conditional on covariates is not unique to my application. This assumption, or some alternative, is needed any time we want to extrapolate from a dataset to a new situation. Unless information on preferences is explicitly solicited, one must use exchangeability assumption or some other equally strong assumption specifying the form of the lack of exchangeability.

pleted high school and three years of some additional training. Her earnings history shows she was employed in 2 of 10 quarters prior to the experiment. The second individual ("Ms. One") is a 32-year old woman, the head of a single-parent household; has one child between the age of 6 and 11, and two children between 12 and 18; and has an educational attainment of grade 9. Her earnings history shows zero earnings in each of the 10 pre-treatment periods. I consider each individual in turn.

Table 8a shows the probability of positive earnings and the mean and standard deviation of the predictive distribution of earnings (including the mass point at zero earnings) for each period under both treatment and control for Ms. Thirteen Fifty-Three. Her mean earnings under treatment are lower than under the control in the first three periods, but her probability of positive earnings in these periods is higher under the treatment. Intuitively, even though she is more likely to find job under treatment, in the first three periods it is not likely to pay more than a job she might find without having been treated. However, from the fourth period on both earnings and the probability of employment are higher under the treatment. The profile of the treatment effect is increasing, in a pattern similar to that depicted in Figure 1 for Alameda county on average. However, the standard deviation of control earnings is higher than that of treatment earnings, and the difference between the treatment and control earnings is small compared to the magnitude of the standard deviation.²² One might wonder whether her risk attitude would affect her preference between the two programs.

²² Of course, the standard deviation of the predictive distribution is not very informative, because of the mass point in the distribution. This is another reason to examine the entire distribution of earnings, which we do below.

Based on the underlying predictive distributions, we could compute expected utilities to determine what choice this individual would make. Figure 6 depicts the cumulative distribution functions for the predictive distribution under treatment and control for each of the 13 periods. They summarize all the information available for the individual to consider, by combining the predicted probability of positive earnings with the distribution of positive earnings. The figures are very revealing. In period 1 to 3, the control earnings (almost) first-order stochastically dominate the treatment earnings; and in periods 3 to 13 the treatment unambiguously dominates the control. This is a simple illustration of the fact that even when the means of the two distributions under consideration are not very different in terms of t-statistics, the underlying decision between the two may be clear-cut. In this case, as long as the individual does not have an extremely high discount rate, we would advise her to join the treatment.²³ Whether she, in fact, would follow this advice depends on the factors she considers, which may or may not be part of the model.

For the second individual, matters are different. We see in Figure 7 that her distribution of earnings under the control first-order stochastically dominates her distribution under treatment in each period. As long as she prefers more earnings to less, she unambiguously would be advised not to participate in GAIN.

Of course, first-order stochastic dominance does not suffice to compare all the distributions which arise. In general, expected utility comparisons would be required.

²³ What would we advise on whether she should go to school, train, etc.? Since our data contains no information regarding this issue, we assume that the choices she makes would not be too different from those of similar people represented in the sample.

6.3 The Importance of Accounting for Uncertainty

A natural question which arises from the preceding analysis is: would similar decisions have been reached if uncertainty had not been accounted for as comprehensively? In particular, one might imagine using the model described in Section 5, but, rather than using the full posterior distribution of the parameters, using point estimates and treating them as though they were the true parameters. Of course, even ignoring parameter uncertainty, the intrinsic uncertainty embodied in the likelihoods of (P) and (M) has to be taken into account. Columns 1 to 4 of Table 8b consider such an exercise for Ms. Ten, whose characteristics are given in Table 7c. Columns 1 to 4 present the distribution of her earnings in each of the 13 quarters, ignoring parameter uncertainty, but still accounting for the uncertainty conditional on parameters. In contrast columns 5 to 8 present the posterior distribution of her earnings, in which parameter uncertainty is accounted for. The means of the two sets of predictions are broadly similar, as are the standard deviations. Of course, since the underlying distributions are highly non-normal, the first two moments are only partially informative. A direct means of comparing the distributions is through expected utilities. An expected utility comparison for log utility and CRRA preferences (with coefficient of relative risk aversion equal to 3) reveals that the advice to Ms. Ten sharply differs depending on which distribution is used. Ignoring parameter uncertainty, her expected utility is higher under GAIN; instead her expected utility under the predictive distribution is higher under AFDC.

Of course this example was chosen precisely for the reason that ignoring uncertainty leads to a different decision (or to different advice) than accounting for uncertainty. In cases where the two distributions are starkly different, ignoring uncertainty would not, typically, lead to a change in decision. For the overall sample from Alameda, uncertainty affects the decisions of about ten percent of individuals.

6.4 Heterogeneity and the Value of Choice

We could consider such decision problems for a wider array of individuals. The differences in the results would be based on the underlying heterogeneity in the treatment effect. One view of this is presented in Table 9. Table 9 supposes that each of the 1360 individuals in the Alameda sample were offered the choice between GAIN and AFDC, rather than having been randomly assigned.²⁴ Like the previous examples, imagine that these individuals are advised on the choice between GAIN and AFDC based on the predictive distributions of their incomes in each period under each program (where income includes both labor earnings and welfare receipts).²⁵ For the moment, I suppose that this advice is based on (time-additive) constant relative risk aversion preferences (with a coefficient of relative risk aversion equal to 3).

²⁴ Seven individuals are excluded from the original sample, because of apparent coding errors in their covariates. These seven individuals are either coded as having seventy children or a previous hourly wage of more than \$300.

²⁵ As mentioned in the previous section, I simplify the analysis by drawing only from the posterior distribution for earnings. When needed, I use a deterministic rule to fill in suitable amounts for AFDC and food stamps. For AFDC, only six individuals in the Alameda sample are earning both zero income and zero AFDC benefits. About 10 percent of the sample have both positive earnings and positive AFDC benefits. The balance, about 90 percent of the sample, are receiving full AFDC benefits, since they have zero earnings. For the moment I use a fixed-effects regression model to predict AFDC and food stamp benefits for these individuals.

Table 9 presents the mean of expected post-treatment earnings under GAIN and AFDC and the mean of pre-treatment covariates for two sub-groups of the sample: the group that “chooses” (i.e. is advised to enter) GAIN over AFDC (for whom expected utility from post-treatment income is higher under GAIN than under AFDC) and the complement of this group.²⁶ Only 40 percent of the sample fare better (in an expected utility sense) under GAIN than AFDC. The comparison of these two groups is revealing. Those benefiting from GAIN have fewer children on average (except between the age of 6 and 11), have higher scores on the reading and mathematics tests, have a higher level of educational attainment, are five years younger, and half of them have previously participated in training programs. Of particular note is the dramatic difference in the level of pretreatment earnings. Individuals who are advised to go into GAIN if given the choice, enter the program with average quarterly earnings on the order \$300, compared with those who would not opt for GAIN whose quarterly earnings are on the order of \$50.

In addition, Table 9 gives one possible view of the value of choice to individuals in the sample. For individuals with high pre-treatment earnings (who are the primary beneficiaries of GAIN), the value of choice is high: their expected post-treatment earnings under GAIN are more than \$300 per quarter higher than under AFDC. In contrast, the value of choice is much lower for individuals with low pre-treatment earnings. The difference between their expected quarterly earnings in the two programs is \$20 per quarter.

²⁶ The administrative cost of GAIN is estimated at \$3638 for 13 quarters, based on chapter 3 of Riccio, *et. al* (1996).

7. The Social Choice Problem

Thus far the analysis has focused on the individual choice between GAIN and AFDC. Sections 7 and 8 take the next step by asking how the policymaker would decide which program or combination of programs to make available, given the pattern of individual choice. There are two aspects to this decision. The first is the set of policies that the policymaker has under consideration, where policies are rules which determine each individual's assignment to treatment. The second aspect is the set of criteria (referred to as social welfare functions) that the policymaker uses to reach this decision. Section 8 will use both of these, evaluating a range of policies under a range of social welfare functions.

7.1 The Policy Alternatives

A policy is a rule that determines each individual's assignment to treatment (here, whether they go into GAIN or AFDC), hence the vector $\{T_i\}_{i=1}^P$ (where i indexes the population of interest, $1, \dots, P$). Of course, such rules may or may not allow the individual to choose her assignment to treatment (again where choice is in the sense of Section 6, i.e., being advised based on expected utility comparisons). Four alternatives are considered:

- (1) All units are required to participate in GAIN;
- (2) All units remain in AFDC;

(3) Each individual is assigned to the one, of AFDC or GAIN, which is most likely to give the individual positive earnings, where the decision is informed by the model of the previous sections (a policy “mandated” to increase employment); and

(4) Each unit is allowed to choose which of AFDC or GAIN she participates in, but is required to remain in the program once she has chosen. Agents are assumed to use the models of Sections 4 and 5 in making their choice. I consider three variants: individuals are risk neutral; they have constant relative risk aversion (CRRA) preferences, with coefficient of relative risk aversion, q , equal to 1; and CRRA preferences with $q=3$ (see Manski [1995]).²⁷

7.2 Social Welfare Functions

Suppose that the policymaker is concerned with a group of individuals, $i=1, \dots, P$, identified by their pre-treatment covariates (including earnings and employment histories), so that this group can be represented as a distribution over a space spanned by the set of pre-treatment covariates. In my analysis, I confine the policymaker to two types of concerns: outcomes such as earnings and welfare costs to the government; and inequality in the distribution of earnings. I address each in turn.

Most of the attention in the program evaluation literature has been focused on the outcome(s) of assignment to treatment. A useful starting point is the case in which there is no uncertainty, so that earnings (and other outcomes) for each individual take

²⁷ Caballero (1991) suggests that a reasonable value for the coefficient of relative risk aversion is between 2 and 3. He cites Hall (1988) and Friend and Blume (1975).

on some given value, $(Y_{i T_t}^t)^0$. A typical set of social concerns is captured in the following social welfare functions (SWFs):

(1) Average (labor) earnings per individual,

$$\frac{1}{\tau P} \sum_{i=1}^P \sum_{t=1}^{\tau} (Y_{i T_t}^t)_1^0,$$

where $(\cdot)_j$ refers to the j -th element in the outcome vector, earnings;

(2) The size-of-the-pie,

$$\frac{1}{\tau P} \sum_{i=1}^P \sum_{t=1}^{\tau} \sum_{m=1}^M (Y_{i T_t}^t)_m^0,$$

where the concern is the average (undiscounted) value of expected income (including labor earnings, AFDC, and food stamps, hence $m=1,2,3$) across individuals and time;

(3) The get-them-to-work criterion,

$$\frac{1}{\tau P} \sum_{i=1}^P \sum_{t=1}^{\tau} \mathbb{1}((Y_{i T_t}^t)_1^0 > 0),$$

considers the fraction of individuals employed (having positive income), averaged across time periods;

(4) The fiscal criterion considers the average expenditure per period per person entailed by a given program,

$$\frac{1}{\tau P} \left(\sum_{i=1}^P \sum_{t=1}^{\tau} \sum_{m=2}^M (Y_{i T_t}^t)_m^0 + \left(\sum_{i=1}^I T_i \right) C \right),$$

where $m=2,3$ are AFDC and Food Stamp payments respectively, and C reflects the per unit costs of GAIN administration (in excess of AFDC administration costs);

(5) The net gain criterion considers total average earnings per individual minus total cost in excess of standard AFDC services,

$$\sum_{i=1}^P \sum_{t=1}^T (Y_{i \tau_t}^t)^0 - \left(\sum_{i=1}^I T_i \right) C$$

(the idea being that when comparing two vectors of treatment assignments $\{T_i\}$ and $\{T_i'\}$, the difference in the net gain criterion can be interpreted as the increase in earnings minus the increase in cost); and

(6) The fiscal criterion with taxes is a variant of (4) in which additional earnings realized through participation in training are taxed at 15 percent.

