Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Econometrics 125 (2005) 141–173
JOURNAL OF Econometrics
www.elsevier.com/locate/econbase

Program evaluation as a decision problem

Rajeev H. Dehejia[a,b,*]

[a] *Department of Economics and SIPA Columbia University 420 W. 118th Street, Room 1022, New York, NY 10027, USA*
[b] *NBER, 1050 Massachusetts Avenue, Cambridge, MA 02138, USA*

Available online 7 June 2004

Abstract

I argue for thinking of program evaluation as a decision problem. There are two steps. First, a counselor determines which program (treatment or control) each individual joins, based for example on maximizing the probability of employment or expected earnings. Second, the policymaker decides whether: to assign all individuals to treatment or to control, or to allow the counselor to choose. This framework has two advantages. Individualized assignment rules (known as profiling) can raise the average impact, improving cost effectiveness by exploiting treatment-impact heterogeneity. Second, it accounts systematically for inequality and uncertainty, and the policymaker's attitude toward these, in the evaluation.

JEL classification: C11; I38; J31

Keywords: Program evaluation; Profiling; Bayesian econometric

1. Introduction

This paper re-examines the Alameda portion of the Greater Avenues for Independence (GAIN) program with the aim of offering new methodological perspectives on program evaluation. Program evaluation is carried out by comparing the values of a range of outcomes of interest for a treatment and a control

*Corresponding author. Tel.: +1-212-854-4659; fax: +1-212-854-8059.
E-mail address: rd247@columbia.edu (R.H. Dehejia).

0304-4076/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2004.04.006

group, typically by considering the average treatment effect and its statistical significance.[1] For example, compared with the alternative—Aid to Families with Dependent Children (AFDC)—the GAIN program has a positive but not statistically significant average treatment effect on earnings and the probability of employment. Usually, the average treatment effect is also considered for subsets of the sample, defined based on pre-treatment characteristics.

The methodology that I adopt differs because it models program evaluation as a decision problem, and allows us to go beyond the average treatment effect along three dimensions. First, I consider how the program being evaluated will be made available subsequent to the evaluation. For example, will all individuals be required to participate in either the treatment program or the control program? These are the two options normally considered in evaluations. I also allow for the possibility that a counselor (caseworker in the context of welfare programs) can decide to which program—treatment or control—each individual will be assigned. This is a reasonable option to consider, because the practice of profiling program participants in order to determine receipt of services has become widespread (for example, in unemployment insurance, see Berger et al., 2001; and Runner, 1996).

Second, in choosing among the available options—which at this point include treatment, control, and assignment by a caseworker—I pay particular attention to how uncertainty about the outcome of interest affects the choice among programs. It is well known that a t-statistic does not embody all of the information relevant for a rational decisionmaker.[2] I therefore use predictive distributions—distributions which capture all of the uncertainty about the outcome of interest—which then allow for the use of standard expected utility theory in comparing the distribution of outcomes under the available programs. Finally, I allow for the policymaker to exhibit inequality aversion, which also entails looking beyond the average treatment effect.

Such issues have been largely ignored in the evaluation literature, with a few notable exceptions. Heckman and Smith (1998) (see also Heckman et al., 1997) rigorously consider the data requirements for evaluating various social welfare functions. Within the framework of their paper, the current paper focuses on social welfare functions that do not require information on the joint distribution of earnings under treatment and control. Manski (1999) (also Manski, 1995) develops non-parametric bounds for the expected welfare from different post-evaluation

[1] In randomized trials, such comparisons give unbiased estimates of the treatment effect (see Fisher (1935) and Neyman (1935)). In a non-experimental setting, the comparison would have to control for potential sources of sample selection bias. See, inter alia, Dehejia and Wahba (1999, 2002), Heckman (1989, 1990, 1992), Heckman and Hotz (1989), Heckman and Robb (1985, 1986), Heckman et al. (1998), Heckman and Smith (1995), Lalonde (1986), Manski (1989, 1993), and Manski and Garfinkel (1992).

[2] The finance literature has made a similar point in a very different context. See Kandel and Stambaugh (1996). In addressing the question "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 decision of interest. See also Barberis (2000) and Chamberlain (2000).

assignment rules, and Manski (2000) extends the analysis to the case where the policymaker's objective is not well defined. Both of his papers are complementary to the current research, because they explore related issues using non-Bayesian econometrics or non-standard decision theory. The contribution of this paper is that it offers an approach that unifies an analysis of individual-level heterogeneity with an analysis of the impact of risk and inequality aversion at the level of the policymaker.

Using the GAIN data, I demonstrate that the methodological contributions just outlined are important in understanding the impact of the GAIN treatment. I show first that a caseworker who maximizes participants' post-assignment probability of employment will assign less than half of the individuals into GAIN. In terms of the evaluation, this implies that the policy of assignment by a caseworker yields higher average post-assignment earnings than either of the other two policies (assigning all individuals into either GAIN or AFDC) that are normally considered. When it is selectively available through a caseworker, GAIN emerges as viable in a cost–benefit sense as well; this overturns the traditional evaluation of the program. More generally, whenever there is heterogeneity in the treatment impact, allowing for assignment by a caseworker will be of central interest.

Further, I show that the evaluation of the GAIN program changes significantly when one consistently accounts for uncertainty. In particular, the ranking that emerges between policies—for example, that assignment by a caseworker dominates GAIN, which in turn dominates AFDC, in terms of post-assignment earnings—is economically significant in the sense that the predictive distribution of earnings under one program first-order stochastically dominates the earnings distribution under the other program. In contrast, the ranking that emerges from a more standard *t*-test on the difference in means is equivocal; the difference is positive but not statistically significant.[3] Finally, I show that allowing for inequality aversion changes the ranking between GAIN and AFDC, with the latter preferred for moderately inequality-averse preferences.

GAIN is an interesting program to study not only because it is very similar to California's current welfare program (CalWORKs) but also because similar welfare-to-work programs have been initiated by many states since the 1980s (Greenberg and Wiseman, 1992, survey 24 such programs). At another level, GAIN is one in a long line of social experiments (see Burtless (1995) for a recent survey) and methodological conclusions about evaluating GAIN will be broadly relevant.

The paper is organized as follows. Section 2 briefly describes the GAIN program and experiment. Section 3 describes the econometric model that I use. Section 4

[3] A highly relevant issue that 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. See Dehejia (2003) and Hotz et al. (1999).

examines the decision problems for two typical individuals. Section 5 discusses the social decision problem and the choice of social welfare functions. Section 6 examines the results of the model at the social level, and Section 7 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. In this paper, we will confine ourselves to the Alameda County portion of the data. A companion paper (Dehejia, 2003) examines all six counties and the issues that arise in evaluating programs implemented across multiple sites.

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. For its evaluation, Alameda further confined itself to long-term welfare recipients (individuals already having received welfare for 2 years or more).[4] 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 was recorded for both treatment and control units, including: demographic characteristics, results of a reading and mathematics proficiency test, and data on 10 quarters of pre-treatment earnings. Table 1 summarizes the characteristics of the Alameda sample: 85 percent are women, who on average have more than two children; the mean level of education is grade 10; a quarter have previously participated in training programs; the average level of pre-treatment earnings is low, ranging from \$150 to \$190 per quarter, but because 87 percent of pre-treatment earnings are zero, the average of non-zero pre-treatment earnings is higher, on the order of \$1110 per quarter.[5]

Of those who attended the orientation session, half were randomly assigned into the GAIN program. These individuals continue to receive AFDC benefits, but face additional requirements and receive additional services (described below). The other

[4] This implies that the Ashenfelter (1978) "dip" in earnings cannot be observed in pre-assignment earnings.

[5] Seven individuals are excluded from the original sample because of apparent coding errors in their covariates. These seven individuals are either coded as having 70 children or a previous hourly wage of more than \$300.

Table 1
Data description, Alameda county

Variable	Mean	Standard deviation
Number of children less than age 4	0.19	0.49
Number of children between ages 4 and 5	0.23	0.46
Number of children between ages 6 and 11	1.16	4.68
Number of children between ages 12 and 18	0.88	2.29
Number of children aged 19 and greater	0.25	0.60
Score on reading test	206.27	98.00
Score on mathematics test	192.44	94.96
Grade	10.79	3.02
Most recently recorded hourly wage	3.74	2.73
Indicator for households with single head	0.62	
Age	35.39	8.85
Indicator for treatment status	0.50	
Indicator for female participants	0.86	
Indicator of refugee status	0.09	
Indicator for receiving AFDC in pre-assignment time	0.99	
Indicator for previous training or job search activities	0.24	
Ethnicity indicator, White	0.18	
Ethnicity indicator, Hispanic	0.08	
Earnings 10 quarters prior to experiment	165.02	740.14
Earnings 9 quarters prior to experiment	153.17	675.96
Earnings 8 quarters prior to experiment	154.53	747.70
Earnings 7 quarters prior to experiment	187.67	1036.91
Earnings 6 quarters prior to experiment	156.83	615.03
Earnings 5 quarters prior to experiment	170.37	771.74
Earnings 4 quarters prior to experiment	185.30	726.89
Earnings 3 quarters prior to experiment	151.60	685.37
Earnings 2 quarters prior to experiment	153.64	642.86
Earnings 1 quarter prior to experiment	167.17	714.04
Zero earnings 10 quarters prior to experiment	0.87	
Zero earnings 9 quarters prior to experiment	0.88	
Zero earnings 8 quarters prior to experiment	0.87	
Zero earnings 7 quarters prior to experiment	0.87	
Zero earnings 6 quarters prior to experiment	0.87	
Zero earnings 5 quarters prior to experiment	0.87	
Zero earnings 4 quarters prior to experiment	0.87	
Zero earnings 3 quarters prior to experiment	0.87	
Zero earnings 2 quarters prior to experiment	0.87	
Zero earnings 1 quarter prior to experiment	0.87	

half is assigned to a control group that is prohibited from receiving GAIN services.[6] Because assignment to treatment was random, the distribution of pre-

