SRDC Working Paper Series 03-05

Assessing the Impact of Non-response on the Treatment Effect in the Canadian Self-Sufficiency Project

Thierry Kamionka

Centre national de la recherche scientifique (CNRS),
Critical Research on Economic and Social Transformation (CREST), and
the Institut d'Economie Industrielle (IDEI)

Guy Lacroix

Département d'économique, Université Laval,
Centre interuniversitaire sur le risque, les politiques économiques et l'emploi (CIRPÉE) and
the Center for Interuniversity Research and Analysis on Organizations (CIRANO)

October 2003

SOCIAL RESEARCH AND DEMONSTRATION CORPORATION

The Social Research and Demonstration Corporation is a non-profit organization and registered charity with offices in Ottawa, Vancouver, and Sydney, Nova Scotia. SRDC was created specifically to develop, field test, and rigorously evaluate social programs. SRDC's two-part mission is to help policy-makers and practitioners identify social policies and programs that improve the well-being of all Canadians, with a special concern for the effects on the disadvantaged, and to raise the standards of evidence that are used in assessing social policies. As an intermediary organization, SRDC attempts to bridge the worlds of academic researchers, government policy-makers, and on-the-ground program operators. Providing a vehicle for the development and management of complex demonstration projects, SRDC seeks to work in close partnership with provinces, the federal government, local programs, and private philanthropies.

Contents

Tables and Figures

Acknowledgements

The authors gratefully acknowledge financial support from le Fonds de recherche sur la société et la culture du Québec (FRSCQ) and the Social Research Demonstration Corporation (SRDC), and the excellent comments of the two anonymous referees.

Abstract

In Canada, a policy to help single parents on social assistance become self-reliant was implemented on an experimental basis. The Self-Sufficiency Project's "applicant study" randomly selected a sample of 4,134 single parents who had applied for welfare between January 1994 and March 1995. Only 3,315 took part in the experiment despite a 50 per cent chance of receiving a generous, time-limited, earnings supplement conditional on finding a full-time job and leaving income assistance within a year.

This paper determines whether a non-response rate as high as 20 per cent is likely to bias the measurement of the treatment effect. Using experimental data only, we compare the estimated impact of the program to those results obtained using additional data on individuals not taking part in the experiment. We write the likelihood function based on different sets of information concerning the sample and obtain relevant estimates of program impact on welfare spell durations. We find strong evidence of non-response bias in the data. When we correct for the bias, we find the estimates that rely on experimental data only significantly underestimate the true impact of the program.

Introduction

In seeking to alleviate the problems that plague particularly disadvantaged groups when integrating into the labour market, governments have traditionally turned to skill-enhancing training programs. By enhancing skills, it is hoped individuals will receive attractive job offers and thus reduce their reliance on transfer programs.

Over the past 20 years the evaluation literature has generally found training programs to have had limited success in achieving these goals. (See Heckman, Lalonde, & Smith, 1999, for a recent and detailed survey and Gilbert, Kamionka, & Lacroix, 2001, for results pertaining to Canada.) Indeed, only very focused programs targeted at specific groups seem to have had any significant impact on reliance on support programs. Yet, any decrease in reliance has not generally translated into significant reductions in poverty rates. It may be inferred from such poor performance that the training programs implemented over that period did not increase productivity to a level that would make work a better alternative to social assistance.

Many governments have responded to such disappointing results by shying away from traditional training-only programs to contemplate policies that directly address the relative attractiveness of work. By directly subsidizing wage rates, it is believed many will be induced to accept jobs offers that would not normally be good alternatives to transfer programs such as social assistance. Inducing individuals to work is motivated by two separate but complementary goals. First, by raising total income such policies may be more effective at addressing poverty than traditional programs. Second, holding a regular job may be more conducive to the acquisition of the skills and attitudes necessary for self-reliance.

In Canada a policy to help single parents on social assistance become self-reliant was implemented on an experimental basis. The Self-Sufficiency Project (SSP) is a research and demonstration project that provides a generous, time-limited, earnings supplement to welfare recipients who find full-time jobs and leave income assistance (IA). SSP consists of two main studies: the SSP “recipient study” and the SSP “applicant study.” The former focuses on welfare recipients who have been on welfare for at least a year. The latter focuses on newly enrolled recipients.