In practice, there will be two modifications. First, I take the population of interest to be the sample which we observe for Alameda county ($P=I$); an alternative interpretation is that I use the empirical distribution of pre-treatment covariates, Z_i , as if it were the population distribution. Second, $(Y_{i \tau_t}^t)$ is, of course, uncertain. The social welfare functions introduced above could be computed for any of the values that $(Y_{i \tau_t}^t)$ might possibly take on. Thus, I compute the posterior distribution of these social welfare functions. Rather than using a single value for $(Y_{i \tau_t}^t)$, this amounts to using the predictive distribution of earnings for each individual. Computationally, I approximate this by simulating a reasonable number of draws from the predictive distribution (100 to 1000 depending on the context), as described in Section 5.3. Thus, the policymaker ultimately considers the (posterior) distribution of possible values of each of the SWFs.

A second type of concern which motivates the policymaker is with inequality. In Section 8, I will first take up the case in which the policymaker is concerned with *ex post* inequality in earnings, by examining the posterior distributions of a number of measures of income inequality for each of the policies under consideration. In this section, I set up the framework for the case in which the policymaker is concerned with *ex ante* inequality. Each individual faces a distribution of possible earnings. The policymaker collapses each distribution into a single number: $E(u(Y_i^t \tau_i))$, the individual's expected utility (or alternatively the certainty equivalent), where the expectation is over the predictive distribution of earnings.²⁸ Her concern for inequality is with the distribution of these expected utilities or certainty equivalents across individuals. In this formulation, the distribution each individual faces is assessed through the individual's preferences. I consider four examples.

Three of these SWFs are of the form:

$$\sum_i g(E(u(Y_i^t \tau_i))),$$

where $g(\cdot)$ is specified according to the given social welfare function. I consider: utilitarian with $g(x)=x$, log with $g(x)=\log(x)$, and an intermediate case with $g(x) = (x^{1-\varepsilon} / 1 - \varepsilon)$, where $\varepsilon=3$ (see Deaton and Muellbauer [1980]). I also consider a Rawlsian social welfare function of the form:

$$\min_i \{E(u(Y_i^t \tau_i))\},$$

²⁸ It is easier for me to deal with certainty equivalents because, unlike the expected utilities, certainty equivalents are denominated in dollars, and hence are comparable even when underlying individual utilities vary.

where the same individual utilities discussed in Section 7.1 are adopted.

8. What We Learn from GAIN

I implement the social welfare analysis outlined in the previous section on the sample from Alameda county; i.e., I consider the welfare implications of assigning the 1360 individuals in the Alameda sample to treatment in various ways. As in Section 6, the key input to the analysis is the (posterior) predictive distribution of the outcomes (earnings per individual per quarter under both treatment and control).

Tables 10a, 10b, and 10c apply the social welfare analysis outlined in Section 7 to the predictive distributions of outcomes under treatment and control. I simulate the predictive distribution of the earnings under treatment and control for 13 quarters of post-treatment earnings for each of the 1360 individuals in the sample, and use deterministic rules to impute values for AFDC receipts and food stamps. Individuals are assigned to treatment and control as outlined in Section 7.1, where (when they are allowed to choose) their decisions are based on the predictive distributions. Even though my modeling attention is concentrated on earnings rather than AFDC and food stamps, the conclusions presented with respect to AFDC receipts and food stamps are robust to the method of imputation.

Table 10a uses SWFs (1) to (6) to assess the policies outlined in Section 7.1. Each SWF-policy combination defines a cell, within which the social welfare function is computed for the entire predictive distribution of earnings. Table 10a presents the mean and the 2.5 and 97.5 percentiles of this distribution for each cell. Consider first

SWF(1), average post-treatment earnings per person per quarter. From the first two cells of column 1 note that the mean prediction from the posterior of the model is similar to that obtained from the empirical distribution, within \$80 for GAIN and within \$30 for AFDC. The 95 percent posterior confidence intervals of expected quarterly earnings under treatment and control just overlap. From cells 3 and 4 we can see that the policies of mandating individuals to enroll in either GAIN or AFDC based on their probability of post-treatment employment, or simply allowing them to choose (if they are risk neutral), both yield substantially higher average quarterly earnings than the policy of enrolling everyone in GAIN, though at the individual level the impact of the policies is quite different. The mandated policy assigns 300 more individuals into treatment than the (risk neutral) voluntary policy. This difference arises because for some individuals, whereas GAIN does increase the probability of employment, it does not increase average earnings; it steers some individuals toward lower paying jobs than they might have found without having their job search or employment disrupted by training. When individuals exhibit a low degree of risk aversion ($q \approx 1$), the pattern of their choices is similar to the risk neutral case. With $q=3$, more individuals prefer GAIN to AFDC. In both cases the average value of expected earnings is similar to the risk neutral case.²⁹

It is not surprising that the voluntary mechanism yields higher earnings than GAIN (or AFDC). All three voluntary mechanisms are able to steer away many (about

²⁹ The findings are much sharper than would be obtained using Manski's (1995) extreme bounds analysis of what he calls the "mixture problem". Of course, the sharper findings come at a price: the willingness to specify a likelihood model. But having paid the price the advantage is a full posterior distribution for the outcomes of interest, allowing for a richer analysis of individual decisions.

700) individuals who are not expected to benefit from GAIN. These results match up with those of Table 9, where it was first observed that a large number of individuals do not benefit from GAIN. The difference in earnings between the voluntary policies and the non-voluntary ones, \$50 to \$200 per person per quarter, provides one measure of the value to participants of being permitted to choose the program in which they participate. The difference is very large relative to average quarterly earnings (which are on the order of \$300).

The advantage of working with the entire distribution of social welfare values is seen in Figure 8, which is analogous to Figures 6 and 7. Based on comparing means and their 95 per cent confidence intervals one might conclude that no decisive statement could be made about the differences among policies: even though the means of the policies are different, their 95 per cent confidence intervals overlap to varying degrees. But when we look at the entire distribution in Figure 8, we note that GAIN first-order stochastically dominates AFDC, and in turn is dominated by the voluntary and mandated policies. Thus, in comparing the entire distribution of values a very strong ranking emerges based on SWF(1).

In column 2, we see that adding AFDC receipts and food stamps to the analysis does not alter the ranking. GAIN produces "a much larger pie" than AFDC, with the mandated and voluntary mechanisms doing slightly better than GAIN. When individuals are allowed to choose, we see a similar pattern to column 1.

Social welfare function (3) ranks the policy alternatives by the criterion of putting people to work. GAIN (in keeping with its stated mandate) does succeed in putting

more people to work than AFDC, although the magnitude of the difference is not large (the finding in Table 5). The mandated policy succeeds in putting more people to work than GAIN and slightly more than the voluntary (1) policy. Again, when individuals are risk averse, we see a pattern similar to columns 1 and 2.

Social welfare function (4) (and SWF(6) which is similar) considers the fiscal criterion, and reveals that GAIN is clearly the most expensive of the policies considered and AFDC the cheapest; the posterior confidence intervals for these two policies do not overlap. To some extent this is not surprising, because GAIN differs from AFDC precisely in offering costly education and training, services that AFDC recipients would have to pay for at their own expense.³⁰ When individuals are allowed to choose, the pattern is similar to column 1. SWF(5) reveals that the increased earnings enjoyed by individuals participating in GAIN do not offset the increased costs when compared to AFDC. Figure 9 illustrates the fact that the ranking among the first four policies that emerges through SWF(5) in fact amounts to first-order stochastic dominance.

The implication of the preceding discussion is that, to the extent we believe that constant relative risk aversion preferences are reasonable, combining SWF (1), (2) and (3) with SWF (4), (5) and (6) suggests that policies allowing individuals to choose between the two programs dominate the policy in which they are assigned to GAIN: such policies are both cheaper and result in higher average earnings per person. But voluntary mechanisms do not dominate AFDC. This finding is not trivial. It is obvious

³⁰ Of course, this ignores the issue of non-GAIN sources of funding.

(indeed, vacuously true) that (rational) individuals always will prefer to choose between alternatives rather than being forced to choose. But Table 10a establishes that under the case of CRRA utility, the (utility maximizing) interests of the participants are aligned with those of a policymaker who values both higher average expected earnings and a lower fiscal burden, at least in preferring a voluntary mechanism to GAIN (AFDC is not dominated).

In summary, Table 10a yields the following results. First, even though GAIN does not yield significantly higher mean earnings than AFDC, looking at the entire distribution we discover that GAIN dominates AFDC, and in turn is dominated by the voluntary mechanism. Second, the policy in which individuals are assigned into treatment and control based on expected utility comparisons dominates GAIN. Third, AFDC is the cheapest of the policies considered in terms of total welfare expenditures, though in terms of increased earnings net-of-costs the voluntary policies are cost effective, whereas GAIN is not, when compared to AFDC. Thus the policymaker is left to decide whether the increase in earnings realized through a voluntary mechanism is worth the additional cost compared to AFDC.

Tables 10b and 10c explore the issue of inequality in the impact of the various policies. I begin in Table 10b by exploring the differences in the degree of *ex post* inequality generated by the policies under consideration. Then in Table 10c, I examine the conclusions reached by using four particular social welfare functions. Table 10b examines inequality in the *ex post* earnings distribution by considering percentiles of the earnings distribution and the difference between the 90th and 10th percentiles.