[6] 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 services compared with pre-existing employment services. To the extent that both groups receive AFDC benefits, the comparison is between the presence and absence of supplementary services and requirements.

assignment covariates is balanced across the treatment and control groups; Table 1 lists each of the covariates. In terms of the chronology of data gathering, "experimental" time (which I also refer to as post-assignment time) begins when individuals attend the GAIN orientation session. The early stages of post-assignment time thus coincide with the education and training part of the GAIN program.[7]

There are several components to the GAIN treatment: basic education, for those deemed to be in need of it (this includes either preparation for the General Educational Development certificate, Adult Basic Education, or English as a Second Language); job search; and job training, for those who do not find jobs (includes on-the-job training and paid or unpaid work experience). Participants were exempted from the requirement to participate in GAIN activities if they found work on their own.[8]

The outcome that we consider is earnings, which is observed for 13 quarters following assignment to treatment. From Tables 2 and 3, we see that GAIN's impact on both earnings and the probability of employment is negative in the first quarter; this is not surprising, since treatment units are participating in training activities in the first quarter. The treatment effect subsequently increases, ranging from 2 to 4 percent for the probability of employment and $200 for earnings (both are statistically significant).[9]

Finally, we should note that the assumption of a constant treatment effect across all individuals is very restrictive. The average treatment effect potentially embodies an array of heterogeneous treatment effects. Two examples illustrate this point. Fig. 1 depicts the interaction between the treatment effect and the score on the reading test: individuals who score 200 or more enjoy a higher treatment effect, although the standard error is quite large. In Fig. 2, we see that individuals who have previously participated in training programs also enjoy a higher treatment effect. Although these interactions are not statistically significant, we will see that they have a substantial impact on the decision problem.[10]

[7] 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 from 1 or 2 months prior to the commencement of the experiment. 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-assignment data would cover the 10 quarters from July 1986 to December 1988.

[8] Only about 85 percent of the treated units actively participated in any GAIN activities; the balance satisfied the requirements of the GAIN program on their own (in most cases by 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 the impact of the GAIN program as a whole rather than the components of the treatment, because some portion of the impact is through participants who find work in order to avoid the burden of participating in treatment activities. See Black et al. (1999).

[9] An earlier version of this paper (Dehejia 1997) examines the impact of lagged employment status on the treatment impact, and shows that the treatment increases the probability of transition from non-employment to employment. For a more detailed analysis of this issue, see Ashenfelter and Card (1985) and Eberwein et al. (1997).

[10] It is well known that statistical significance is not the criterion that a rational decisionmaker considers in choosing between alternatives. For example, in the finance literature, Barberis (2000) notes that the predictability of stock market returns is not statistically significant, yet it is sufficient to influence an agent's portfolio choice (relative to bonds) over a sufficiently long horizon.

Table 2
Treatment effect on probability of employment

Post-experimental period	Treatment effect	Standard error
1	−0.024	0.017
2	0.011	0.019
3	0.019	0.020
4	0.034	0.021
5	0.044	0.021
6	0.018	0.021
7	0.021	0.022
8	0.048	0.022
9	0.062	0.023
10	0.061	0.023
11	0.044	0.023
12	0.027	0.022
13	0.063	0.022

A probit is used; covariates include variables for the number of children, reading and writing test scores, grade, age, sex, ethnicity, and earnings histories. 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 3
Regression coefficients of treatment indicator for post-assignment earnings

Post-assignment period	OLS treatment effect,with covariates	Standard error
1	−47.6	22.8
2	−9.2	39.2
3	35.1	45.9
4	67.6	56.8
5	111.3	54.1
6	85.0	61.1
7	84.8	66.5
8	95.3	68.6
9	203.1	76.0
10	232.1	79.5
11	194.5	86.4
12	150.7	88.8
13	206.6	90.8

All earnings are in 1988 dollars.

3. A model of the earnings data

3.1. The statistical model

Let Y_{ij}^t denote earnings, where $j = 1$ (GAIN) or 0 (AFDC), i = 1,…, 1360, and t = 1,…,13. Y_{i1}^t is interpreted as individual i's earnings in period t if she was in

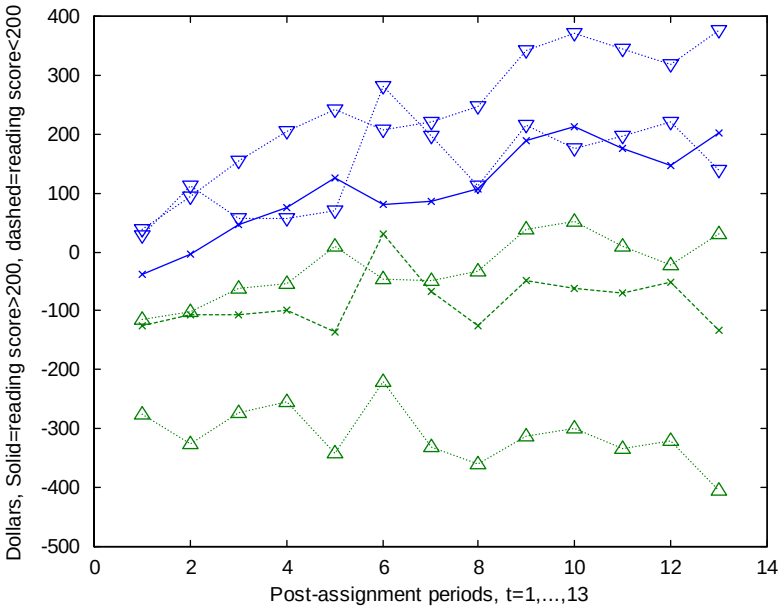

Fig. 1. Comparing individuals with reading score > 200 to those with reading score < 200: Treatment effect (+,−2 standard errors).

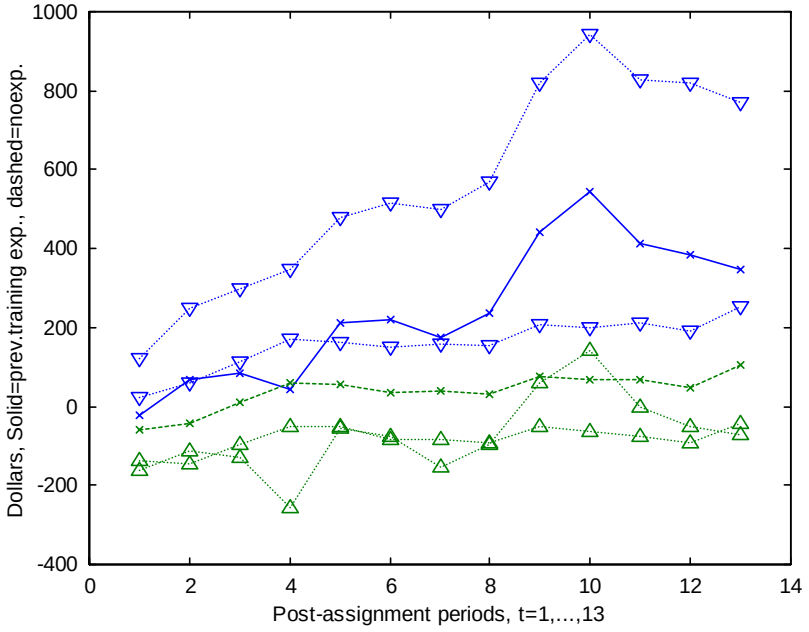

Fig. 2. Comparing individuals with previous training experience to those without previous experience: treatment effect (+/− two standard errors).

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:

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

where T_i is a treatment indicator ($=1$ if individual i was in fact 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 2, is the mass point in the distribution of earnings at zero. The strategy adopted is to use a censored normal likelihood, the Tobit model. Following Chib (1992), define a latent variable, y_{it}^*, which determines which value y_{it} takes on;

$$y_{it} = \begin{cases} y_{it}^* & if \quad y_{it}^* \geqslant 0, \\ 0 & if \quad y_{it}^* < 0. \end{cases} \tag{2}$$

For the Tobit model

$$Y_{it}^* \left| \{X_{it} = x_{it}\}_{t=1}^{13}, \beta, \sigma \sim \mathrm{N}(x_{it}\beta, \sigma). \right. \tag{3}$$

The vector of explanatory variables is given by $x_{it} = (1_{1\,it}, \ldots, 1_{13\,it}, [1_{1\,it} \ldots 1_{13\,it}] \cdot \mathrm{T}_i, \mathrm{Z}_i, \mathrm{Z}_i \cdot \mathrm{T}_i, \mathrm{R}_{it})$. $[1_{1\,it} \ldots 1_{13\,it}]$ is a set of indicator variables for each quarter of post-assignment time ($1_{k\,it} = 1$ if $t = k$, $=0$ otherwise), giving each period its own intercept. The treatment indicator is interacted with $[1_{1\,it} \ldots 1_{13\,it}]$. Since each period corresponds to experimental, rather than calendar, time, the treatment dummies produce a profile of the treatment effect over 13 quarters. Exogenous regressors, Z_i and their interactions with the treatment indicator are also included, which allow the treatment effect to vary with observable pre-treatment characteristics. These characteristics include: indicators for the age and number of children, race and ethnicity, educational attainment, score on the reading and mathematics tests, sex, an indicator for previous participation in other training programs, and 10 periods of pre-assignment earnings history.[11] A calendar time trend, R_{it}, is also included.[12]

I use diffuse priors for the parameters of the model. Appendix A discusses the estimation procedure in detail, and Appendix B summarizes the posterior distributions of the parameters.

[11] A fixed set of pre-assignment earnings is used; thus these are pre-determined rather than auto-regressive variables.