The recipient study began in 1992 and enrolled over 5,600 volunteers. About half were randomly offered SSP. The other half — the experimental control group — were not offered the supplement. The applicant study, on the other hand, aimed at documenting so-called delayed exit effects. Since new entrants had to stay on welfare for at least 12 months to qualify for SSP, it was feared the supplement would entice some to remain on welfare longer. The applicant study randomly selected a sample of single parents who had applied for welfare between January 1994 and March 1995. Half of those selected were offered the supplement. Most evaluations of SSP are based on the recipient study. Nearly all of them concluded that the program had sizable impacts on exits from welfare (Michalopoulos, Card, Gennetian, Harknet, & Robins, 2000). Others found the program beneficial to children (Morris & Michalopoulos, 2000) but that it had ambiguous results on marital behaviour (Harknett & Gennetian, 2001).

There is little doubt the program had significant impacts on individual behaviour. Because both the recipient study and the applicant study use classical random assignment designs, estimates of program impacts rest on simple comparisons between mean responses of treatment and control groups. Such comparisons provide appropriate estimates of "treatment effects on the treated" only under a number of relatively stringent assumptions. One states that individuals taking part in the experiment constitute a true random sample of the population of interest. There is little discussion of experimental biases in the literature partly because the data obtained from social experiments simply cannot confirm or deny that behaviour has been disrupted in one way or another. The evidence brought to bear is almost always indirect or inferential at best.[1] It is thus important to determine whether behaviour has indeed been affected by the experimentation and, if so, whether behavioural disruptions have contaminated the estimated impacts.

This paper documents the extent of non-response bias in the SSP experiment and proposes a measure of the impact of such bias, if any. Our analysis focuses on the applicant study, because the non-response rate was much higher than in the recipient study (20 vs. 5 per cent).[2] Our strategy is to compare the estimated impact of the program, using experimental data only, to those results obtained using additional data on individuals not taking part in the experiment. Reasons for not participating are threefold. First, some recipients were not selected at baseline. This sample can be thought of as a legitimate control group. Second, some were selected but refused to participate. Finally, some were selected but could not be reached at baseline. Since we know the probability of being in each sample, we can write the likelihood of various sets of information and obtain relevant estimates of program impact on welfare spell durations. Our results are consistent with those of Bancroft, Card, Lin, and Robins (1998) in finding little evidence of delayed exits, if any. Furthermore, we find strong evidence of non-response bias in the data. When we properly correct for the bias, we find the estimates that rely on experimental data alone underestimate the true impact of the program. The remainder of the paper is organized as follows. The next section provides a detailed description of the applicant study, including data on both participants and non-participants in the applicant study. Non-parametric evidence on delayed exits is presented as well. The third section discusses the statistical model and the treatment of unobserved individual heterogeneity. The fourth section reports our main findings. Finally, the last section concludes the paper.

[1] See Heckman (1992) for a discussion of randomization biases.

[2] As many as 4,134 individuals were contacted for the applicant study. Yet, only 3,326 completed the baseline survey, and an additional nine asked to be removed from the experiment after completing the survey. Thus, the response rate is about 80 per cent.

The Applicant Study

Economists have long recognized that policies providing for a conditional earnings supplement may have the unintended consequence of inducing some to modify their behaviour to become eligible. There is very little empirical evidence to support this claim. Most studies that focus on so-called entry effects are based on simulation models (Moffitt, 1992, 1996) that have been shown to perform relatively well at predicting inflows and outflows from welfare caseloads (Garasky & Barnow, 1992).

The Self-Sufficiency Project (SSP) was introduced in Canada in 1992 to measure the response of long-term welfare recipients to a financial incentive that made work pay better than welfare. SSP offered a generous, time-limited, monthly cash payment to eligible single parents in British Columbia and New Brunswick who found full-time jobs and left welfare. The supplement was available only to those who had remained on welfare for at least 12 months. This feature of the program and the (relative) generosity of the supplement were thought to give rise to two types of entry effects. The first, unconditional effect, is to induce single parents to join the welfare rolls and become eligible. The second, conditional effect, is to induce those on the rolls to delay their exit from welfare to become eligible.