These are based on a distribution which averages individual predictive distributions across the 13 post-treatment periods; these averaged individual predictive distributions are then used to generate the posterior distributions of the various percentiles. The 5th and 25th percentiles of the distribution for each of the five policies is \$0. The differences among the policies emerge in the upper percentiles. The 75th percentile is lower for AFDC compared to the voluntary other policies, but is higher than GAIN. This underscores the point which arose in earlier discussion, namely that the benefits from GAIN are far from uniform. For many individuals, the expected earnings they achieve through GAIN are lower than they would achieve through AFDC. The difference between GAIN and AFDC arises in the upper tail of the distribution, for example at the 95th percentile; this is depicted graphically in Figure 10. Looking at the difference between the 90th and 10th percentiles, we note that in general voluntary policies produce more inequality than either GAIN or AFDC.

Table 10c addresses the question of how a policymaker would combine an aversion to inequality with a concern for the utility of those individuals who benefit from GAIN and the voluntary policies. The utilitarian, Rawlsian, log, and intermediate social welfare functions discussed in Section 7.2 are employed. Because many of the policy variations are generated by varying individuals' utility functions, I apply the social welfare functions to the certainty equivalents of the distributions of earnings for each individual. Thus, the social welfare rankings do not have any standard errors or confidence intervals; uncertainty is already accounted for in the certainty equivalents. Table 10c reveals that if the policy maker is sufficiently inequality averse (as embodied

in SWF (8) to (10)), then she will prefer AFDC to GAIN. But, under both the inequality-neutral and inequality-averse specification the voluntary policies are preferred to both GAIN and AFDC. This finding is not trivial. Though each individual must have weakly higher utility under a voluntary mechanism as compared to being assigned to GAIN or AFDC, depending on the policymaker's aversion to inequality -- similar to risk aversion -- the social ranking may not mirror the individual ranking.

In summary, from Table 10a we learn that, looking at outcomes such as earnings and welfare expenditures, the voluntary mechanism dominates GAIN, though not AFDC. In Table 10b it was observed that the benefits from GAIN are enjoyed at the upper end of the earnings distribution, and that a voluntary mechanism produces more inequality in the distribution of earnings than either AFDC or GAIN. Finally, from Table 10c, we learn that even a policymaker who is inequality-averse would prefer the distribution under a voluntary mechanism to *both* GAIN and AFDC.

9. Conclusion

In this paper I have examined the implications of shifting the emphasis in program evaluation from examining treatment effects and their statistical significance to looking at the underlying decision problems. Two main differences emerge.

First, by looking at the entire distribution of earnings under treatment and control for a wide range of individuals, I conclude that for most individuals the choice between GAIN and AFDC is clear-cut, with over half the sample preferring GAIN to

AFDC. The group benefiting from GAIN was identified as those individuals with higher pre-treatment earnings.

Second, the policy of allowing individuals to choose which program they enter was discovered to dominate GAIN; it yields higher average earnings and leads to lower government expenditure.

The applicability of this framework extends beyond the case of randomized experiments considered here. In this regard the key observation is that in non-experimental settings, if a sufficiently rich set of covariates is observed (so that the assumption of selection on observable covariates is reasonable) and one is able to condition on them in a flexible way, the same methodology could be adopted (see Rubin [1977, 1978]). In fact, in the Bayesian framework that I employ, a hierarchical prior can readily be fitted to the parameters, allowing a large number of interactions and coefficients to be estimated under the assumption that they are draws from distributions that depend on a much smaller set of parameters.

The model adopted thus far can be extended in a number of respects. First, there is scope to add greater heterogeneity, perhaps by using a hierarchical model to incorporate many more interactions. Second, the model could be modified to forecast beyond the 13 quarters included in the dataset to extend the evaluation to longer horizons. These are subjects for future research.

References

- Albert, J. and S. Chib (1993). "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669-679.
- Ashenfelter, O., and David Card (1985). "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," *Review of Economics and Statistics*, 67, 648-660.
- Barberis, Nicholas (1996). "Investing for the Long Run when Returns are Predictable," Harvard University, unpublished.
- Burtless, Gary (1995). "The Case for Randomized Field Trials in Economic and Policy Research," *Journal of Economic Perspectives*, 9, 63-84.
- Caballero, Ricardo (1991). "Earnings Uncertainty and Aggregate Wealth Accumulation," *American Economic Review*, 81, 859-871.
- Chamberlain, Gary, and Guido Imbens (1996). "Hierarchical Bayes Models with Many Instrumental Variables," Harvard Institute of Economic Research, Paper Number 1781.
- Chib, Siddhartha (1992). "Bayes Inference in the Tobit Censored Regression Model," *Journal of Econometrics*, 51, 79-99.
- and Edward Greenberg (1994). "Markov Chain Monte Carlo Simulation Methods in Econometrics," Washington University, manuscript.
- Clements, Nancy, Jeffrey Smith, and James Heckman (1994). "Making the Most Out of Social Experiments: Reducing the Intrinsic Uncertainty in Evidence From Randomized Trials With an Application to the National JTPA Experiment," National Bureau of Economic Research Technical Paper 149.
- Deaton, Angus, and John Muellbauer (1980). *Economics and consumer behavior*. Cambridge: Cambridge University Press.
- Dehejia, Rajeev, and Sadek Wahba (1996). "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs," Harvard University, unpublished.
- Friend, Benjamin, and Marshal Blume (1975). "The Demand for Risky Assets," *American Economic Review*, 65, 900-22.
- Fisher, R. (1935). *The Design of Experiments*. London: Oliver and Boyd.

Gelfand, A.E., and A.F.M. Smith (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398-409.

Gelman, Andrew, John Carlin, Hal Stern, and Donald Rubin (1996). *Bayesian Data Analysis*. London: Chapman and Hall.

Geman, S., and D. Geman (1984), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.

Geweke, John, and Michael Keane (1996). "An Empirical Analysis of the Male Income Dynamics in the PSID: 1968-1989," University of Minnesota, unpublished.

Greenberg, David, and Michael Wiseman (1992). "What Did the OBRA Demonstrations Do," in Manski and Garfinkel (1992).

Hall, Robert (1988). "Intertemporal Substitution," *Journal of Political Economy*, 96, 339-57.

Heckman, James (1989). "Choosing among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training," *Journal of the American Statistical Association*, 84, 862-74.

----- (1990). "Varieties of Selection Bias," *American Economic Review*, 80, 313-318.

----- (1992). "Evaluating welfare and training programs," in Manski and Garfinkel (1992).

----- and Jeffrey Smith (1995). "Assessing the Case for Social Experiments," *Journal of Economic Perspectives*, 9, 85-110.

----- and Richard Robb (1985). "Alternative Methods for Evaluating the Impact of Interventions," in James Heckman and Burton Singer (eds.), *Longitudinal Analysis of Labor Market Data*. Econometric Society Monograph No. 10. Cambridge: Cambridge University Press.

----- and ----- (1986). "Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatment on Outcomes," in Howard Rainer (ed.), *Drawing Inferences from Self-Selected Samples*. New York: Springer-Verlag.

Kandel, Shmuel, and Robert Stambaugh (1996). "On the Predictability of Stock Returns: An Asset-Allocation Perspective," *Journal of Finance*, LI, 385-424.

- Lalonde, Robert (1986). "Evaluating the Econometric Evaluations of Training Programs," *American Economic Review*, 76, 604-620.
- Manski, Charles (1989). "Anatomy of the Selection Problem," *Journal of Human Resources*, 24, 343-360.
- (1993). "The Selection Problem," in C. Sims (ed.), *Advances in Econometrics*. Cambridge: Cambridge University Press.
- (1995). "The Mixing Problem in Program Evaluation," *Identification Problems in the Social Sciences*, Chapter 3. Cambridge: Harvard University Press.
- (1996). "Learning about Treatment Effects from Experiments with Random Assignment of Treatments," *Journal of Human Resources*, 31, 709-733.
- and Irwin Garfinkel (1992). *Evaluating Welfare and Training Programs*. Cambridge: Harvard University Press.
- Neyman, J. (with K. Iwazkiewicz and S. Kolodziejczyk) (1935), "Statistical Problems in Agricultural Experimentation" (with discussion), *Supplement of Journal of the Royal Statistical Society*, 2, 107-180.
- Robert, Christian (1996). "Mixtures of Distributions: Inference and Estimation," in W.R. Gilks, S. Richardson, and D.J. Spiegelhalter (eds.), *Markov Chain Monte Carlo in Practice*, London: Chapman and Hall.
- Rubin, Donald (1977). "Assignment to a Treatment Group on the Basis of a Covariate," *Journal of Educational Statistics*, 2, 1-26.
- (1978). "Bayesian Inference for Causal Effects: The Role of Randomization," *The Annals of Statistics*, 6, 34-58.
- Riccio, James, Daniel Friedlander, and Stephen Freedman (1994). *GAIN: Benefits, Costs, and Three-Year Impacts of a Welfare-to-Work Program*. New York: Manpower Demonstration Research Corporation.
- Rossi, Peter, Robert McCulloch, and Greg Allenby (1995). "Hierarchical Modelling of Consumer Heterogeneity: An Application to Target Marketing," in C. Gatsonis, J. Hodges, R. Kass, and N. Singpurwalla (eds.), *Case Studies in Bayesian Statistics*, Volume II, *Lecture Notes in Statistics*, 105. New York: Springer-Verlag.
- Tanner, M., and W. Wong (1987). "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82, 528-550.

Data Legend

Variable	Description
CHILD4	Number of children less than age 4
CHILD45	Number of children between ages 4 and 5
CHILD611	Number of children between ages 6 and 11
CHILD18	Number of children between ages 12 and 18
CHILD19	Number of children ages 19 and greater
CAREAD	Score on reading test
CAMATH	Score on mathematics test
GRADE	Educational attainment (grade 0 to 20)
HRWAGE	Most recently recorded hourly wage
FAM.TYPE	Indicator for households with single head
AGE	Age
EXPER.	Indicator for experimental unit
CONTROL	Indicator for control unit
SEXFEMALE	Indicator for female participants
REFUGEE	Indicator of refugee status
CUR.AFDC	Indicator for receiving AFDC in pre-experimental time
PREVTR	Indicator for previous training or job search activities
ETH. WHITE	Ethnicity Indicator, White
ETH. HISP.	Ethnicity Indicator, Hispanic
ETH.BLACK	Ethnicity Indicator, Black
ETH.NATIVE	Ethnicity Indicator, Native Indian
ETH.IND.CH	Ethnicity Indicator, Indo-Chinese
ETH.OTH.AS	Ethnicity Indicator, Other Asian
ETH.PACF.	Ethnicity Indicator, Pacific Islander
ETH.FILIP.	Ethnicity Indicator, Filipino
ETHOTHR	Ethnicity Indicator, Other
AVG.UNEMP	Average county employment rate, at registration
PEARN x	Earnings in quarter x of pre-experimental time
PEARN x Z	Indicator of zero earnings, pre-experimental quarter x
EARN x	Earnings in quarter x of post-experimental time
EARN x Z	Indicator of zero earnings, post-experimental quarter x
PAFDC x	AFDC receipts, pre-experimental quarter x
AFDC x	AFDC receipts, post-experimental quarter x
PFDSTMP x	Food Stamps receipts, pre-experimental quarter x
FDSTMP x	Food Stamps receipts, post-experimental quarter x
EMPPQ x	Employment status, pre-experimental quarter x
EMPQ x	Employment status, post-experimental quarter x

Table 1: Comparing GAIN and AFDC, Average of Outcomes Per Person^a

Policy	Labor earnings per quarter	Total income per quarter	Probability of employment	Total fiscal expenditures	Earnings net-of-costs per quarter ^b	Total fiscal expenditures less tax receipts
GAIN	463	2,425	0.2042	1,883	202	1,181
AFDC	372	2,359	0.1843	1,636	387	1,630
standard error	56	67	0.018	85	-48	-156

Notes:

(a) Means are computed from the empirical distribution.