[12] Note that the earnings process is i.i.d., conditional on covariates. This specification allows for persistent differences in earnings across individuals through the permanent, rather than transitory, component. The source of heterogeneity is individual exogenous characteristics, which are also interacted with the treatment indicator. Note also that the model is not interpreted structurally; it is used predictively. It would be an interesting extension to consider more general specifications for earnings processes (see, for example, Hirano, 2000), which might improve predictions to some extent, but for the questions which I examine the current model produces predictions that are robust to generalizations of the model (e.g., to a mixture of normals or to allow for additional serial correlation in earnings).

3.2. The predictive distribution

Because the decision problems associated with program evaluation are in the space of outcomes, not the space of the parameters of the model, it is important to construct a distribution in the outcome space which embodies all of the uncertainty from the model (i.e., conditional on parameters) and from parameter estimation. This is called the (posterior) predictive distribution.

Imagine predicting earnings for an $(I+1)$st individual. This individual is identified by Z_{I+1}, a set of exogenous variables. By specifying the time dummies and the treatment indicator, we construct $x_{I+1\,t}$. Conditional on parameters, $(\beta^{(i)},\sigma^{(i)})$, we can simulate the outcome distribution by drawing for Y_{I+1t} (from the likelihood (3), $N(\beta^{(i)}x_{it},\sigma^{(i)})$). To obtain the predictive distribution, we must account for parameter uncertainty; thus, we use draws from the posterior distribution of the parameters (obtained from the Gibbs sampler outlined in the appendix): $\beta^{(i)},\sigma^{(i)} \sim p(\beta,\sigma|Data)$. For each draw, we simulate the outcome distribution from the likelihood. Using this procedure, we obtain the joint predictive distribution of earnings for individual $I+1$ from periods $1,\ldots,13$. To vary the treatment status, we re-specify x_{it} by switching the treatment indicator.

3.3. The choice and fit of the model

A major issue is the choice of likelihood. The predictive distribution only captures uncertainty correctly if the model is specified correctly. Figs. 3 and 4 give a sense of the fit of the model. These figures show the density of the empirical distribution of average earnings for treated and control units (estimated through a histogram), and plot the density of the predictive distribution of average earnings.

As we can see from the figures, the empirical distributions of earnings for treated and control units are well approximated by the truncated normal density. The model also fits the mass point with reasonable accuracy: an empirical probability of employment of 0.2047 for treated units compared to a predictive probability of 0.2040, and 0.1852 vs. 0.1756 for control units.[13]

4. The individual-level impact

This section studies the individual-level impact of the GAIN treatment. It not only provides a detailed view of the impact of the program, but also lays the foundations for the analysis in Section 6 of the social welfare problem.

[13] A non-parametric model is another option. I use a parametric model because it allows me to incorporate many explanatory variables, which is important in this application and is difficult in a non-parametric setting. An alternative would be to use a flexibly parametric model. The results would be similar to those presented here (see, for example, Dehejia, 1999b).

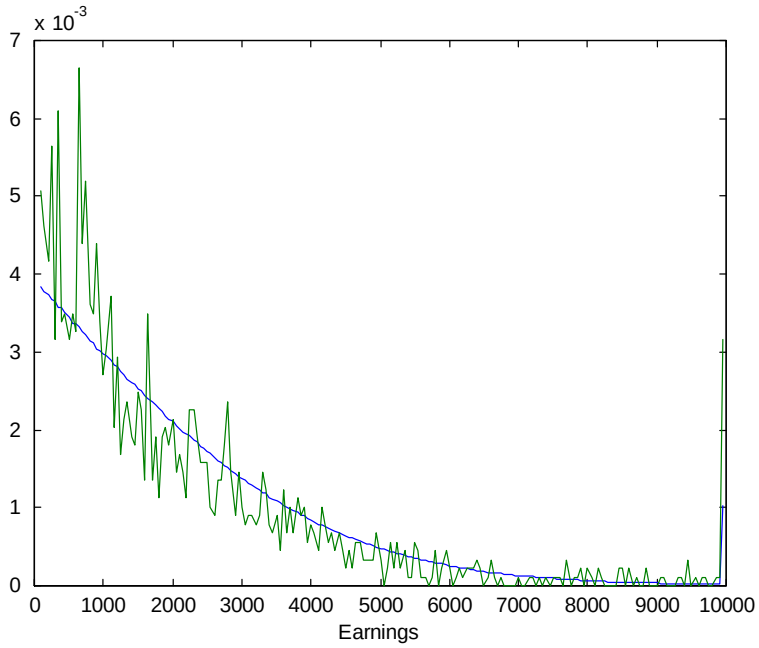

Fig. 3. Probability densities: empirical vs. predictive, treated group.

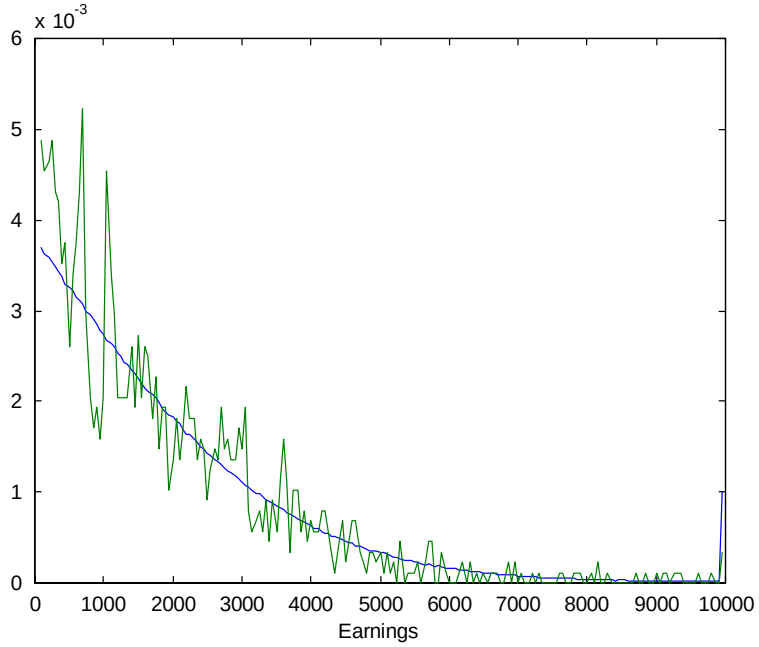

Fig. 4. Probability densities: empirical vs. predictive, control group.

Imagine that a caseworker has to choose whether to assign an individual into GAIN or AFDC. For the caseworker, an individual is identified by her pre-assignment characteristics. Thus, the key assumption is that the individual under consideration is exchangeable with those in the data; that is, earnings for individuals with the same covariates are taken to be drawn from the same distribution. A rather strong implication of this assumption is that the caseworker does not have (or use) any private information—i.e., information that is not observed by the researcher—in the assignment decision.[14]

4.1. Two typical examples

Table 4 shows 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. Ms. Ten Forty-Three is a clear winner from GAIN, and Ms. Eight Twenty-Two is a clear loser. Ms. Ten Forty-Three is of age 23, heads a single-parent household, has one child between the age of 6 and 11, and has completed high school. Her earnings history shows that she was employed in each of the quarters prior to the experiment. Ms. Eight Twenty-Two is a 41-year-old woman, the head of a single-parent household, has one child between the age of 12 and 18, and has completed high school. Her earnings history shows substantially higher earnings in all but one of pre-assignment periods. Let us consider each individual in turn.

Table 5a shows the probability of positive earnings, and the mean and standard deviation of the predictive distribution of earnings, for each period under both treatment and control for Ms. Ten Forty-Three. For each of the 13 periods, the probability of positive earnings and the mean of earnings are higher in the GAIN treatment than in the control program. The profile of the treatment effect is increasing, in a pattern similar to that depicted in Table 3 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 with the magnitude of the standard deviation (i.e. there is substantial overlap in the predictive distributions of earnings).[15] Is the difference between treatment and control earnings significant?

Within a decision framework, we could say that the difference is significant if a wide range of decisionmakers would opt for the treatment distribution over the control distribution. Fig. 5 depicts the cumulative distribution

[14] 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. If the individual herself is making the choice, but any private information she has is independent of the observed covariates, then there would be no systematic errors in terms of average earnings for the group of interest.

[15] The difference in means is not significant in the sense that the 95 percent probability intervals of the posterior distributions substantially overlap. But 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.

Table 4
Characteristics of three typical individuals

Variable	Ms. 1043	Ms. 822	Ms. 397
Number of children less than age 4	0	0	0
Number of children between ages 4 and 5	0	0	0
Number of children between ages 6 and 11	1	0	0
Number of children between ages 12 and 18	0	1	3
Number of children aged 19 and greater	0	0	1
Score on reading test	253	218	227
Score on mathematics test	228	212	222
Grade	12	12	9
Age	23	41	16
Female	1	1	0
Indicator of refugee status	0	0	0
Indicator for previous training or job search activities	0	0	0
Ethnicity indicator, White	0	1	0
Ethnicity indicator, Hispanic	0	0	0
Earnings 10 quarters prior to experiment	5687	11,598	0
Earnings 9 quarters prior to experiment	2992	11,124	0
Earnings 8 quarters prior to experiment	5397	15,729	0
Earnings 7 quarters prior to experiment	4391	29,852	0
Earnings 6 quarters prior to experiment	6232	0	0
Earnings 5 quarters prior to experiment	3186	11,660	0
Earnings 4 quarters prior to experiment	4171	11,660	0
Earnings 3 quarters prior to experiment	4577	15,000	0

functions (CDFs) for the predictive distribution for each of the 13 periods. In each of the 13 post-assignment periods, treatment earnings first-order stochastically dominate control earnings. Any risk-neutral or risk-averse agent (whose preferences are increasing in earnings) would prefer the treatment distribution. This is a simple illustration of the fact that even when the means of the two distributions under consideration are not very different, the underlying decision may be clearcut.