Designing an experiment to measure unconditional entry effects is not feasible since it would require a very large sample and involve huge implementation costs. On the other hand, measuring delayed exit behaviour through a social experiment is much more feasible. The applicant study used a random sample of single parents who had applied for and received income assistance (IA) between January 1994 and March 1995 in British Columbia.[3] Selected individuals who agreed to be part of the experiment were interviewed at home to complete the baseline survey. They were also asked to sign an informed consent that explained the nature of the experiment, described the random assignment process, and stated that all individual-level data would be kept confidential. The agreement also gave researchers access to IA administrative records from the British Columbia Ministry of Social Services. Immediately after the baseline interview, individuals were randomly assigned to either the program or the control group. Program members were sent a letter and brochure explaining their potential eligibility for an earnings supplement. They were reminded they had to remain on welfare for at least 12 months to qualify for the supplement and that on qualification, they had to find a full-time job within the next 12 months. They were also mailed a reminder notice six to seven months after their baseline interview.

DATA

As mentioned earlier, our empirical strategy consists of using information on individuals who were not in the experiment to assess the existence of non-response bias. Statistics Canada, the data collection contractor, agreed to provide individual IA histories on participants and non-participants alike using administrative files.

[3]To be considered as new entrants, applicants could not have received IA in the six previous months. A significant minority (31 per cent) had received IA at some time in the two years before their current application (Berlin, Bancroft, Card, Lin, & Robins, 1998).

The original sample was fielded between January 1994 and March 1995. Each month, an independent random sample from the population of welfare applicants was selected. To be included in the experimental sample, individuals could not have received welfare payments for at least six months before applying for benefits. Statistics Canada used the same algorithm to generate the sample of non-participants.[4] For confidentiality reasons, the data were restricted in two ways. First, only information on the first welfare spell was made available. Second, those who refused to take part in the experiment were included in the population not sampled at baseline.[5]

The sampling scheme and the data at our disposal are illustrated in Figure 1. The original sample comprised over 4,337 individuals. Of those, 139 were declared out-of-scope (i.e. they were sampled by mistake), 56 were eventually excluded for the same reason, and an additional 8 asked to be removed from the study. This left 4,134 individuals. Of these, 3,315 agreed to sign the informed consent and complete the baseline survey. The response rate is thus approximately 80 per cent. Of the original sample, 694 individuals could either not be contacted at baseline (307) or were not followed up (387). We refer to this group as sample C.[6] Finally, 122 individuals refused to take part in the experiment.[7] The randomization procedure yielded the experimental treatment and control groups, samples A and B, respectively.

Statistics Canada provided a sample of 3,073 individuals who were not contacted at baseline or refused to be in the experiment. We refer to this group as sample D.[8] Those who refused are not identifiable in the data. As such, sample D is a complex mix of groups A, B, and C. Indeed, among those in D, some would have joined the experiment $(A + B)$ had they been selected, others would not have been contacted for different reasons (C), and still others would have refused to take part in the experiment. Under the null assumption that the data is void of non-response bias, groups B and D should behave in a similar manner. If systematic differences are found, it will be necessary to investigate whether the treatment effect is biased.

DESCRIPTIVE STATISTICS

Table 1 provides descriptive statistics for each sample separately.[9] The first two columns show that the experimental treatment and control groups are very similar in terms of

[4] Randomization occurred during the first month following application for benefits in most cases. As many as 2,464 individuals had either received no or one IA payment at randomization. Another 653 individuals had received two monthly payments. Finally, 92 individuals had received as many as three or four payments before assignment. We use the randomization date as the starting date for the experimental sample, since this corresponds to the beginning of the treatment. We acknowledge, though, that this tends to decrease the average duration of the experimental sample.

[5] Statistics Canada estimates that eight per cent of the original sample either refused to sign the informed consent, asked to be removed from the project, or did not agree to have their data included in any part of the study. These observations are included in the population not sampled at baseline.

[6] Although Statistics Canada documents show that 694 individuals were not contacted or followed up at baseline, the sample we were provided contains only 637 observations. Further, we have no information on individual status in the sample.

[7] It is very likely those who were not followed up also refused to take part in the experiment.

[8] The total population of welfare applicants over the period covered by the applicant study is 7,390. Thus, samples A, B, C, and D represent over 95 per cent of the total population.