(b) Costs are normalized to zero for AFDC, and are an additional \$3638 for 13 quarters of GAIN.

Table 2: The Sample

	Alameda	Butte	Los Angeles	Riverside	San Diego	Tulare
GAIN:						
Treated Group	685	1717	3730	5808	8711	2693
Control Group	682	458	2124	1706	1810	1146
Total	1367	2175	5854	7514	10521	3839
AFDC:						
Total	30305	5663	231356	24000	50958	14673

Notes: The GAIN sample sizes are from the public use file of the GAIN data. The AFDC total represents the number of AFDC cases (both single-parent and two-parent households) in the six evaluation counties in December 1990 (see Riccio, *et al.*, (1994), Table 1.1). GED is the General Educational Development certificate.

Table 3: Data Description, Alameda County

Variable	Mean	Standard deviation
CHILD4	0.1931	0.49
CHILD45	0.2312	0.46
CHILD611	1.1639	4.68
CHILD18	0.8830	2.29
CHILD19	0.2480	0.60
CAREAD	206.27	98.00
CAMATH	192.44	94.96
GRADE	10.79	3.02
HRWAGE	3.74	2.73
FAM.TYPE	0.6218	
AGE	35.39	8.85
EXPER.	0.5011	
SEXF	0.8574	
REFUGEE	0.0914	
CUR.AFDC.	0.9898	
PREVTR	0.2407	
ETH.WHITE	0.1792	
ETH.HISP.	0.0775	
PEARN10	165.02	740.14
PEARN9	153.17	675.96
PEARN8	154.53	747.70
PEARN7	187.67	1036.91
PEARN6	156.83	615.03
PEARN5	170.37	771.74
PEARN4	185.30	726.89
PEARN3	151.60	685.37
PEARN2	153.64	642.86
PEARN1	167.17	714.04
PEARNZ10	0.87	
PEARNZ9	0.88	
PEARNZ8	0.87	
PEARNZ7	0.87	
PEARNZ6	0.87	
PEARNZ5	0.87	
PEARNZ4	0.87	
PEARNZ3	0.87	
PEARNZ2	0.87	
PEARNZ1	0.87	

Table 4: Proportion of Units With Zero Earnings

Period	Treated, Fraction with 0 Earnings			Control, Fraction with 0 Earnings		
	Total	With pos. earnings last period	With 0 earnings last period	Total	With pos. earnings last period	With 0 earnings last period
1	0.87	0.04	0.82	0.83	0.03	0.80
2	0.84	0.03	0.80	0.83	0.04	0.79
3	0.82	0.04	0.78	0.82	0.04	0.78
4	0.80	0.03	0.76	0.82	0.04	0.78
5	0.80	0.04	0.76	0.83	0.04	0.78
6	0.81	0.06	0.75	0.82	0.04	0.78
7	0.80	0.04	0.76	0.81	0.02	0.79
8	0.78	0.03	0.75	0.82	0.04	0.78
9	0.76	0.03	0.73	0.81	0.03	0.78
10	0.76	0.03	0.73	0.80	0.03	0.78
11	0.77	0.04	0.72	0.80	0.02	0.78
12	0.78	0.04	0.74	0.80	0.03	0.77
13	0.76	0.03	0.73	0.82	0.04	0.78

Table 5: Treatment Effect on Probability of Unemployment

Period	Treatment effect	Standard error
1	0.0238	0.0172
2	-0.0114	0.0193
3	-0.0187	0.0203
4	-0.0344	0.0207
5	-0.0444	0.0209
6	-0.0176	0.0210
7	-0.0214	0.0218
8	-0.0480	0.0216
9	-0.0624	0.0225
10	-0.0613	0.0228
11	-0.0444	0.0225
12	-0.0269	0.0221
13	-0.0627	0.0223

Note: A probit is used; covariates include variables for the number of children (CHILD4-CHILD19), reading and writing test scores, grade, age, sex, ethnicity, and earnings histories (PEARN10-PEARN1). The treatment effect is computed as the discrete difference between the probability of unemployment with the treatment indicator set to 0 and 1, where the value of other covariates is set to their sample mean. The delta method is used to compute standard errors.

Table 6: Tobit and OLS Coefficients of Treatment Indicator for Post-experimental Earnings

Post-experimental period:	Tobit treatment effect, no covariates ^a (standard error)	OLS treatment effect, no covariates ^b (standard error)	Tobit treatment effect at mean, with covariates ^c (standard error)	OLS treatment effect, with covariates ^c (standard error)
1	-53.8132 (31.5943)	-55.3965 (38.1249)	-19.4872 (11.7806)	-47.5891 (22.8029)
2	-14.4639 (40.9349)	-21.5642 (49.9266)	7.61671 (24.0324)	-9.22507 (39.1862)
3	15.2279 (45.8743)	27.9581 (53.8577)	28.3159 (30.3611)	35.0674 (45.9238)
4	61.9781 (53.8393)	59.2831 (63.9863)	60.1949 (36.5732)	67.5878 (56.7998)
5	80.4062 (51.3482)	102.034 (59.8254)	85.4666 (36.9396)	111.275 (54.1284)
6	46.1670 (56.4287)	81.7121 (66.2047)	56.2586 (43.5834)	84.9766 (61.1139)
7	48.2229 (60.6675)	76.2227 (70.9794)	61.8553 (48.6967)	84.8176 (66.5290)
8	96.7405 (63.9757)	89.3386 (73.5930)	99.3480 (49.2329)	95.2800 (68.5889)
9	169.134 (70.2234)	188.850 (80.8695)	169.315 (57.4620)	203.061 (76.0224)
10	162.144 (72.7343)	211.317 (84.2982)	180.639 (59.8569)	232.069 (79.4980)
11	131.728 (78.1387)	176.331 (90.0036)	150.025 (65.5091)	194.477 (86.4388)
12	99.6775 (79.7070)	150.746 (92.3857)	106.719 (66.9410)	150.716 (88.8401)
13	206.186 (82.4205)	196.275 (93.8432)	196.552 (71.0454)	206.565 (90.7750)

Notes:

(a) These are the treatment effects for observed income. In a general set-up with covariates:

$$\begin{aligned} E(Y|X) &= E(Y|X, Y > 0) \Pr(Y > 0|X) + E(Y|X, Y = 0) \Pr(Y = 0|X) \\ &= X\beta\Phi(X\beta/\sigma) + \sigma\phi(X\beta/\sigma) \quad (+ 0), \end{aligned}$$

where ϕ and Φ are the standard normal density and c.d.f., X are the regressors in the Tobit, and β their coefficients. If $X = (X_{-T} \ T)$ then the treatment effect is computed as:

$E(Y|\bar{X}_{-T}, T = 1) - E(Y|\bar{X}_{-T}, T = 0)$, where \bar{X}_{-T} is the sample mean of the other covariates. Standard errors are computed by the delta method. In column 1, $X = (1, T)$; otherwise the same method is used.

(b) These are equivalent to a difference in means across treated and control groups for each period.

(c) Covariates are entered linearly.

Table 7a: Characteristics of an Individual ("Ms. Thirteen Fifty-Three")

CHILD4	CHILD45	CHILD611	CHILD18	CHILD19	CAREAD	CAMATH	GRADE	HRWGE	FAM.TYPE
0	0	0	1	0	241	225	15	8	30
AGE	SEXF	REFUGEE	ETH.WHT	ETH.HSP	AVG.UNEM	PEARN10	PEARN9	PEARN8	PEARN7
42	1	0	1	0	4.1	0	0	0	3667
PEARN6	PEARN5	PEARN4	PEARN3	PEARN2	PEARN1	PAFDC7	PAFDC6	PAFDC5	PAFDC4
425	0	0	0	0	0	1605	1605	1551	1443
PAFDC3	PAFDC2	PAFDC1							
1518	1518	1518							

Table 7b: Characteristics of an Individual ("Ms. One")

CHILD4	CHILD45	CHILD611	CHILD18	CHILD19	CAREAD	CAMATH	GRADE	HRWGE	FAM.TYPE
0	0	1	2	0	221	191	9	2.5	30
AGE	SEXF	REFUGEE	ETH.WHT	ETH.HSP	AVG.UNEM	PEARN10	PEARN9	PEARN8	PEARN7
32	1	0	0	1	3.07	0	0	0	0
PEARN6	PEARN5	PEARN4	PEARN3	PEARN2	PEARN1	PAFDC7	PAFDC6	PAFDC5	PAFDC4
0	0	0	0	0	0	2259	2364	2364	2364
PAFDC3	PAFDC2	PAFDC1							
2364	2472	2472							

Table 7c: Characteristics of an Individual ("Ms. Ten")

CHILD4	CHILD45	CHILD611	CHILD18	CHILD19	CAREAD	CAMATH	GRADE	HRWGE	FAM.TYPE
0	0	2	0	0	223	240	10	5.85	30
AGE	SEXF	REFUGEE	ETH.WHT	ETH.HSP	AVG.UNEM	PEARN10	PEARN9	PEARN8	PEARN7
27	1	0	1	0	3.7	0	0	0	0
PEARN6	PEARN5	PEARN4	PEARN3	PEARN2	PEARN1	PAFDC7	PAFDC6	PAFDC5	PAFDC4
0	0	0	0	0	0	1899	1899	1926	1989
PAFDC3	PAFDC2	PAFDC1							
1989	1989	2082							