For Ms. Eight Twenty-Two, it is a different matter. In Fig. 6 we see that her distribution of earnings in the control first-order stochastically dominates her distribution of earnings in treatment in each period. As long as more earnings are preferred to less, the caseworker unambiguously would not assign her to participate in GAIN. Of course, first-order stochastic dominance does not suffice to compare all the distributions that arise. In general, expected utility comparisons would be required.

4.2. *The importance of accounting for uncertainty*

A natural question that arises from the preceding analysis is: would similar decisions have been reached if uncertainty had not been accounted for as

Table 5
Mean and variance of predicted earnings: (a) Ms. Ten Forty-Three; (b) Ms. Eight Twenty-Two. Predicted earnings, with and without uncertainty, Ms. Three Ninety-Seven

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
(*a*)						
1	0.81	3327	2827	0.68	2329	2474
2	0.85	3773	2893	0.7	2496	2512
3	0.87	3942	2927	0.73	2689	2618
4	0.9	4185	2897	0.72	2524	2637
5	0.9	4341	3037	0.71	2438	2519
6	0.9	4219	2931	0.74	2665	2617
7	0.89	4386	3096	0.75	2800	2698
8	0.9	4499	3035	0.77	2852	2639
9	0.92	4819	3121	0.74	2623	2587
10	0.91	4747	3142	0.74	2755	2685
11	0.91	4717	3125	0.77	2896	2640
12	0.92	4765	3136	0.75	2794	2638
13	0.93	4823	2973	0.79	2969	2670
(*b*)						
1	0	12,406	3355	0	16,739	4487
2	0	12,818	3546	0	16,804	4432
3	0	13,236	3346	0	16,907	4413
4	0	13,652	3361	0	17,065	4355
5	0	13,683	3412	0	16,865	4420
6	0	13,582	3440	0	16,924	4448
7	0	13,695	3365	0	17,230	4300
8	0	13,678	3376	0	17,074	4302
9	0	14,259	3354	0	17,071	4349

10	0	14,186	3460	0	17,145	4238
11	0	14,271	3413	0	17,369	4413
12	0	14,265	3356	0	17,388	4337
13	0	14465	3295	0	17,133	4398

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
(*c*)								
1	129	539	191	690	107	554	208	739
2	217	802	218	716	137	568	231	795
3	190	659	208	683	163	634	234	791
4	209	736	202	689	257	841	222	741
5	198	709	257	778	210	751	223	750
6	215	745	193	711	261	836	191	652
7	239	820	289	843	267	892	260	823
8	224	713	271	823	276	866	256	802
9	340	966	298	915	302	944	242	776
10	323	948	333	998	336	955	255	777
11	324	923	311	905	309	920	295	867
12	352	962	301	863	354	999	327	970
13	293	864	266	860	368	1010	249	837
Probability of employment	0.8575		0.8491		0.8494		0.8532	
Average earnings	250		257		257		245	
Expected utility (CRRA, $q = 3$)	0.5000		0.5000		0.5000		0.5000	

All earnings are in 1988 dollars

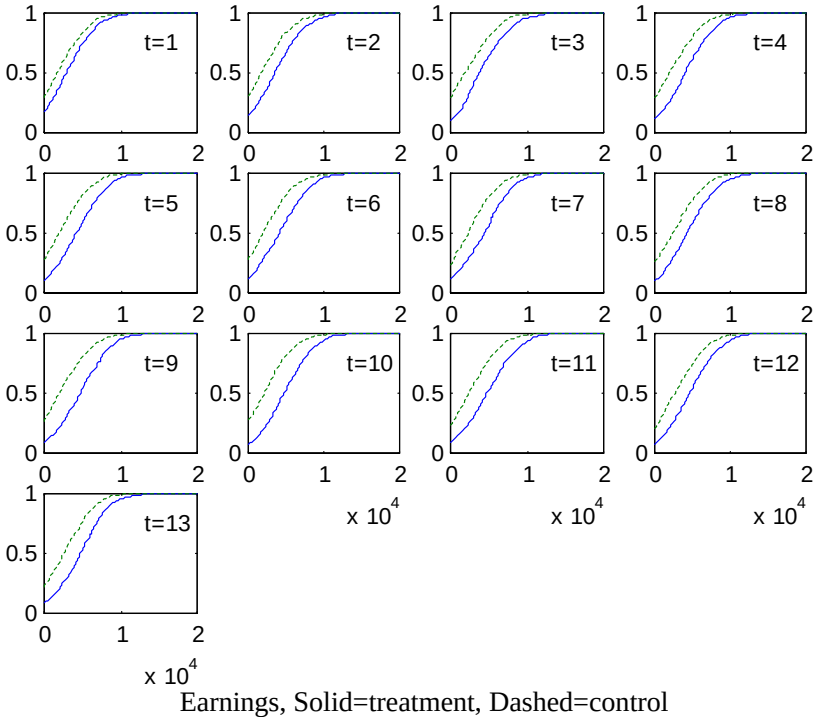

Fig. 5. CDFs of predicted earnings for Ms. Ten Forty-Three.

comprehensively? In particular, one might imagine using the model described in Section 3 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 without parameter uncertainty, the intrinsic uncertainty embodied in the likelihood (3) has to be taken into account. Table 5c presents such an exercise for Ms. Three Ninety-Seven, whose characteristics are given in Table 4. Columns 1–4 of Table 5c 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 5–8 summarize 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, but the sign of the treatment effect is reversed. When uncertainty is ignored, the mean of predicted control earnings is higher; when we account for uncertainty, the reverse is true. If the decisionmaker exhibits risk aversion, then the differences between the two programs are not as extreme. The final row of the table shows that the expected utility (with log preferences) of each program is the same.

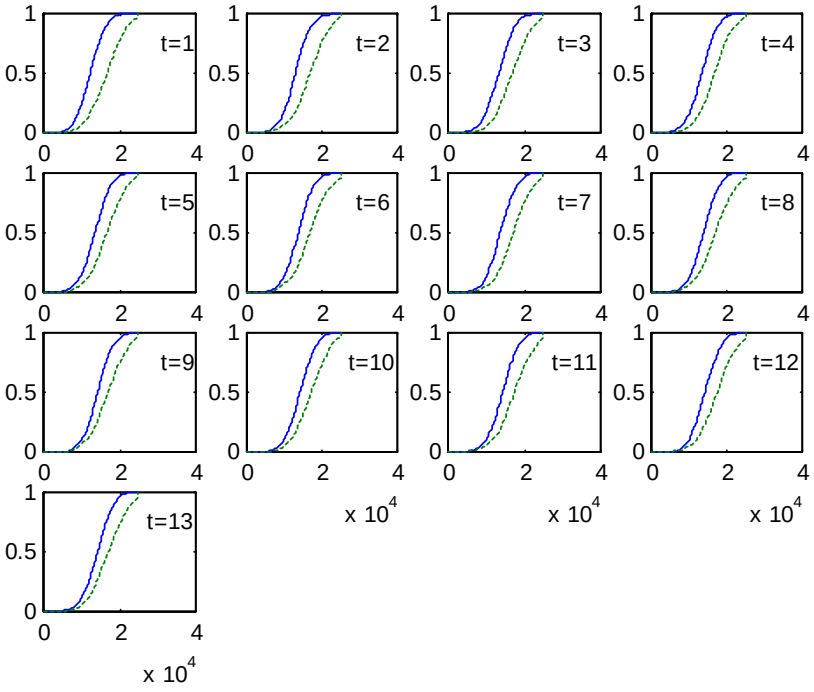

Fig. 6. CDFs of predicted earnings for Ms. Eight Twenty-Two.

Of course, this example was chosen precisely because ignoring uncertainty leads to different advice than accounting for it. In cases where the two distributions are starkly different, ignoring uncertainty typically would not lead to a change in the decision. For the overall sample from Alameda, uncertainty affects decisions for about 10 percent of individuals.

4.3. *The importance of heterogeneity*

We could consider such decision problems for a wider array of individuals. The differences in the results would reflect the underlying heterogeneity in the treatment effect. One view of this is presented in Table 6. Assume that each of the 1360 individuals in the Alameda sample will be assigned to either GAIN or AFDC and, as in the previous examples, imagine that the decision is made by a caseworker based on the predictive distributions of their earnings in each period under each program. Two criteria are considered: maximizing the probability of post-assignment employment and assigning to

Table 6
Groups benefiting the most and least from GAIN

Variable	Prob. of employment		Expected earnings exceed costs	
	Higher in GAIN	Higher in AFDC	Yes	No
Number	766	594	249	1111
Average earnings GAIN	567	250	912	320
Average earnings AFDC	323	363	380	331
Number of children less than age 4	0.16	0.24	0.23	0.18
Number of children between ages 4 and 5	0.23	0.23	0.19	0.24
Number of children between ages 6 and 11	0.85	0.93	0.83	0.9
Number of children between ages 12 and 18	0.77	0.89	0.73	0.84
Number of children aged 19 and greater	0.15	0.37	0.11	0.28
Score on reading test	214	196	214	204
Score on mathematics test	199	183	199	191
Grade	11.98	9.25	12.43	10.42
Age	33.52	37.76	33.16	35.87
Female	0.90	0.80	0.90	0.85
Indicator of refugee status	0.07	0.12	0.09	0.09
Indicator for previous training or job search activities	0.40	0.04	0.8	0.12
Ethnicity indicator, White	0.16	0.20	0.15	0.19
Ethnicity indicator, Hispanic	0.05	0.10	0.05	0.08
Earnings 10 quarters prior to experiment	181	146	293	137
Earnings 9 quarters prior to experiment	132	182	186	147
Earnings 8 quarters prior to experiment	177	127	295	124
Earnings 7 quarters prior to experiment	195	180	324	158
Earnings 6 quarters prior to experiment	182	126	362	112
Earnings 5 quarters prior to experiment	218	111	507	96
Earnings 4 quarters prior to experiment	226	136	422	133
Earnings 3 quarters prior to experiment	135	174	270	125