[9] The administrative files contain more information on individual characteristics than those reported in Table 1. To ensure confidentiality, we were only provided with information on characteristics reported in the table.

observable characteristics. This is not surprising, since treatment is randomly assigned among those who agreed to take part in the experiment. Individuals in sample D are also very similar to those of samples A and B. On the other hand, sample C stands out as containing proportionately more men, and slightly younger individuals with fewer children. Although not reported in the table, women in sample C are somewhat younger than those of other samples, whereas the converse holds for men. In all samples, male-headed households have significantly fewer children than female-headed households.

Table 1 indicates the mean IA spell duration is relatively similar for individuals in samples A, B, and D. Those in sample C have significantly shorter mean and median durations. Finally, although we observed individual IA histories for over 65 months, more than 9.6 per cent of all spells are censored.

To better ascertain the extent to which observable characteristics differ among samples A, B, C, and D, we report simple logit regressions of belonging to a given sample in Table 2. For example, column (1) reports the parameter estimates of the probability of belonging to sample A when samples A and B are pooled together. As expected, all parameter estimates turn out not to be statistically significant. Likewise, columns (2) and (3) show that samples A, B, and D are very homogeneous. Indeed, only the intercepts are statistically significant in both regressions. The intercepts reflect the relative weight of the samples in the regression. On the other hand, sample C appears to be quite different from the other samples. Column (4) indicates that women are less likely to belong to sample C, as are households with more children, as well as those with older heads.[10]

NON-PARAMETRIC EVIDENCE

The applicant study aimed to determine whether IA applicants might be induced to delay their exit from welfare to qualify for the (relatively) generous earnings supplement. To qualify for the supplement, IA recipients had to stay on welfare for at least 12 months. Once qualified, those in sample A had to find a full-time job within 12 months to receive the supplement. Those in sample B continued to receive the standard IA benefit.

Behavioural response to the applicant study is best investigated through the use of hazard and survival functions.[11] Figure 2 plots smoothed hazard rates of IA spells for experimental samples A and B.[12] The first noteworthy feature of the figure is that the treatment sample appears to be sensitive to the parameters of the applicant study. Indeed, the hazard rates increase in the first eight months for both groups on entry into IA. The hazard rates of the treatment group keep increasing up until the 25th month, while those of the control decrease steadily.[13]

[10] We did not report the results using samples A, B, and C for the sake of brevity. They are very similar to those reported in column (4) of Table 2.

[11] Only brief non-parametric evidence on non-response bias in the applicant study is presented here. More extensive analyses using non-parametric permutation tests can be found in Lacroix & Royer (2001).

[12] Recall that approximately 20 per cent of the sample had been on welfare for at least two months before randomization. If we use the first month on IA instead of the randomization date as the start of the spell, the figure is basically unchanged. We use the Epanechnikov kernel with optimal bandwidth to draw the hazard functions.

[13] The rise in the hazard rates in the first few months has been observed in many studies using Canadian data. See for instance Drolet, Fortin, & Lacroix (2002) and Fougère, Fortin, & Lacroix (2002).

Weak delayed exit behaviour is evidenced by the difference between the hazard functions during the first seven months. Indeed, the hazard function of sample A lies below that of sample B during the first seven months, then crosses it and remains above it for the next 30 months or so. The underlying survival functions are plotted in Figure 3. Not surprisingly, the survival function of sample A lies above that of sample B up until Month 16. This is consistent with Michalopoulos and Hoy (2001) who found the individuals in sample A were proportionately more likely to receive IA than those in sample B up until the fifth quarter of the experiment. Based on Figure 3, it seems reasonable to claim that the earnings supplement first induces individuals to delay their exits in the beginning months and then provides a relatively strong incentive to leave IA. It is worth investigating whether these differences are statistically significant. Figure 4 plots the confidence intervals of the two survival curves. The confidence intervals of both survival functions overlap for the first 24 months. Thus, delayed exit from welfare, although evidenced from the survival functions, seems to lack statistical support. This can be formally tested by means of a simple non-parametric test. Indeed, it can be shown that the estimated mean duration over the interval $[0, \tau]$ is[14]

$$\hat{\mu}_\tau = \int_0^\tau \hat{S}(t)dt, \tag{1}$$

where $\hat{S}(t)$ is the estimated survival rate at time t. The variance of this estimator is