Table 8a: Mean and Variance of Predicted Earnings, Ms. Thirteen Fifty-Three

Post-treatment earnings period	Treatment			Control		
	Probability of positive earnings	Mean post-treatment earnings	Standard deviation	Probability of positive earnings	Mean post-treatment earnings	Standard deviation
1	0.1249	20.95	95.70	0.0539	37.53	294.35
2	0.2617	35.23	134.50	0.0959	37.36	201.24
3	0.3556	215.00	418.18	0.1229	245.98	899.72
4	0.4695	360.45	587.21	0.1469	342.38	1358.29
5	0.5255	442.95	659.03	0.1518	366.65	1290.63
6	0.5524	535.90	985.31	0.1818	430.44	1258.39
7	0.5924	590.03	796.80	0.2158	596.04	1617.93
8	0.6184	650.47	819.86	0.2128	639.57	1924.10
9	0.6553	301.25	1189.85	0.2208	201.58	697.13
10	0.6713	275.89	566.68	0.2308	277.00	1290.74
11	0.6803	890.10	1025.41	0.2288	893.36	2489.61
12	0.6683	881.85	1048.65	0.2208	764.71	2097.09
13	0.6933	367.76	900.34	0.1828	285.40	1431.58

Table 8b: Predicted Earnings, With and Without Uncertainty, Ms. Ten

Post-treatment earnings period	Ignoring parameter uncertainty				Accounting for parameter uncertainty			
	Treated		Control		Treated		Control	
	Predicted earnings	Standard deviation	Predicted earnings	Standard deviation	Predicted earnings	Standard deviation	Predicted earnings	Standard deviation
1	29	217	51	245	23	220	40	233
2	75.90	356	131	581	77	340	119	502
3	208	851	207	701	179	718	227	772
4	316	1014	292	994	302	1127	285	955
5	349	1124	309	986	260	1316	329	1085
6	429	1328	322	1020	376	1277	384	1141
7	443	1427	399	1534	491	1657	417	1330
8	162	1498	200	1116	581	1715	428	1222
9	638	1710	383	1272	640	1857	463	1385
10	674	1813	533	1537	639	2058	478	1448
11	844	1880	514	1501	679	1882	575	1709
12	641	1870	611	1755	604	1912	589	1716
13	898	2570	628	1860	738	2062	484	1836
Log utility [*]	48.85		48.53		48.72		48.75	
CRRA (3) [*]	4.86607		4.86606		4.86606		4.86607	

Note:

^{*} For expected utility comparisons, \$100 is added to the distribution of treated and control earnings.

Table 9: Groups Benefiting the Most and Least from GAIN, Comparing Mean Covariates

Variable	Those who prefer GAIN ^a	Those who prefer AFDC ^b
Number	571	789
Avg. Earnings	854	252
GAIN:		
Avg. Earnings	531	273
AFDC		
CHILD4	0.16	0.21
CHILD45	0.23	0.24
CHILD611	0.94	0.85
CHILD18	0.56	1.02
CHILD19	0.09	0.36
CAREAD	211.99	201.95
CAMATH	197.33	188.69
GRADE	11.85	10.03
HRWAGE	5.05	3.56
FAM.TYPE	30.47	30.74
AGE	32.43	37.50
EXPER.	0.47	0.53
CONTROL	0.53	0.47
SEXFEMALE	0.90	0.82
REFUGEE	0.06	0.11
CUR.AFDC	1.00	0.98
PREVTR	0.46	0.08
ETH. WHITE	0.19	0.17
ETH. HISP.	0.05	0.10
ETH.BLACK	0.69	0.60
ETH.NATIVE	0.00	0.01
ETH.IND.CH.	0.05	0.09
ETH.OTH.AS	0.01	0.02
ETH.PACF.	0	0.00
ETH.FILIP.	0.01	0.01
ETHOTHR	0.02	0.04
AVG.UNEMP	3.95	4.05
PEARN10	323.02	52.14
PEARN9	281.42	61.71
PEARN8	296.80	52.93
PEARN7	399.16	36.29
PEARN6	274.97	72.71
PEARN5	362.57	32.78
PEARN4	407.32	26.27
PEARN3	333.22	20.75
PEARN2	299.38	47.28
PEARN1	268.82	95.10

Note:

(a) Individuals whose expected utility is higher from total earnings under GAIN than AFDC, where preferences are CRRA, with relative risk aversion equal to 3.

(b) The complement of (a).

Table 10b: Considering *Ex Post* Inequality, Quantiles of the Earning Distribution

Policy	0.05	0.25	0.5	0.75	0.95	(0.90-0.10)
GAIN	0 [0,0]	0 [0,0]	39 [5,74]	491 [392,606]	1961 [1677,2251]	1252 [1097,1478]
AFDC	0 [0,0]	0 [0,0]	55 [0,101]	601 [492,704]	1784 [1573,1993]	1312 [1153,1469]
Mandated	0 [0,0]	0 [0,0]	122 [79,185]	702 [587,820]	2156 [1979,2335]	1508 [1378,1694]
Voluntary (1)	0 [0,0]	0 [0,0]	116 [66,171]	772 [668,892]	2271 [2071,2546]	1616 [1504,1840]
Voluntary (2)	0 [0,0]	0 [0,0]	121 [80,182]	768 [666,880]	2257 [2050,2522]	1607 [1481,1804]
Voluntary (3)	0 [0,0]	0 [0,0]	122 [79,187]	738 [644,860]	2206 [1993,2500]	1559 [1426,1757]

Note: Each cell presents the median of the posterior distribution of the percentile, and in parentheses the 5th and 95th posterior percentiles.

Table 10c: Expected Utility Comparisons, Alameda

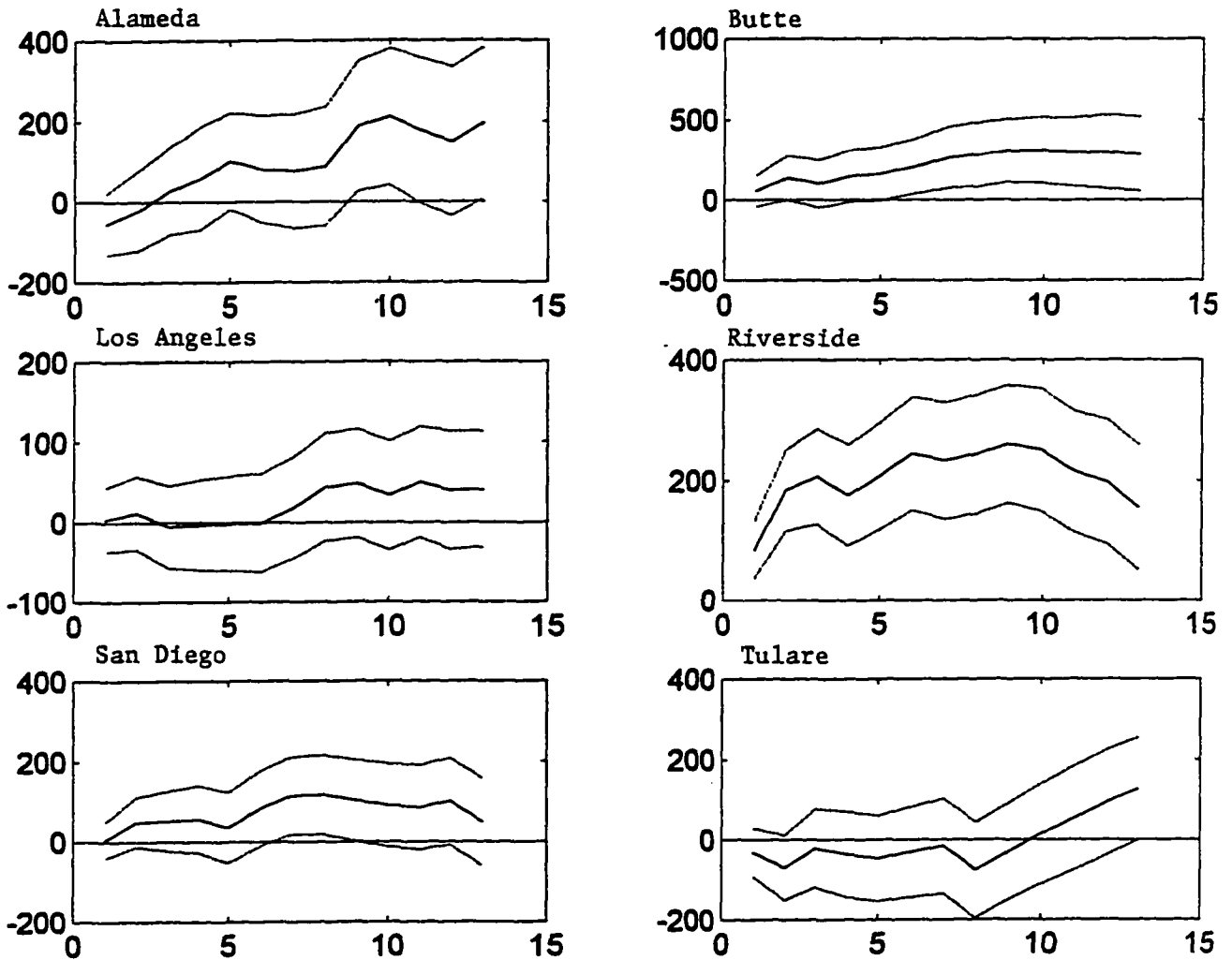
Policy	SWF(7)	SWF(8)	SWF(9)	SWF(10)
Risk Neutral:				
GAIN	1.0483	0.5492	0.9998	-1.0357
AFDC	1	1	1	-1
Mandated	1.0817	0.8205	1.0036	-0.9760
Voluntary (1)	1.0927	1.0114	1.0049	-0.9614
Risk Averse, $q=1$				
GAIN	0.9892	0.2707	0.9949	-1.1240
AFDC	0.9487	0.1972	0.9936	-1.1593
Mandated	1.0098	0.1972	0.9972	-1.1033
Voluntary(2)	1.0144	0.2707	0.9977	-1.0896
Risk Averse, $q=3$				
GAIN	0.9295	0.1994	1.0065	-1.2127
AFDC	0.9057	0.1153	1.0048	-1.3159
Mandated	0.9397	0.1153	1.0075	-1.2384
Voluntary(3)	0.9411	0.1994	1.0078	-1.1992

Notes:

* Expected utilities are normalized.