All earnings are in 1988 dollars.

treatment individuals for whom the expected increase in earnings exceeds the training costs.[16]

Table 6 presents the mean of expected post-assignment earnings under GAIN and AFDC and the mean of pre-assignment covariates. We see that depending on the criterion either 18 or 56 percent of the sample are faring better under the treatment. The average treatment effect is \$323 (−\$113) for those (not) assigned to treatment under the first criterion, and \$532 (−\$11) for those (not) treated under the second criterion. Comparing those assigned to treatment and control for each criterion is revealing. For both criteria, those benefiting from GAIN generally have fewer

[16] The GAIN public use file does not contain information on which services individuals received if they participated in the treatment. In the absence of this data, we assume that the cost of the GAIN program is the same across participants, which is estimated at \$3638 for 13 quarters (Riccio et al., 1994). Hotz et al. (2000) obtain data on which treatment participants received.

children (except under age 4 for the second criterion), have higher scores on the reading and mathematics tests, have a higher level of educational attainment, are younger, and often have participated previously in training programs. Of particular note is the difference in the level of pre-assignment earnings, which are by and large higher for those benefiting from treatment (though Ms. Eight Twenty-Two is an exception). Comparing the two criteria, we see that those who are assigned to treatment by the second criterion have on average an even higher level of pre-assignment earnings, and a greater proportion have previously participated in training programs (0.8 compared to 0.4).

Table 6 in essence arrives at profiles of the beneficiaries and non-beneficiaries of GAIN. These profiles are not a substitute for individual predictions (as in Section 4.1) or an overall evaluation (as in Section 6, below). But they do allow us to generalize to some extent about the attributes of those who benefit from the treatment. These profiles are of great relevance in the contemporary policy environment, because welfare agencies in fact are now profiling program participants to determine who should receive supplemental services (see inter alia Berger et al., 2001). The method described in this section achieves this profiling in a systematic manner.

5. The social choice problem

Thus far, the analysis has focused on the individual-level decision between GAIN and AFDC. This section takes the next step by asking: how can the policymaker decide which program or combination of programs to make available, given the pattern of individual effects? There are two steps in this decision.

First is choosing the set of policies under consideration in the post-evaluation environment, where policies determine each individual's assignment to treatment. I consider the following alternatives: (1) All individuals are required to participate in GAIN; (2) All individuals remain in AFDC; (3) A caseworker assigns each individual into the program in which she is most likely to "succeed". Success is defined by a range of criteria that include the probability of finding employment, expected earnings, the increase in earnings net of program costs, and the expected utility of earnings. I consider two expected utility functions: log and constant relative risk aversion (CRRA) with the coefficient of relative risk aversion equal to 3.[17] Note that these are the preferences that the caseworker uses to assess individual earnings.

The second choice is the set of criteria (that is, social welfare functions or SWFs) that the policymaker uses to decide which policy to adopt. I consider two sets of alternatives. The first set ignores issues of inequality and focuses on outcomes averaged over the 13 post-assignment periods that are under consideration. These outcomes include average earnings, the probability of employment, and the increase

[17] The literature has suggested a range of values between 0 and 5. Friend and Blume (1975) obtain indirect evidence from individual asset holdings. They estimate a value between 2 and 3.

in earnings net of program costs.[18] In the absence of uncertainty, each SWF would produce a single number (for example, average earnings) for each program. Because there is uncertainty, we integrate out the unknown parameters and the intrinsic uncertainty, producing a predictive distribution of each SWF for each program. In practice, this amounts to drawing for the outcome from each individual's predictive distribution of earnings, computing the SWFs, and repeating this procedure until the distribution is well approximated.

The second set of SWFs allows for inequality aversion. Four standard SWFs are considered: utilitarian (the inequality-neutral benchmark); and exponential with coefficient of inequality aversion ranging from one (log preferences, slightly inequality averse), to three (intermediate), to infinity (Rawlsian) (see Deaton and Muellbauer, 1980). For these SWFs, we account for uncertainty by first collapsing each individual's predictive distribution of earnings in each program to its certainty equivalent, and then applying the SWFs to the certainty equivalents.[19]

6. Accounting for risk and inequality aversion

A useful benchmark for the social evaluation of GAIN is the conclusion reached using differences in means rather than predictive distributions; I consider this below. Table 7 presents the three social welfare criteria discussed in Section 5. For post-assignment earnings and the probability of employment, there is a positive but insignificant treatment impact. For earnings net-of-costs, the impact is significant.

Tables 8–10 apply the social welfare analysis outlined in Section 5 to the predictive, rather than the empirical, distributions of outcomes under treatment and control, allowing for a range of post-evaluation assignment mechanisms. I simulate the predictive distribution of earnings under treatment and control for 13 quarters of post-assignment earnings for each of the 1360 individuals in the sample.

Table 8 displays the first set of social welfare criteria discussed in Section 5. Consider first average post-assignment earnings per person per quarter. From the first two cells of column 1, the mean predictions from the model ($428 for GAIN and $340 for AFDC) are similar to those obtained from the empirical distribution, within $35 for both GAIN and AFDC. The 95 percent posterior confidence intervals do not overlap. Cells 3–6 show that the policies of assigning individuals based on their probability of post-assignment employment, or based on expected (or the expected utility of) earnings, yield substantially higher average quarterly earnings than the policy of enrolling everyone in GAIN ($478 compared with $428).[20] This is not surprising in light of Table 6, which reveals substantial heterogeneity underlying the

[18] These criteria ignore the disutility that might be experienced by program participants from reduced leisure or a change in job attributes. See Dehejia (1999a) and Greenberg (1997).

[19] A zero discount rate is assumed, but the results are not sensitive to this choice.

[20] The findings are sharper than would be obtained from an extreme bounds analysis (see Manski, 1995, 1999, 2000). 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.

Table 7
Comparing GAIN and AFDC, average of outcomes per person[a]

Policy	Labor earnings per quarter	Probability of employment	Earnings net-of-costs per quarter[b]
GAIN	463	0.2042	183
AFDC	372	0.1843	372
Difference	91	0.0199	−189
Standard error on difference	56	0.018	56

[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 8
Social welfare comparisons for Alameda

Policy(number in treatment)	Social welfare functions		
	Average earnings	Probability of employment	Average increase in earnings net of costs[a]
GAIN (1360)	428 [402,456]	0.2042	−192 [−227,−156]
AFDC (0)	340 [317,364]	0.1758	0 [0,0]
Maximize probability of employment (766)	478 [451,505]	0.2258	−20 [−47,6]
Maximize expected earnings (773)	478 [451,506]	0.2257	−21 [−48,5]
Maximize expected utility (log preferences)(756)	478 [451,505]	0.2257	−18 [−44,8]
Maximize expected utility (CRRA, $q = 3$)(648)	476 [450,502]	0.2249	2 [−22,26]
Expected increase in earnings must exceed cost (249)	437 [411,463]	0.2092	46 [28,64]

Each set of values of the parameters from the posterior distribution defines a state of the world. For each state of the world the social welfare functions are computed. Thus, there is a distribution of these SWFs over the various states of the world. These are summarized by the mean and the 2.5 and 97.5 percentiles of the distributions.
[a] The estimated costs of the GAIN treatment are $3638 for 13 quarters.

average treatment effect. Even though the heterogeneity may not be statistically significant, it is economically significant in the sense that a decisionmaker (with preferences ranging from risk-neutral to moderately risk-averse) would opt not to assign a significant fraction of the sample to the treatment.

From column 1 of Table 8, it is not possible to determine to what extent the policymaker's risk attitude would affect the ranking of the programs. The fact that the predictive distributions of the GAIN and AFDC options do not overlap suggests that the difference is significant, but the policies with individual assignment do overlap with the policy of assigning all individuals to GAIN. The advantage of working with the predictive distribution of the social welfare values is seen in Fig. 7, which plots the CDFs of the predictive distributions for the first four assignment rules. We note that GAIN first-order stochastically dominates AFDC, and in turn is

Table 9
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]	42 [25,62]	178 [156,202]	333 [303,364]	501 [464,541]	453 [418,488]
AFDC	0 [0,0]	51 [36,68]	160 [140,181]	276 [251,301]	393 [362,426]	361 [331,390]
Maximize probability of employment	0 [0,0]	96 [79,114]	233 [212,255]	386 [357,414]	546 [511,582]	494 [461,528]
Maximize expected earnings	0 [0,0]	96 [79114]	233 [212,257]	386 [358,414]	546 [511,584]	495 [463,528]
Maximize expected utility (log preferences)	0 [0,0]	96 [79,114]	233 [212,257]	386 [357,414]	546 [510,583]	495 [462,528]
Maximize expected utility (CRRA, $q=3$)	0 [0,0]	95 [77,112]	231 [209,253]	383 [355,411]	543 [508,580]	493 [461,524]
Expected increase in earnings must exceed cost	0 [0,0]	72 [54,90]	196 [175,220]	336 [309,364]	486 [451,523]	443 [410,476]

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

Table 10
Expected utility comparisons, Alameda

Individual Preferences/ Policy	Risk neutral	Log	Exponential[a]	Rawlsian
Risk Neutral				
GAIN	428.3	317.6	65.3	2.1
AFDC	340.1	294.7	271.5	70.9
Mandated	477.7	394.3	332.2	156.2
Choice (1)	477.8	394.4	332.4	156.2
Risk Averse, log				
GAIN	416.8	307.8	64.5	2.1
AFDC	337.1	291.7	268.7	70.2
Mandated	467.9	386.0	326.1	154.5
Choice (2)	468.0	386.1	326.2	154.5
Risk Averse, CRRA ($q = 3$)				
GAIN	390.7	285.7	75.2	3.0
AFDC	330.4	285.2	262.8	74.1
Mandated	445.6	366.8	311.8	153.0
Choice (3)	445.5	366.6	311.4	153.0

Expected utilities are normalized.
[a] $\mathrm{SWF} = \frac{1}{(1-\varepsilon)}\left(\sum_i (u_i)^{(1-\varepsilon)}\right)$, $\varepsilon = 3$, applied to the certainty equivalent of the individual income distribution.