$$\hat{V}\left[\hat{\mu}_\tau\right] = \sum_{i=1}^{T} \left[\int_{t_i}^{\tau} \hat{S}(t)dt\right]^2 \frac{n_i}{Y_i(Y_i - n_i)}, \tag{2}$$

where T is the number of distinct discrete intervals over $[0, \tau]$, n_i is the number of individuals who leave welfare at time t_i, and Y_i is the number of individuals at risk of leaving welfare at time t_i. The mean duration of samples A and B over the first 12 months are found to be 8.69 and 8.48, respectively, a difference approximately equal to 2.5 per cent in favour of sample B. A simple $\chi^2(1)$ test cannot reject the null assumption that both durations are equal. This finding is similar to that of Berlin et al. (1998) who report an average impact of approximately three per cent. On the other hand, mean durations computed over $[0, 65]$ equal 20.3 and 21.8, respectively. This time, the $\chi^2(1)$ test ($= 4.38$) rejects the null assumption that mean durations are equal.

One could conclude that the treatment reduces mean duration by approximately 7.4 per cent. Even though such an estimate does not account for individual characteristics, it is very unlikely the program impact will be affected by such variables given the results of Table 2. The more interesting question that must be addressed is whether our estimates are plagued with non-response biases. Before we address this question formally, we present informal evidence that such biases may be present in the data.

Figure 5 plots the survival functions of samples B, C, and D. Notice that the survival function of group D lies everywhere below that of group B. Standard Log-rank and Wilcoxon tests strongly reject equality of the two curves. Hence, individuals in sample B have longer spells than those in sample D. In the absence of a non-response bias, sample D would

[14] See Klein & Moeschberger (1997) for a formal derivation.

normally constitute a proper control group since the two differ only in that the individuals in the former D were not sampled while those in the latter B were sampled and agreed to participate in the experiment. Yet, the difference between D and B may be partly explained by the fact that sample D includes individuals with unusually short spells that are excluded from B. Those are individuals who could not be contacted when this group was sampled. They probably share similar characteristics with, and behave similarly to, those in sample C. Incidentally, the survival function of sample C lies well below that of sample D. Yet, according to the figure, as many as a third would have qualified for the earnings supplement had they been contacted at baseline, notwithstanding potential delayed exit effects.

The above discussion indicates that the experimental control group likely suffers from non-response bias. It does not necessarily follow that the comparison between samples A and B yields a biased estimator of the treatment effect. Indeed, sample A may just as well be plagued with similar non-response bias that increases mean durations in the same proportion as that of sample B. To measure the program impact correctly, non-response must be modelled explicitly and accounted for in a regression framework.

Modelling Individual Spell Durations

To derive an appropriate estimator of the treatment effect, the non-response bias must be explicitly taken into account. The framework within which the experiment took place is illustrated in Figure 6, which depicts a hypothetical sample of individuals drawn from the flow of welfare applicants. The inner circle is the set of those who are sampled with probability p at baseline. Those in the population not willing a priori to participate in such an experiment are located below the dashed line. Those who could not be contacted are located in the ellipse. Among the latter, an unknown fraction would agree to be part of the experiment (above the dashed line) and another unknown fraction would refuse (below the dashed line).

The treatment group is located inside the inner circle to the left of the vertical line. Members of this group have all agreed to participate (above the dashed line) and have been contacted (outside the ellipse). The control group is located inside the inner circle to the right of the vertical line. The space between the inner and outer circles is the set of applicants who were not selected at baseline. This set can be broken down into subsets similar to those of the experimental samples (e.g. acceptance, refusal, contacted, not contacted).

Our task is to model all the information available in Figure 6. To do this, we need to determine the probability of belonging to the experimental samples. The experimental samples comprise 3,315 individuals. According to Statistics Canada, these individuals represent 45 per cent of all claimants over the enrolment period.[15] If we consider those who could not be contacted as well as those who refused to participate in the experiment, then we can establish that the average probability of being sampled each month ranges between 60 and 65 per cent. Thus, each applicant faces a probability $p = 0.65$ of being sampled.[16]

To model individual contributions to the likelihood function, we need to define a number of dummy variables. Let

$$E = \begin{cases} 1, & \text{if the individual was sampled at baseline,} \\ 0, & \text{otherwise.} \end{cases}$$

$$A = \begin{cases} 1, & \text{if the individual is willing to participate in the experiment,} \\ 0, & \text{otherwise.} \end{cases}$$

$$R = \begin{cases} 1, & \text{if the individual could be contacted at baseline,} \\ 0, & \text{otherwise.} \end{cases}$$

$$T = \begin{cases} 1, & \text{if the individual belongs to the treatment group,} \\ 0, & \text{otherwise.} \end{cases}$$

[15] See Footnote 8.