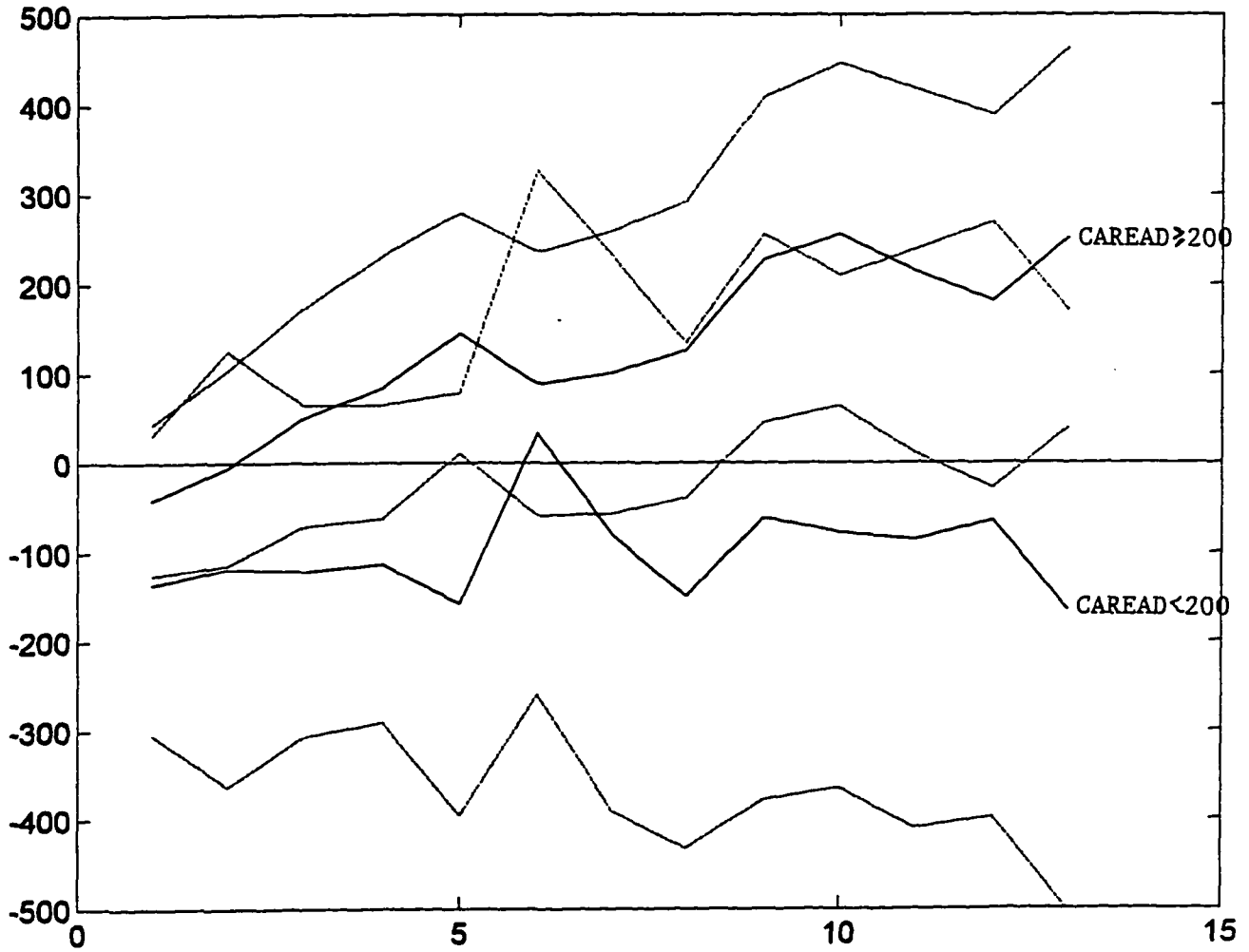
SWF(7): Utilitarian SWF, applied to certainty equivalent of income distribution.SWF(8): Rawlsian SWF, applied to certainty equivalent of income distribution.SWF(9): log SWF, applied to certainty equivalent of income distribution.SWF(10): $SWF = \left(\frac{1}{1-\varepsilon}\right) \left(\sum_i (u_i)^{1-\varepsilon}\right)$, $\varepsilon=3$, applied to certainty equivalent of income distribution.