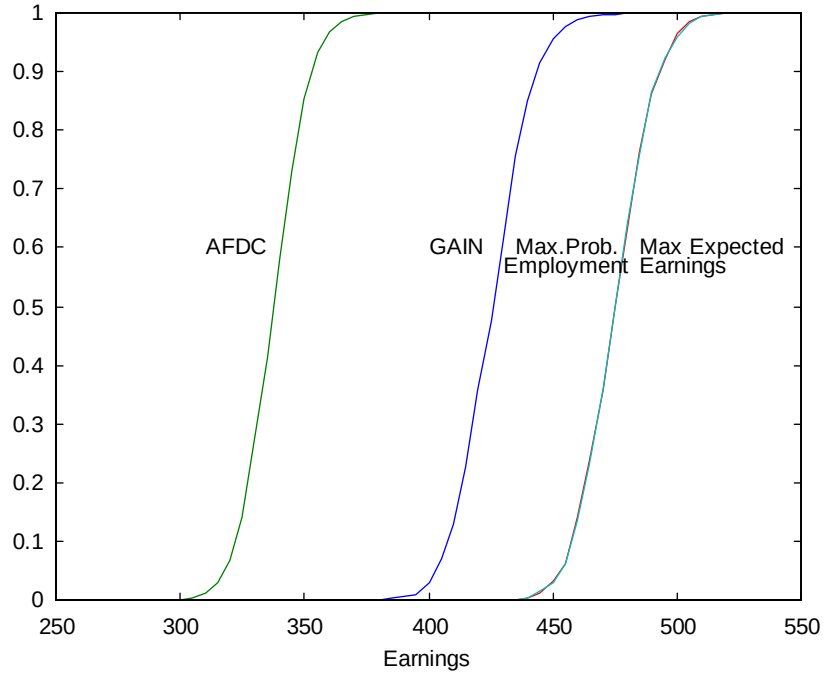

Fig. 7. Predictive distributions for average earnings.

dominated by individual assignment policies. The ranking between the two individual assignment policies is unclear, since they produce an almost identical assignment.

The second social welfare function ranks the policy alternatives by post-assignment probability of employment. GAIN (in keeping with its stated mandate) does succeed in increasing the probability of employment relative to AFDC, although the magnitude of the difference is not large (0.21, compared with 0.18). The policy of individual assignment based on probability of employment, by definition, maximizes the post-assignment probability of employment (0.22, compared with 0.20 for GAIN). Again, the outcome is very similar when individuals are assigned based on expected post-evaluation earnings.

The third column reveals that the increased earnings realized by assigning all individuals into GAIN do not offset the increased costs when compared with AFDC (a net difference of −$192 per person per quarter), nor are the costs of treatment offset by increased earnings when individuals are assigned by a caseworker based on the probability of employment or expected earnings. For assignment based on risk-averse (CRRA(3)) preferences, the program appears to break even. The final row in Table 8 highlights this point by considering assigning only those individuals to treatment for whom the expected increase in earnings offsets the increased costs. This maximizes the third social welfare criterion, which achieves a value of $46. Fig. 8 illustrates that these policies, again, can be ranked by stochastic dominance.

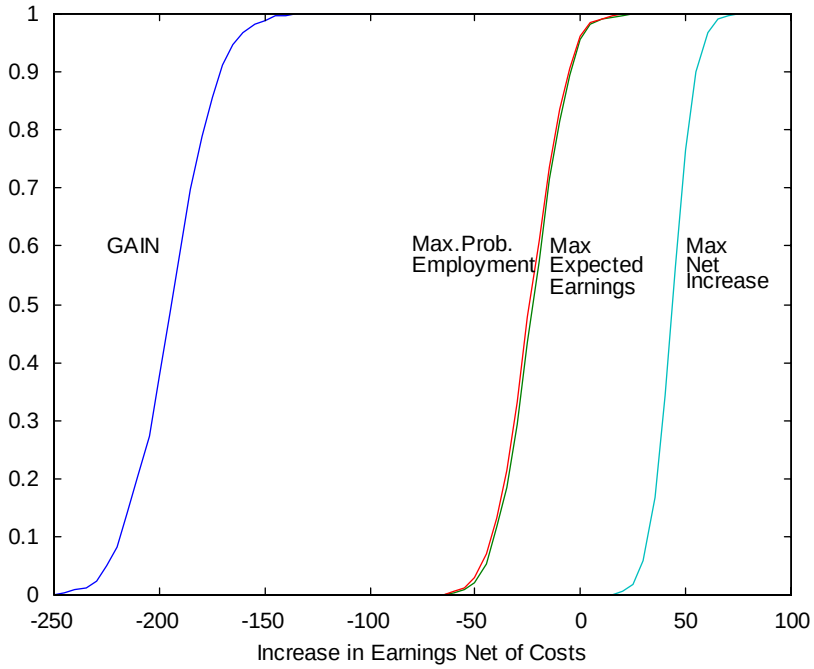

Fig. 8. Predictive distributions for increased earnings net-of-costs.

Thus, combining the three columns of Table 8 suggests that policies which allow individualized assignment dominate the policy in which all individuals are assigned to GAIN: the former policies are cheaper and result in higher average earnings per person. These policies also dominate AFDC in terms of average earnings. The only policy that dominates AFDC in terms of increased earnings net-of-cost is explicitly assigning to treatment only those individuals for whom increased earnings exceed training costs. We conclude that allowing a caseworker to assign individuals into GAIN and AFDC makes both individuals and the social planner better off; thus, we reach a positive assessment of the treatment. In contrast, ignoring the possibility of individual assignment, one would conclude that GAIN has a mixed and limited impact on individual earnings, with its benefits more than offset by the increased costs of the program.

Another set of concerns for the policymaker is the distribution of the benefits from GAIN. Presumably not all forms of inequality are of concern to the policymaker. Indeed, an increase in the upper percentiles of the earnings distribution of GAIN relative to AFDC would be one of the aims of training. However, if GAIN were to reduce earnings in the lower percentiles, then this might be a source of concern. Table 9 presents percentiles of the predictive distributions of earnings in each program (averaged over the 13 quarters and 1360 individuals). The 5th percentile for each of the policies is zero. From the 25th to the 50th percentiles AFDC overtakes GAIN, and from the 50th to the 90th percentiles GAIN overtakes AFDC. This is depicted in Fig. 9, which shows the differences in the percentiles of earnings between GAIN and AFDC. The figure reveals that AFDC once again overtakes GAIN for very high percentiles of earnings. (This fits into the pattern of Ms. Eight Twenty-Two in Table 5b, who had very high pre-assignment earnings, but fared poorly in the treatment.) For individual assignment to treatment, each of the percentiles exceeds the corresponding percentile for GAIN or AFDC. The final column in Table 9 presents a more synoptic view of inequality by examining the 90-10 difference for each program. The 90-10 spread increases from \$361 for ADFC, to \$453 for GAIN, to \$494 for the individual assignment policies.[21]

In Table 10, we explicitly examine the role of the policymaker's attitude toward ex ante inequality by applying a range of SWFs to the predictive distribution of post-assignment earnings. The table reveals that for a sufficient degree of inequality aversion (either exponential or Rawlsian) the ranking between GAIN and AFDC is reversed, with AFDC preferred. However, even for extreme inequality-aversion, the policy of individual assignment is preferred to either GAIN or AFDC. The lower two panels of the table demonstrate that this conclusion does not depend on the individual preferences used to compute the certainty equivalents. As noted in Section 5, the social welfare rankings do not require standard errors or confidence intervals;

[21] Because the joint distribution of earnings is not identified, we cannot make claims about how particular individuals fare relative to the distribution in each program. See Heckman and Smith (1998) and Heckman et al. (1997) for an approach to doing this.

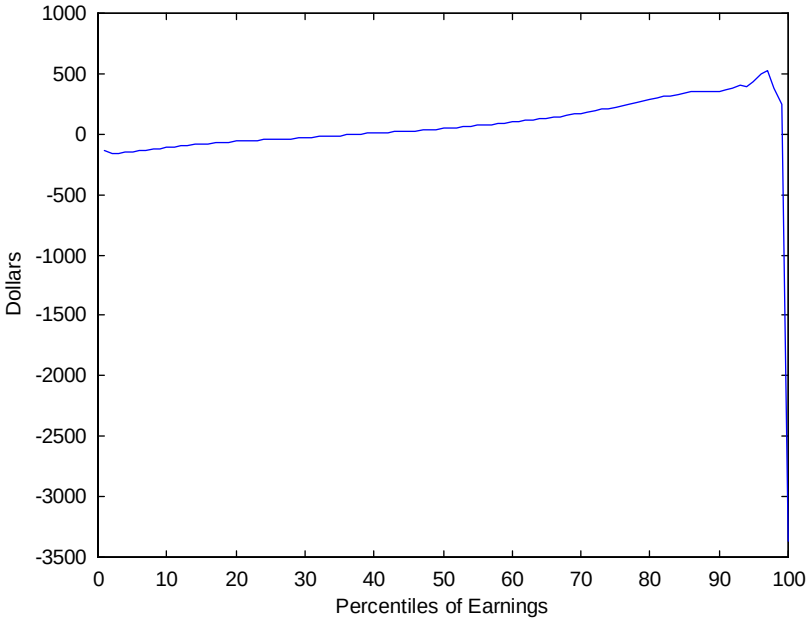

Fig. 9. Difference in percentiles of the earnings distribution GAIN–AFDC.

uncertainty is already accounted for, since the table is based on certainty equivalents of the earnings distribution.