[16] The indeterminacy of the probability of being sampled arises due to some confusion related to sample C. According to private communications with Statistics Canada, our sample C only includes individuals who could not be contacted at baseline. In such a case, the probability of being sampled is roughly equal to 65 per cent. If, on the other hand, the sample includes both those who could not be contacted *and* those who were not followed up, then the probability of being sampled is approximately equal to 60 per cent. The model was estimated with $p = 0.60$ and $p = 0.65$. The main results are very robust to the choice of p.

Finally, let y be a realization of the experiment

$$y = (e, a, r, t, u),$$

where u is the duration of a welfare spell.[17]

Which arguments of $y(\cdot)$ are observable depends on the set to which an individual belongs. Only T and U are observable for all individuals.[18] For those in A, we know they have been sampled in the experiment ($e = 1$), they have agreed to participate ($a = 1$), they could be contacted ($r = 1$), and they are eligible for the supplement ($t = 1$). Table 3 summarizes the realizations of the random variables according to group membership.

LIKELIHOOD FUNCTION

Each individual contributes a sequence $y = (e, a, r, t, u)$ to the likelihood function. The contribution can be written conditionally on a vector of exogenous variables, x, and on an unobserved heterogeneity factor, ν. To simplify the presentation, we assume the components of y that are not observed are equal to -1.

Let $l_v(\theta)$ denote the conditional contribution of the realization y. We have

$$l_v(\theta) = f(y \mid x; \nu; \theta),$$

where $f(y \mid x; \nu; \theta)$ is the conditional density of y given x and ν, and $\theta \in \Theta \subset \mathbb{R}^p$ is a vector of parameters. When the welfare spell is right censored, the contribution to the conditional likelihood function is limited to the survivor function of the observed duration.

The random variable ν is assumed to be independently and identically distributed across individuals, and independent of x. If the unobserved heterogeneity only takes a finite number of values, $\nu_1, \ldots, \nu_J$, the contribution of a realization y to the likelihood function is

$$l(\theta) = \sum_{j=1}^{J} f(y \mid x; \nu_j; \theta)\ \pi_j, \tag{3}$$

where π_j is the probability that $\nu = \nu_j$ with $0 \leq \pi_j \leq 1$ and $\sum_{j=1}^{J} \pi_j = 1$.

If ν is a continuous random variable, then

$$l(\theta) = \int_S f(y \mid x; \nu; \theta)\ g(\nu; \gamma)\ d\,\nu, \tag{4}$$

where $g(\nu; \gamma)$ is a probability density function and S is the support of ν.

The conditional contribution of the realization $y = (e, a, r, t, u)$ to the likelihood function is written using the joint distribution of the components of y with the values of the realization fixed to those observed in the sample for a given individual.

[17] We follow the convention of denoting a random variable by a capital letter and write its realization in lower case.

[18] The welfare durations are right censored at 64 months.

MODELLING INDIVIDUAL CONTRIBUTIONS

We now focus on the conditional distributions of variables A, R, and U. Recall that the probability of being sampled in the experiment is p, and the probability of assignment to the treatment group conditional on acceptance and on being contacted is 0.5. We assume these two probabilities are independent of individual characteristics.

Define $z(x, \nu)$ as the conditional probability the individual agrees to participate in the experiment. We assume that

$$z(x, \nu) = \text{Prob}[A^* \geq 0 \mid x; \nu], \tag{5}$$

where

$$A^* = x' \, \beta_a + \nu + \epsilon_a,$$

where ϵ_a is a normal random variable with mean 0 and variance equal to 1, and is distributed independently of ν. In the model, ν is an unobserved heterogeneity term. In the participation equation, ν can be considered as an individual random effect.