Figure 1: Treatment Effect Six Counties



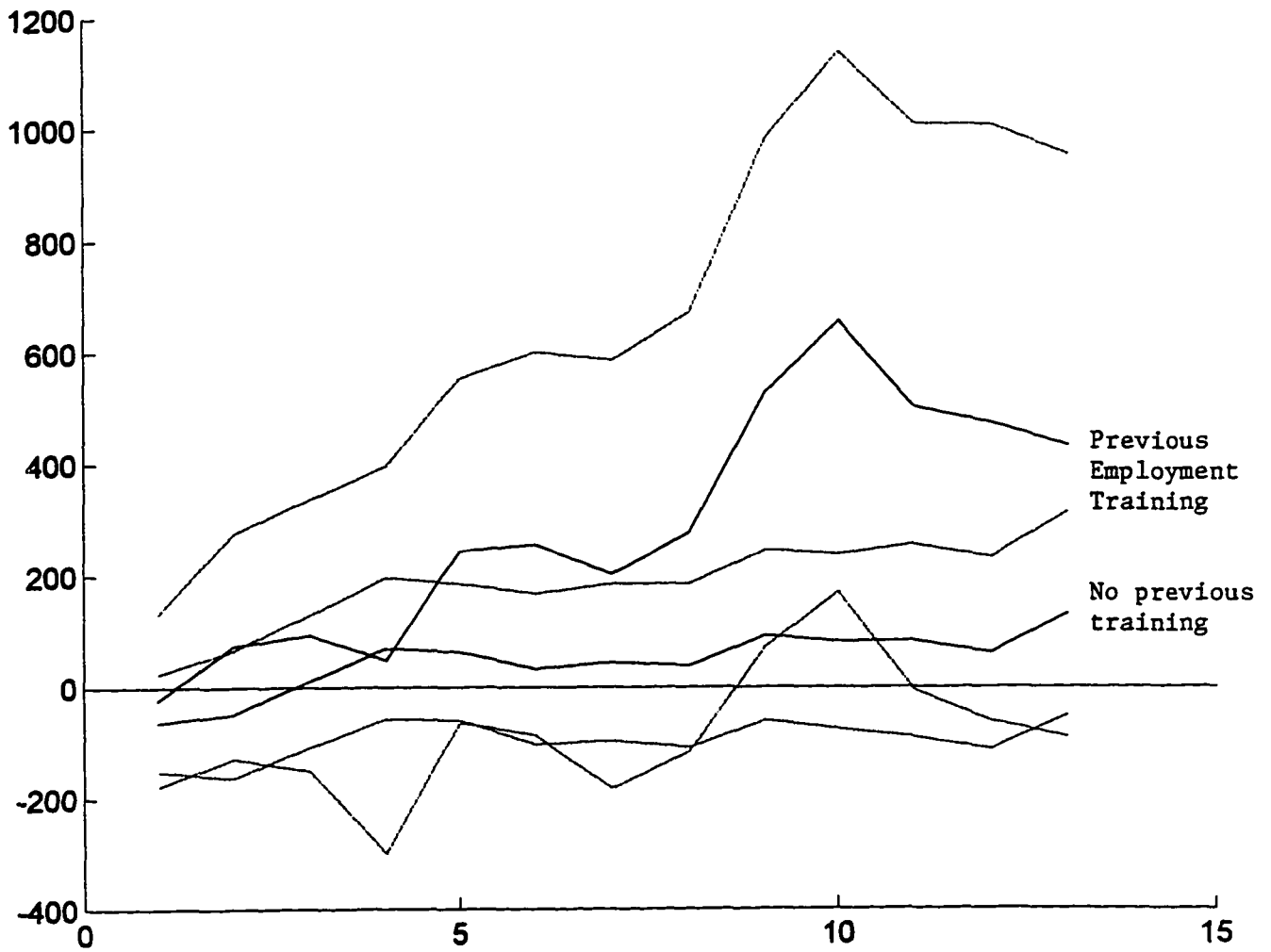
Dashed Lines: \pm two standard errors

Figure 2: Treatment Effects, Alameda



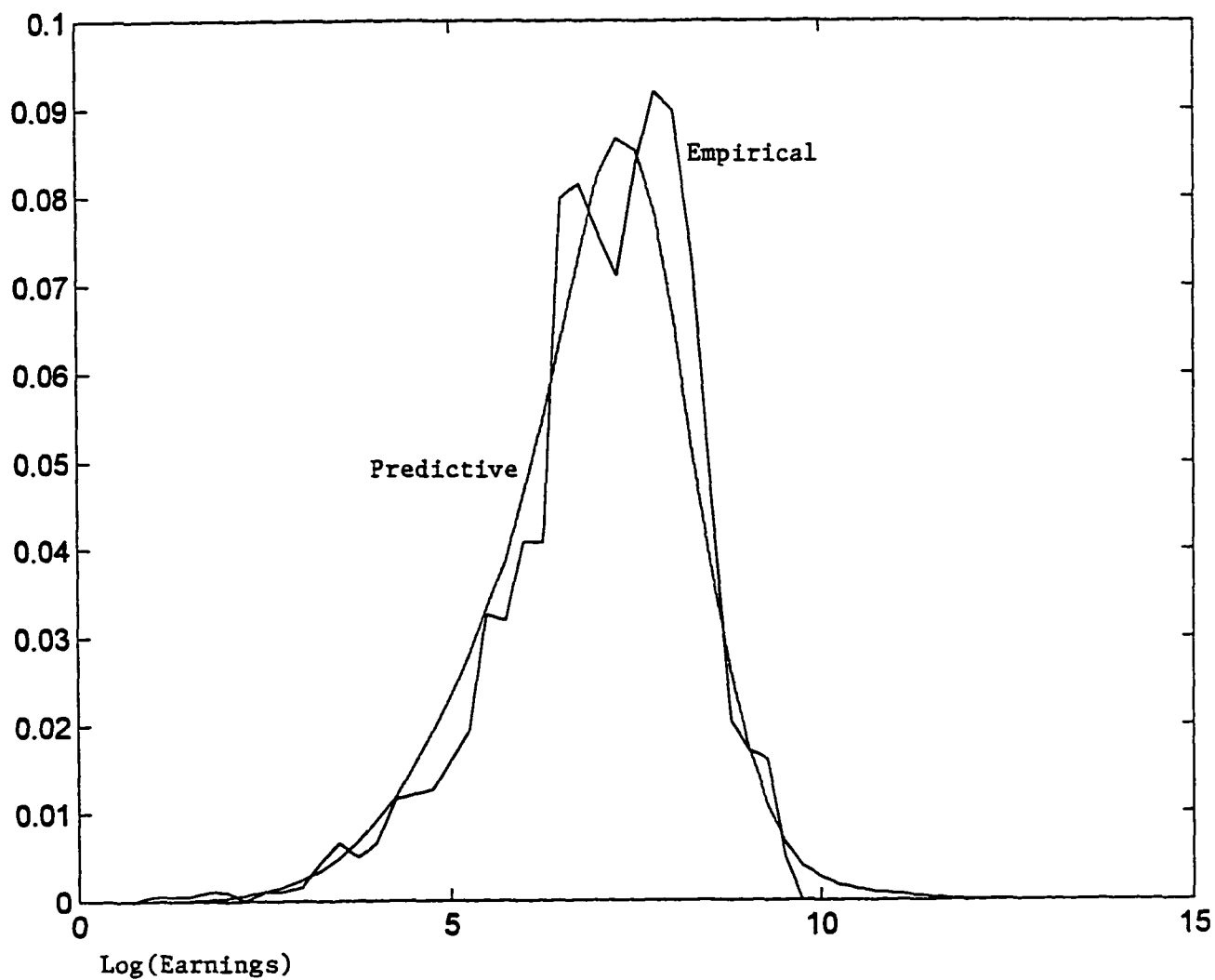
Dashed Lines: +/- two standard errors

Figure 3: Treatment Effects, Alameda



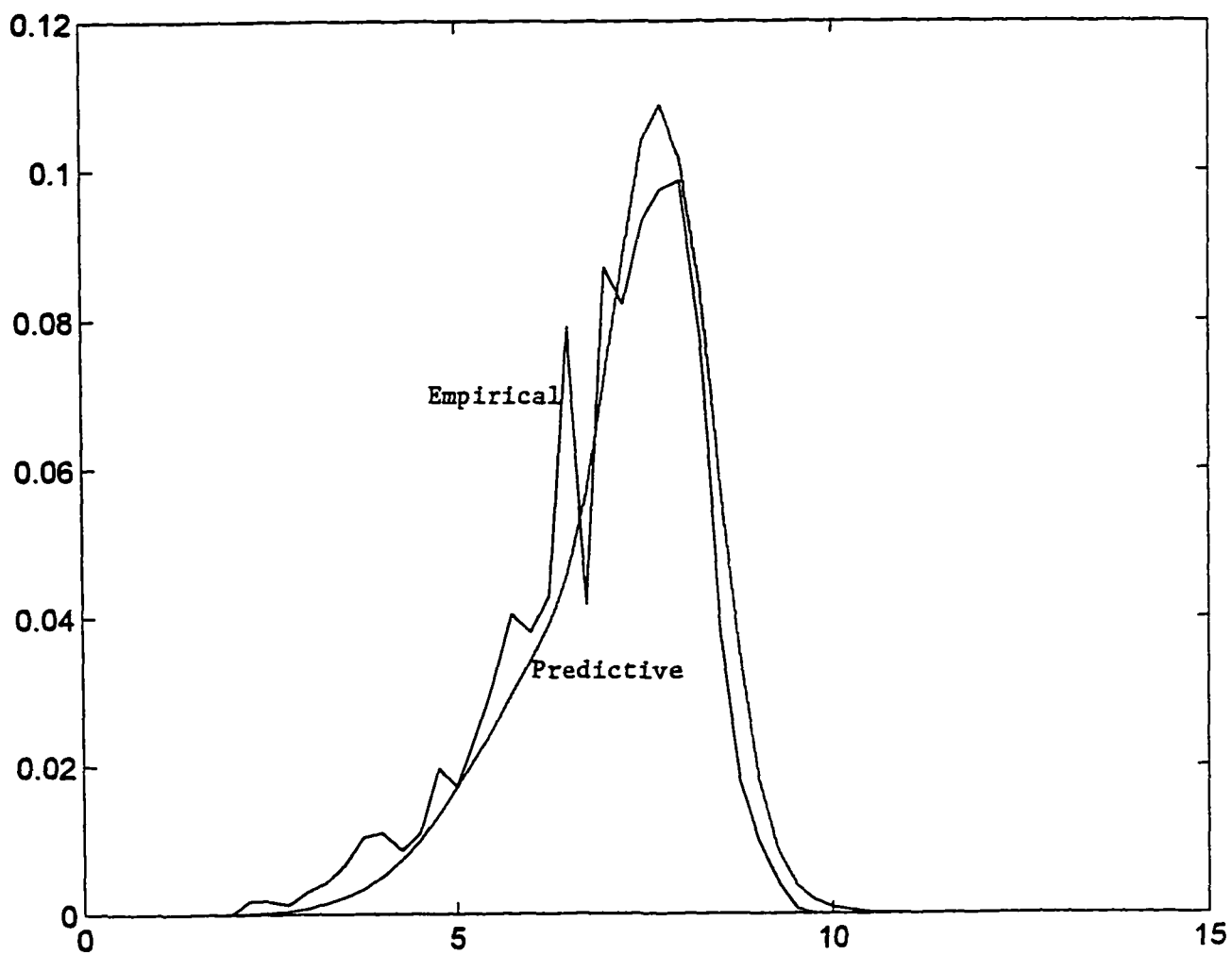
Dashed Lines: +/- two standard errors

Figure 4: Average Post-Treatment Log Earnings
Per Person Per Quarter for Positive Earnings
Empirical and Predictive Densities, Treatment Group



Fraction of Zero Earnings: Empirical 0.80, Predictive 0.79

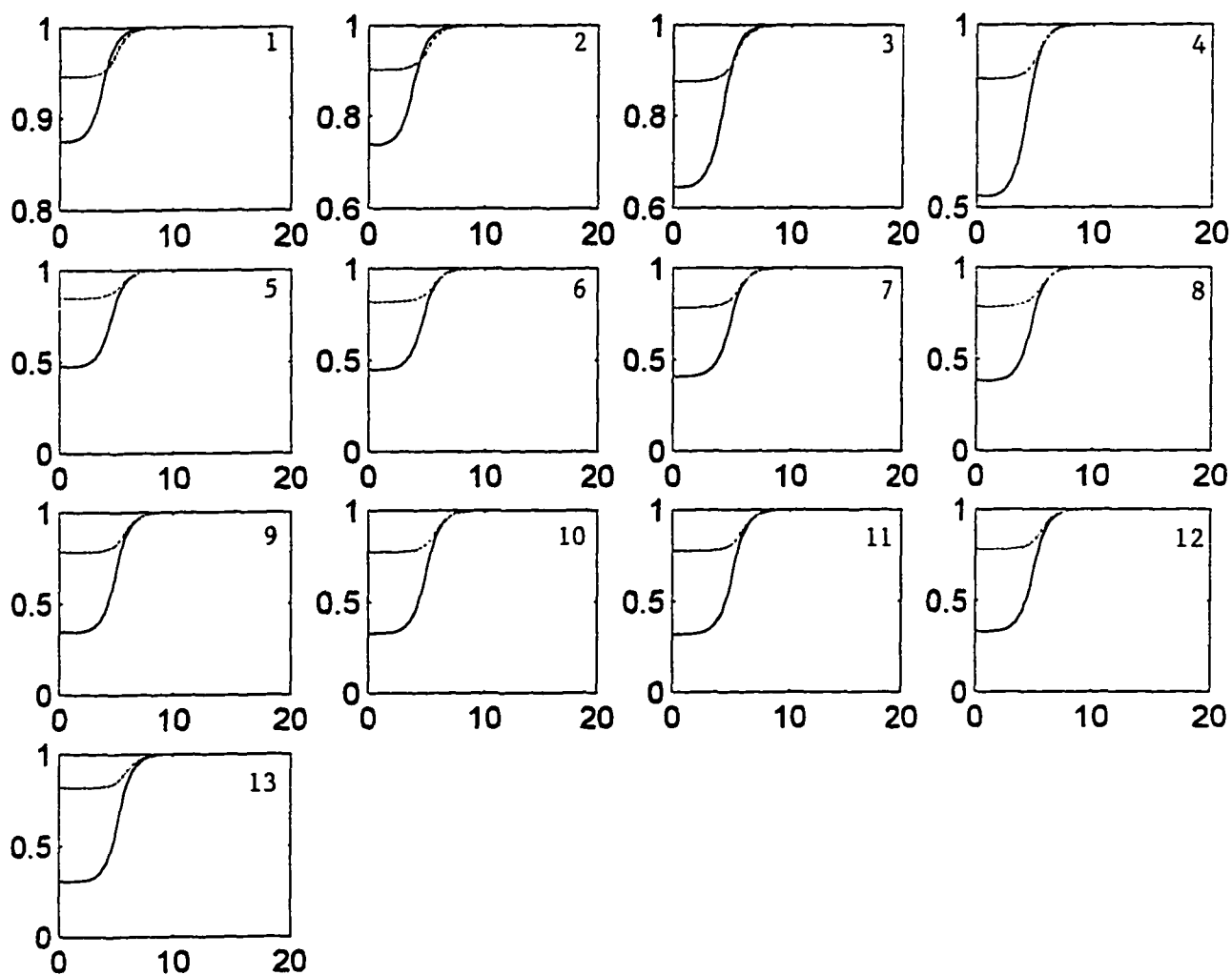
Figure 5: Average Post-Treatment Log Earnings
Per Person Per Quarter for Positive Earnings
Empirical and Predictive Densities, Control Group



Log(Earnings)