The overall picture that emerges from Tables 8–10 is that GAIN is strongly preferred to AFDC in terms of earnings and the probability of employment, but not in terms of earnings net-of-costs or in the presence of a sufficient degree of inequality aversion. The policy of individualized assignment emerges as superior to both GAIN and AFDC in terms of all the criteria considered. Depending on the assignment rule, this policy can also lead to a net gain in terms of increased earnings net-of-increased-costs.

7. Conclusion

This paper examines the implications of shifting the emphasis in program evaluation from examining average treatment effects and their statistical significance to looking at the underlying decision problem. There are several important differences that emerge.

First, by considering the decision problem at the individual level (solved by the caseworker), we can expose the heterogeneity in the treatment impact, and produce a profile of the winners and losers from GAIN. Second, by embedding individual assignment within the policymaker's decision problem, we allow the policymaker to

consider not only the usual policies of assigning all individuals to GAIN or AFDC, but also policies that assign individuals based on a range of criteria such as maximizing earnings or the probability of employment. Third, by considering the full predictive distribution of the evaluation criteria (such as average earnings or increased earnings net-of-costs), we are able to account systematically for uncertainty in the policymaker's decision. The question of whether the difference between two programs is significant now reduces to asking what decision a policymaker would take. When the decision is invariant to the policymaker's preferences (as was the case in the GAIN example), an economically "significant" ranking emerges. Finally, by converting uncertainty regarding individual earnings into certainty equivalents, we are able to examine the importance of inequality aversion in ranking programs.

These results are certainly important for an analysis of the GAIN data, but they are also relevant to other exercises in program evaluation. For any program in which there is heterogeneity in the treatment impact, there is potentially a role for individualized assignment into treatment. This is especially true for programs in which the gains from the treatment do not exceed the costs for some individuals. Also, the importance of comparing outcomes under different programs using their predictive distributions rather than simply the first moments of their empirical distributions applies quite broadly to other evaluations. The relevance of this framework extends beyond the case of randomized experiments considered here. In non-experimental settings, a similar methodology could be adopted, if an appropriate selection correction mechanism were adopted.

The model can be extended in a number of directions. First, in some policy contexts, there may exist substantial prior information regarding the control program. Such information could readily be incorporated into the priors of the model. Second, there is scope to add greater heterogeneity, perhaps by using a hierarchical model to incorporate many more interactions. Third, the model could be modified to forecast beyond the 13 quarters included in the dataset to extend the evaluation to longer horizons. Fourth, the framework of individual assignment by a caseworker could be extended to allow individuals to incorporate private information into their decisions; the policymaker then would not simply offer individuals a choice but would design incentive-compatible assignment mechanisms (see Dehejia, 1999a). These are subjects of ongoing research.

Acknowledgements

I am grateful to Gary Chamberlain, Edward Glaeser, Caroline Minter Hoxby, Guido Imbens, and Lawrence Katz for their support and encouragement; to an Associate Editor and three anonymous referees, Joshua Angrist, Richard Blundell, and Jeffrey Smith for detailed suggestions; to Gordon Anderson, Vivek Dehejia, Roberta Gatti, James Heckman, Kei Hirano, Jeffrey Liebman, Emily Mechner, Carl

Morris, Dale Poirier, Donald Rubin, and Amartya Sen for their invaluable input; and to seminar participants at the Columbia University, the Canadian Econometrics Study Group, the European Econometric Society Meetings, Harvard University, Johns Hopkins University, Mathematica Policy Research, McGill University, the NSF-NBER Bayesian Meetings, Ohio State University, Penn State University, Universitat Pompeu Fabra, University College London, the University of Toronto, and the ZEW Research Conference on Econometric Evaluation of Active Labor Market Policies in Europe for comments. I owe a special debt to Sadek Wahba for bringing the GAIN data to my attention and for many conversations during our ongoing collaboration, and to the Manpower Development Research Corporation (MDRC) for facilitating and permitting the use of these data. Responsibility for any remaining errors and omissions is my own.

Appendix A

A.1. The Tobit Model

The likelihood for the Tobit model given Eqs. (2) and (3) is

$$L(\beta,\sigma)=\prod_{i\in C}\left[1-\Phi(x_i\beta/\sigma)\right](2\pi)^{-n_1/2}\sigma^{-n_1}\exp\left[-\tfrac{1}{2\sigma^2}(y_1-X_1\beta)^2\right],$$

where $C=\{j,\ t|y_{jt}=0\}$, the elements of C are indexed by j, Φ is the standard normal c.d.f., and y_1 denotes the vector of non-zero observations and X_1 the corresponding covariates. See Chib (1992) for further details.

A.2. The estimation procedure

The posterior distribution of the parameters of the Tobit model is obtained through a Gibbs sampling procedure. The Gibbs sampler is a Markov chain Monte Carlo simulation technique that 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 set of starting values (given certain conditions), these draws will eventually converge to draws from the true posterior (see Chib and Greenberg, 1996; Gelman et al., 1996; Geman and Geman, 1984; Gelfland and Smith, 1990; Geweke, 1997; Tanner and Wong, 1987).

In many cases, such as the Tobit, the task of drawing from the joint posterior is simplified by augmenting the parameter space of the model. For the Tobit model, the parameter space is expanded to include the latent variables y^*_{it}; conditional on this variable, the Tobit 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}.

The Gibbs sampling algorithm for the Tobit model has been worked out by Chib (1992) (see also Albert and Chib, 1993). The Gibbs sampling scheme is:

(1) Let y_{it}^z equal y_{it} for the uncensored observations, i.e., $\{i,t|y_{it}>0\}$, and for the censored observations, i.e., $\{i,t|y_{it}=0\}$, draw for y_{it}^z from the negative portion of a truncated normal distribution with mean $x_{it}\beta$ and variance σ^2.

(2) Draw for β from $\mathrm{N}\left(\hat{\beta}, \sigma^2(x'x)^{-1}\right)$, where $\hat{\beta}=(x'x)^{-1}x'y^z$ and $y^z=(y_{1,1}^z, y_{1,2}^z, \ldots, y_{1,13}^z, \ldots, y_{1360,1}^z, y_{1360,2}^z, \ldots, y_{1360,13}^z)'$. and $x=(x_{1,1}, \ldots, x_{1360,13})'$.

(3) Draw for σ^{-2} from Gamma $\left(8840, \|y^z-x\beta\|^2/2\right)$.

From an arbitrary starting value, this is iterated 5,000 times. The first 4000 iterations are discarded, leaving 1,000 draws from the posterior distribution of the parameters.[22]

A.3. Simulating the predictive distribution

For each individual i in the group of interest, we consider $x_{it,1}$ and $x_{it,0}$. The specification of $x_{it,1}$ ($x_{it,0}$) is identical to x_{it} above, except that we impose $T_i=1$ (0). We use our stored draws from the posterior distribution of the parameters, $\{\beta_{(j)}, \sigma^2_{(j)}\}_{j=1}^{1000}$. Given ($\beta_{(j)}$, $\sigma^2_{(j)}$), draw for $y_{it\ 1}^{*(j)} \sim \mathrm{N}(x_{it,1}\ \beta_{(j)}, \sigma^2_{(j)})$ and $y_{it\ 0}^{*\ (j)} \sim \mathrm{N}(x_{it,0}\ \beta_{(j)}, \sigma^2_{(j)})$. Finally, we obtain $y_{it\ 1}^{*(j)}$ and $y_{it\ 0}^{*\ (j)}$ by censoring $y_{it\ 1}^{*(j)}$ and $y_{it\ 0}^{*\ (j)}$ according to Eq. (2).

In Section 4.1, we compare the distributions $\left\{y_{it1}^{(j)}\right\}_{j=1}^{1000}$ and $\left\{y_{it\ 0}^{(j)}\right\}_{j=1}^{1000}$ for $i=822$, 1043.

In Sections 4.3 and 5, we produce predictive draws of earnings under treatment and control for the 1360 individuals in the sample. For each individual, we compute a range of assignment rules over the predictive treatment and control distributions. For example, in Table 8 (row 3), for each individual, we compute the probability of employment under treatment and control from the predictive distributions: $p_{it}=\sum_j 1(y_{it1}^{(j)}>0)/1000$, for $t=1,0$. The policy "maximize probability of employment" will assign individual i to treatment ($\mathrm{A}_i=1$) if $p_{i1}>p_{i0}$. Then along row 3 we compute the mean of the specified outcomes based on this assignment, for each draw from the predictive distribution. For example, for mean earnings: $\sum\left(A_i y_{it1}^{(j)}+(1-A_i)y_{it\ 0}^{(j)}\right)/1360$, for $j=1,\ldots,1000$.

Appendix B

Table 11 summarizes the posterior distributions of the parameters.

[22] Several diagnostics suggest that throwing out the first 4000 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.