Let $\phi(\nu, x, a)$ denote the conditional probability the individual cannot be contacted. We assume

$$\phi(x, \nu, a) = \text{Prob}[R^* \geq 0 \mid x; a; \nu], \tag{6}$$

where

$$R^* = x' \, \beta_r + a \, \xi_a + \nu + \epsilon_r,$$

where a is the realization of the participation decision, and β_r is a vector of parameters and $\xi_a \in I\!R$. We also assume ϵ_r is a normal random variable with mean 0 and variance equal to 1. For simplicity, we further assume ϵ_a, ϵ_r, and ν are independent.

Finally, let $q(e, a, r)$ denote the conditional probability the individual belongs to the treatment group given selection into the experiment ($e = 1$ or 0), given acceptance ($a = 1$ or 0), and given having been contacted ($r = 1$ or 0). Let us assume

$$\text{Prob}[T = 1 \mid e, a, r] = q(e, a, r) = \begin{cases} \frac{1}{2}, \text{ if } e = 1, a = 1 \text{ and } r = 1, \\ 0, \text{ otherwise.} \end{cases}$$

Hence, individuals can be assigned to the treatment group if and only if they have been sampled in the experiment, have agreed to participate, and could be contacted.

The conditional probability density function of the welfare duration is denoted $f(u \mid x; a; r; t; \nu; \theta)$, where θ is a vector of parameters. Therefore, the conditional contribution of a given realization to the likelihood function is

$$\ell_\nu(\theta) = p \; z(x, \nu) \; (1 - \phi(x, a, \nu)) \; 0.5 \; f(u \mid x; a = 1; r = 1; t = 1; \nu; \theta), \tag{7}$$

if the individual belongs to group A;

$$\ell_\nu(\theta) = p \; z(x, \nu) \; (1 - \phi(x, a, \nu)) \; 0.5 \; f(u \mid x; a = 1; r = 1; t = 0; \nu; \theta), \tag{8}$$

if the individual is in group B;

$$\begin{aligned}\ell_\nu(\theta) &= p\, z(x,\nu)\, \phi(x,a,\nu)\, f(u \mid x; a=1; r=0; t=0; \nu; \theta), \\ &+ p\, (1-z(x,\nu))\, \phi(x,a,\nu)\, f(u \mid x; a=0; r=0; t=0; \nu; \theta),\end{aligned} \tag{9}$$

if the individual is in group C;

and

$$\begin{aligned}\ell_\nu(\theta) &= p\, (1-z(x,\nu))\, (1-\phi(x,a,\nu))\, f(u \mid x; a=0; r=1; t=0; \nu; \theta), \\ &+ (1-p)\, z(x,\nu)\, (1-\phi(x,a,\nu))\, f(u \mid x; a=1; r=1; t=0; \nu; \theta), \\ &+ (1-p)\, z(x,\nu)\, \phi(x,a,\nu)\, f(u \mid x; a=1; r=0; t=0; \nu; \theta), \\ &+ (1-p)\, (1-z(x,\nu))\, (1-\phi(x,a,\nu))\, f(u \mid x; a=0; r=1; t=0; \nu; \theta), \\ &+ (1-p)\, (1-z(x,\nu))\, \phi(x,a,\nu)\, f(u \mid x; a=0; r=0; t=0; \nu; \theta),\end{aligned} \tag{10}$$

if the individual belongs to group D.[19]

The contribution of each group to the likelihood function is indicated in Figure 7. Thus groups A and B contribute sections 1 and 2 (equations (7) and (8), respectively). Likewise, group C (equation (9)) corresponds to sections 3 and 4. Group D (equation (10)) corresponds to sections 5, 6, 7, 8, and 9.

Let us consider a given individual. Let S_e denote the set of possible values of E:

$$S_e = \begin{cases} \{1\}, \text{ if the observed value } e=1, \\ \{0\}, \text{ if the observed value } e=0, \\ \{0,1\}, \text{ if } e \text{ is not observed (i.e. } e=-1), \end{cases}$$

Let S_a and S_r denote the sets of possible values of A and R. Both are defined in a similar fashion to S_e. Finally, the contribution to the likelihood function can be written[20]

$$\begin{aligned}\ell_\nu(\theta) = &\sum_{e\in S_e; a\in S_a; r\in S_r} p^e (1-p)^{1-e} z(x,\nu)^a (1-z(x,\nu))^{1-a} \times \\ &\phi(x,a,\nu)^{1-r} (1-\phi(x,a,\nu))^r q(e,a,r)^t (1-q(e,a,r))^{1-t} f(u \mid x; a; r; t; \nu; \theta).\end{aligned}$$