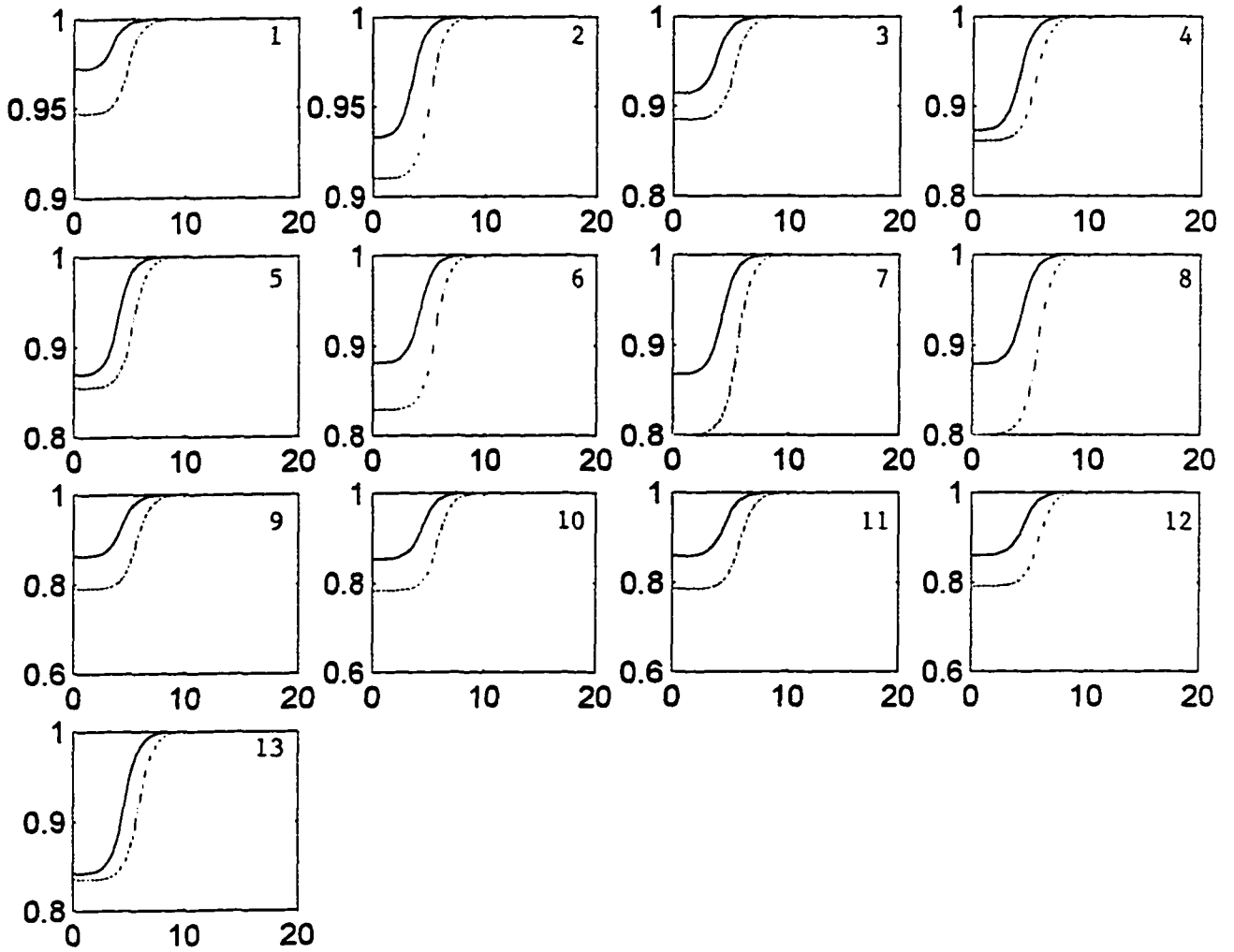
Fraction of Zero Earnings: Empirical 0.82, Predictive 0.82

Figure 6: Cumulative Distribution Functions, Post-Treatment Earnings (Unit 1353)



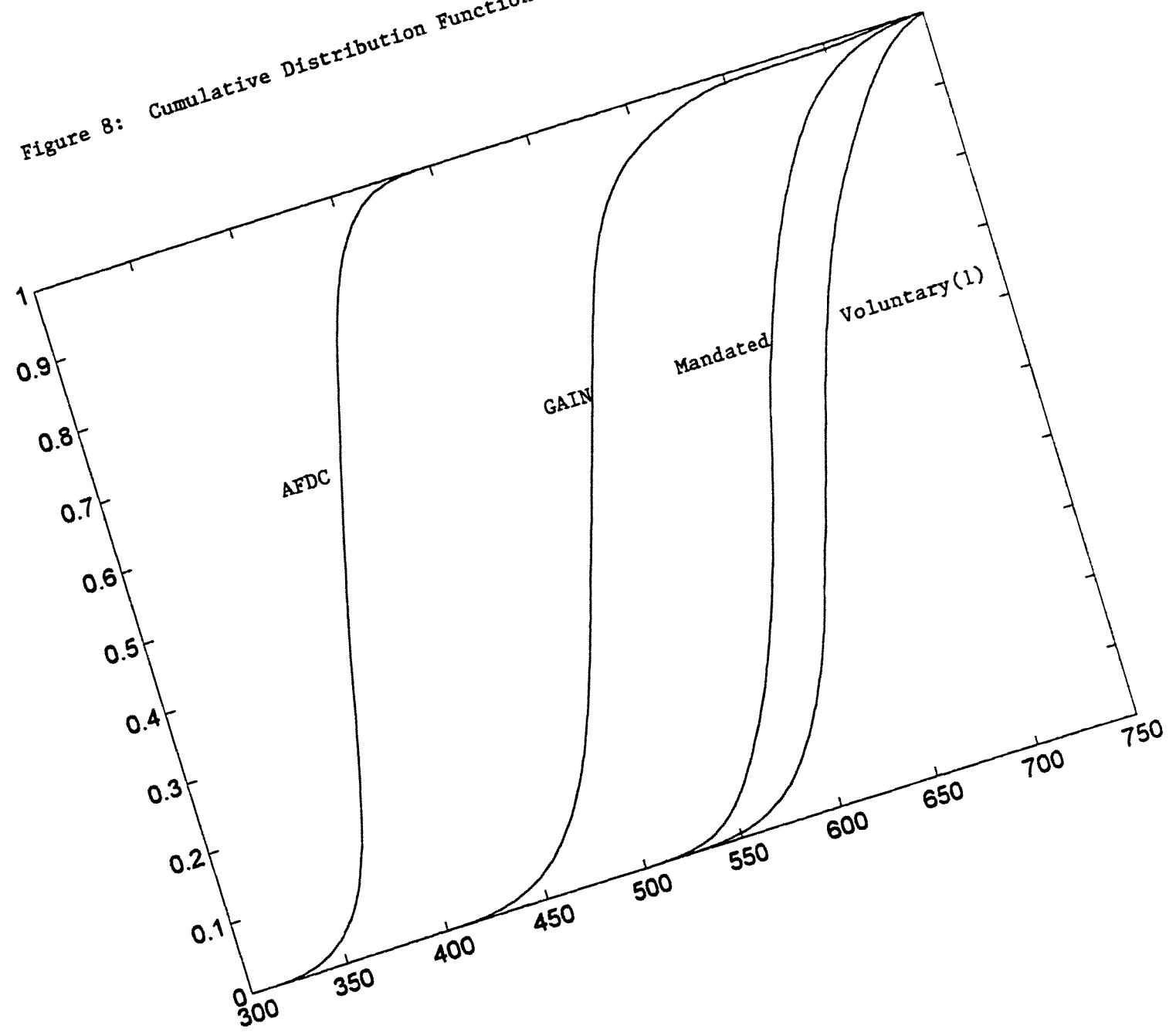
Log(Earnings+1), Solid=Treated, Dashed=Control

Figure 7: Cumulative Distribution Functions, Post-Treatment Earnings (Unit 1)



$\text{Log}(\text{Earnings}+1)$, Solid=Treated, Dashed=Control

Figure 8: Cumulative Distribution Functions, SWF(1)



Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Figure 9: Cumulative Distribution Functions, SWF(5)

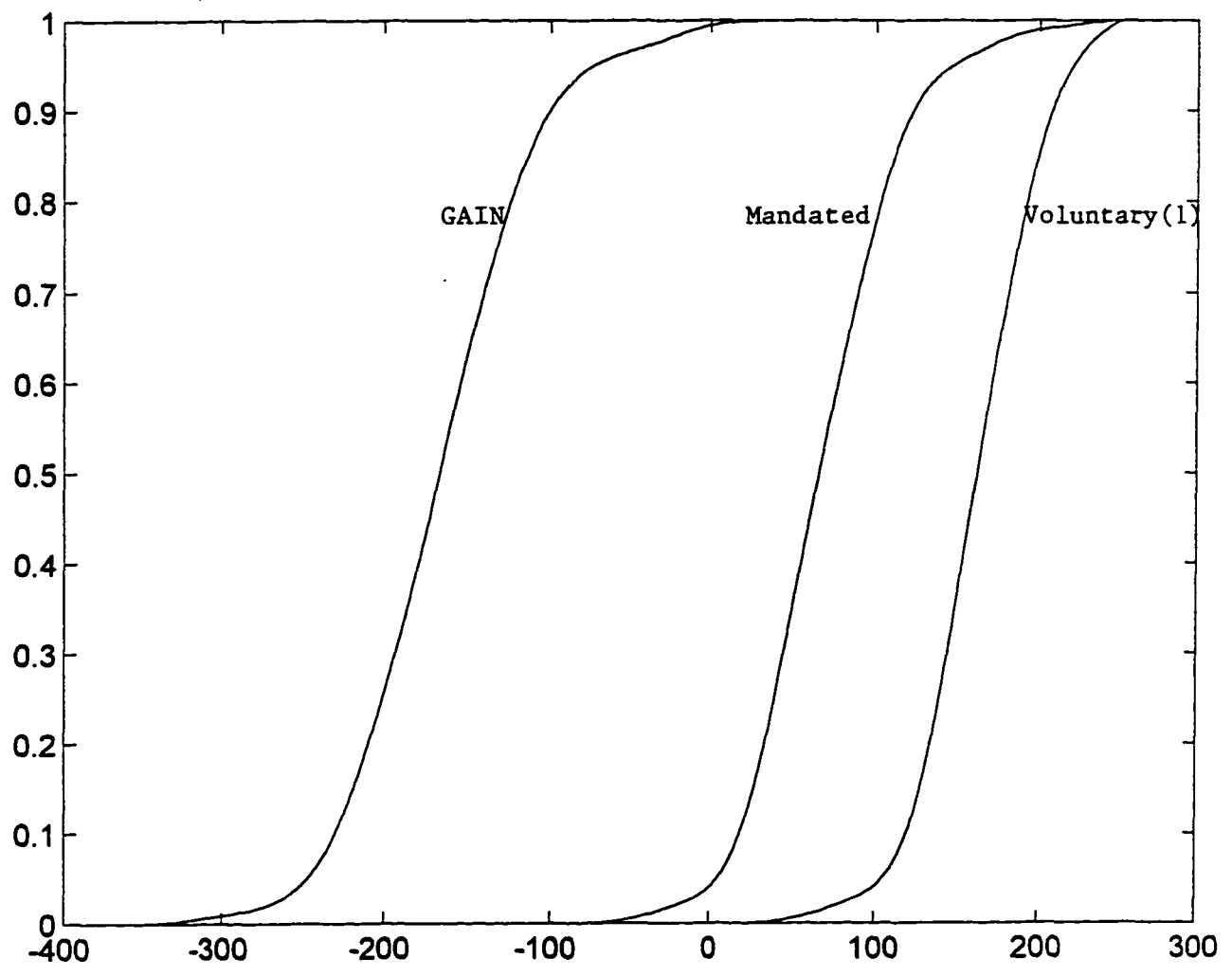


Figure 10: Percentiles of the Earnings Distribution
Average Earnings Per Quarter
(Medians of the Posterior Distributions of the Percentiles)