Table 11
Mean and standard deviation of the posterior distribution of the tobit parameters

Variable	Mean of posteriod distribution (standard deviation)	Variable	Mean of posteriod distribution (standard deviation)	Variable	Mean of posteriod distribution (standard deviation)
Constant, $t = 1$	−4264 (323)	Treatment effect, $t = 13$	−6939 (2038)	Indicator for previous training or job search activities	1374.07 (115.50)
Treatment effect, $t = 1$	−8116 (2038)	Constant, $t= 11$	−4131 (633)	Ethnicity indicator, White	−161.96 (132.83)
Constant, $t = 2$	−4189 (348)	Treatment effect, $t = 11$	−7115 (2044)	Ethnicity indicator, Hispanic	−34.51 (182.11)
Treatment effect, $t = 2$	−7743 (2031)	Constant, $t = 12$	−4158 (667)	Earnings 10 quarters prior to experiment	0.39 (0.15)
Constant, $t = 3$	−4077 (381)	Treatment effect, $t = 12$	−7252 (2053)	Earnings 9 quarters prior to experiment	−0.65 (0.18)
Treatment effect, $t = 3$	−7581 (2047)	Constant, $t = 13$	−4361 (705)	Earnings 8 quarters prior to experiment	0.48 (0.19)
Constant, $t = 4$	−4137 (410)	Number of children less than age 4	−64.34 (115.58)	Earnings 7 quarters prior to experiment	−0.21 (0.14)
Treatment effect, $t = 4$	−7293 (2047)	Number of children between ages 4 and 5	−36.79 (108.68)	Earnings 6 quarters prior to experiment	−0.07 (0.14)
Constant, $t = 5$	−4284 (447)	Number of children between ages 6 and 11	5.00 (54.19)	Earnings 5 quarters prior to experiment	0.45 (0.14)
Treatment effect, $t = 5$	−7196 (2034)	Number of children between ages 12 and 18	161.89 (61.52)	Earnings 4 quarters prior to experiment	0.03 (0.16)
Constant, $t = 6$	−4257 (475)	Number of children aged 19 and greater	−12.37 (92.98)	Earnings 3 quarters prior to experiment	−0.50 (0.16)

Treatment effect, $t = 6$	−7386 (2058)	Score on reading test	−1.47 (3.64)	Treated · earnings 10 quarters prior to experiment	−0.24 (0.10)
Constant, $t = 7$	−4138 (504)	Score on mathematics test	0.67 (3.76)	Treated · earnings 9 quarters prior to experiment	0.50 (0.14)
Treatment effect, $t = 7$	−7388 (2052)	Grade	196.80 (24.39)	Treated · earnings 8 quarters prior to experiment	−0.25 (0.15)
Constant, $t = 8$	−4204 (536)	Most recently recorded hourly wage	141.78 (16.69)	Treated · earnings 7 quarters prior to experiment	0.20 (0.13)
Treatment effect, $t = 8$	−7237 (2047)	Indicator for households with single head	33.81 (50.19)	Treated · earnings 6 quarters prior to experiment	0.14 (0.10)
Constant, $t = 9$	−4220 (570)	Age	−35.78 (7.60)	Treated · earnings 5 quarters prior to experiment	−0.20 (0.10)
Treatment effect, $t = 9$	−6978 (2059)	Indicator for female participants	659.43 (237.47)	Treated · earnings 4 quarters prior to experiment	0.38 (0.09)
Constant, $t = 10$	−4201 (594)	Indicator of refugee status	1125.00 (267.01)	Treated · earnings 3 quarters prior to experiment	0.89 (0.11)
Treatment effect, $t = 10$	−7013 (2050)	Indicator for receiving AFDC in pre-assignment time	4125.59 (1145.99)	Treated · time trend	21.62 (11.49)

The mean of the posterior of σ is 3259 (with a standard deviation of 44).

References

Albert, J., Chib, S., 1993. Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association 88, 669–679.

Ashenfelter, O., 1978. Estimating the effects of training programs on earnings. Review of Economics and Statistics 60, 47–57.

Ashenfelter, O., Card, D., 1985. Using the longitudinal structure of earnings to estimate the effect of training programs. Review of Economics and Statistics 67, 648–660.

Barberis, N., 2000. Investing for the long run when returns are predictable. Journal of Finance 55, 225–264.

Berger, M., Black, D., Smith, J., 2001. Evaluating profiling as a means of allocating government services. In: Lechner, M., Pfeiffer, F. (Eds.), Econometric Evaluation of Labour Market Policies, Physica Heidelberg, pp. 59–84.

Black, D., Berger, M., Smith, J., Noel, B., 1999. Is the threat of training more effective than training itself? Experimental evidence from UI claimant profiling. Mimeo, University of Western Ontario.

Burtless, G., 1995. The case for randomized field trials in economic and policy research. Journal of Economic Perspectives 9, 63–84.

Chamberlain, G., 2000. Econometrics and decision theory. Journal of Econometrics 95, 255–283.

Chib, S., 1992. Bayes inference in the Tobit censored regression model. Journal of Econometrics 51, 79–99.

Chib, S., Greenberg, E., 1996. Markov chain Monte Carlo simulation methods in econometrics. Econometric Theory 12, 409–431.

Deaton, A., Muellbauer, J., 1980. Economics and Consumer Behavior. Cambridge University Press, Cambridge.

Dehejia, R., 1997. A decision—theoretic approach to program evaluation. Ph.D. Dissertation, Harvard University, Chapter 2.

Dehejia, R., 1999a. Effort, incentives, and choice in program evaluation. Mimeo, Columbia University.

Dehejia, R., 1999b. Program evaluation as a decision problem. National Bureau of Economic Research Working Paper No. 6954.

Dehejia, R., 2003. Evaluating multi-site programs. Journal of Business and Economic Statistics 21, 1–11.

Dehejia, R., Wahba, S., 2002. Propensity score matching methods for non-experimental causal studies. Review of Economics and Statistics 84, 151–161.

Dehejia, R., Wahba, S., 1999. Causal effects in non-experimental studies: reevaluating the evaluation of training programs. Journal of the American Statistical Association 94, 1053–1062.

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

Fisher, R., 1935. The Design of Experiments. Oliver and Boyd, London.

Friend, B., Blume, M., 1975. The demand for risky assets. American Economic Review 65, 900–922.

Gelfland, A., Smith, A., 1990. Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association 85, 398–409.

Gelman, A., Carlin, J., Stern, H., Rubin, D., 1996. Bayesian Data Analysis. Chapman and Hall, London.

Geman, S., Geman, D., 1984. Stochastic relaxation, gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence 6, 721–741.

Geweke, J., 1997. Posterior simulators in econometrics. In: Kreps, D., Wallis, K. (Eds.), Advances in Economics and Econometrics: Theory and Applications, Vol. III. Cambridge University Press, Cambridge, pp. 128–165.

Greenberg, D., 1997. The leisure bias in cost-benefit analyses of employment and training programs. Journal of Human Resources 32, 413–439.

Greenberg, D., Wiseman, M., 1992. What did the OBRA demonstrations do? In: Manski, C., Garfinkel, I. (Eds.), Evaluating Welfare and Training Programs. Harvard University Press, Cambridge, MA.

Heckman, J., 1990. Varieties of selection bias. American Economic Review 80, 313–318.

Heckman, J., 1992. Evaluating welfare and training programs, In: Manski and Garfinkel (1992).

Heckman, J., Hotz, J., 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, 862–874.

Heckman, J., Ichimura, H., Smith, J., Todd, P., 1998. Characterizing selection bias using experimental data. Econometrica 66, 1017–1098.

Heckman, J., Robb, R., 1985. Alternative methods for evaluating the impact of interventions. In: Heckman, J., Singer, B. (Eds.), Longitudinal Analysis of Labor Market Data. Econometric Society Monograph No. 10. Cambridge University Press, Cambridge.

Heckman, J., Robb, R., 1986. Alternative methods for solving the problem of selection bias in evaluating the impact of treatment on outcomes. In: Rainer, H. (Ed.), Drawing Inferences from Self-Selected Samples. Springer, New York.

Heckman, J., Smith, J., 1995. Assessing the case for social experiments. Journal of Economic Perspectives 9, 85–110.

Heckman, J., Smith, J., 1998. Evaluating the welfare state. In: Strom, S. (Ed.), Econometrics in the 20th Century: The Ragnar Frisch Centenary. Cambridge University Press for the Econometric Society Monograph Series, Cambridge.

Heckman, J., Smith, J., Clements, N., 1997. Making the most out of programme evaluations and social experiments: accounting for heterogeneity in programme impacts. Review of Economic Studies 64, 487–536.

Hirano, K., 2000. Semiparametric Bayesian inference in autoregressive panel data models Mimeo. University of California, Los Angeles.

Hotz, J., Imbens, G., Klerman, J., 2000. The long-term gains from GAIN: a re-analysis of the impacts of the California GAIN program Mimeo, University of California, Los Angeles.

Hotz, J., Imbens, G., Mortimer, J., 1999. Predicting the efficacy of future training programs using past experiences. National Bureau of Economic Research Working Paper No. T238.

Kandel, S., Stambaugh, R., 1996. On the predictability of stock returns: an asset-allocation perspective. Journal of Finance 51, 385–424.

Lalonde, R., 1986. Evaluating the econometric evaluations of training programs. American Economic Review 76, 604–620.

Manski, C., 1989. Anatomy of the selection problem. Journal of Human Resources 24, 343–360.

Manski, C., 1993. The selection problem. In: Sims, C. (Ed.), Advances in Econometrics. Cambridge University Press, Cambridge.

Manski, C., 1995. the mixing problem in program evaluation. In: Manski, C. (Ed.), Identification Problems in the Social Sciences. Harvard University Press, Cambridge Chapter 3.

Manski, C., 1999. Statistical treatment rules for heterogeneous populations: with applications to randomized experiments. Mimeo, Northwestern University.

Manski, C., 2000. Identification problems and decisions under ambiguity: empirical analysis of treatment response and normative analysis of treatment choice. Journal of Econometrics 95, 415–442.

Manski, C., Garfinkel, I., 1992. Evaluating Welfare and Training Programs. Harvard University Press, Cambridge.

Neyman, J., Iwaszkiewicz, K., Kolodziejczyk, S., 1935. Statistical problems in agricultural experimentation (with discussion). Supplement of Journal of the Royal Statistical Society 2, 107–180.

Runner, D., 1996. Changes in state unemployment insurance legislation in 1995. Monthly Labor Review 119, 73–78.

Riccio, J., Friedlander, D., Freedman, S., 1994. GAIN: Benefits, Costs, and Three-Year Impacts of a Welfare-to-Work Program. Manpower Demonstration Research Corporation, New York.

Tanner, M., Wong, W., 1987. The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association 82, 528–550.