[19] The likelihood function of individuals in sample D is written as if the sample included all the individuals outside the experiment (i.e. as if sample B was the complement of samples A, B, and C). In principle, the likelihood function should be weighted to account for the fact that sample D is a subsample of those outside the experiment. As mentioned in Footnote 8, sample D comprises over 95 per cent of that population. Further, selection into the sample was made using a random procedure. We have thus chosen not to weigh the function to avoid making an already involved function overly complicated.

[20] Whether there is unique mapping between these reduced form equations and the structural model may be questioned. Note that we imposed a number of restrictions on the covariance matrix of the reduced form model. In particular, the dichotomization of the latent variables corresponding to the acceptance and recontact variables imposes variances that are normalized to unity. Further, there are no correlations between the latent variables and the duration variable. It is then possible to show that a generalized order condition holds for each latent equation in the conditional model (see Fomby, Hill, & Johnson, 1984). It should be noted, however, that assuming there is no correlation between the latent variables does not imply they are independent. Indeed, the conditional expectation of the recontact variable depends on the acceptance decision. Consequently, whereas the error terms ϵ_a and ϵ_r are assumed to be independent, the recontact variable R^* and the acceptance variable A^* are correlated. The correlation between the two latent variables is given by the parameter a (see equation (6)).

UNOBSERVED HETEROGENEITY

Estimation of the parameters by means of maximum likelihood requires specification of the distribution of the unobserved heterogeneity terms. We first approximate arbitrary continuous distributions using a finite number of mass points (see Heckman & Singer). Next we investigate the robustness of the slope parameters using various continuous distributions.

Discrete Distributions

Let V denote the random variable associated to the unobserved heterogeneity terms.

Assume that

$$\text{Prob}[V = v] = \begin{cases} p_0, & \text{if } v = \nu_0, \\ (1 - p_0), & \text{if } v = -\nu_0, \end{cases} \tag{11}$$

where the probability p_0 is defined as

$$p_0 = \Phi(d),$$

where $d, \nu_0 \in I\!R$ are parameters, and Φ is the cumulative distribution function of the normal distribution with mean 0 and variance 1.

This unrestricted model is estimated first. Next, we consider a restricted version which imposes $d = 0$ or, equivalently, $p = 0.5$ (i.e. $E(V) = 0$).

Continuous Distributions

The unobserved heterogeneity terms ν are assumed to be independently and identically distributed. Let $g(\nu; \gamma)$ be the pdf of ν, with $g(\nu; \gamma)$ representing any well-behaved probability density function (e.g. the pdf of normal or student distributions).

SPECIFICATION OF CONDITIONAL HAZARD FUNCTION

The conditional hazard function for welfare durations is given by

$$h(u \mid x; a; r; t; \nu; \theta) = h_0(u; \alpha)\ \varphi(x; a; r; t; \beta_d)\ \exp(-\nu), \tag{12}$$

where φ is a positive function of the exogenous variables, x, and of a, r, and t, and where $h_0(u; \alpha)$ is the baseline hazard function. Depending on which version of the model is estimated, x may or may not include a constant. We assume

$$\varphi(x; a; r; t; \beta_d) = \exp(-x'\beta_x - a\ \delta_a - r\ \delta_r - t\ \delta_t),$$

where $\delta_a, \delta_r, \delta_t \in I\!R$ and β_x are vectors of parameters.

The baseline hazard function is

$$h_0(u; \alpha) = \alpha\ u^{\alpha-1},$$

$\alpha \in I\!R^+$. Consequently, welfare duration is assumed to be distributed as a Weibull random variable. If $\alpha > 1$, then the hazard function is increasing with respect to u. If $\alpha < 1$, then the

hazard function is decreasing with respect to u, and if $\alpha = 1$, the conditional hazard function is constant.[21]

For uncensored spells, the contribution of the welfare duration is given by the conditional probability density function

$$\begin{aligned} f(u \mid x; a; r; t; \nu; \theta) &= h(u \mid x; a; r; t; \nu; \theta) \ \exp\left\{-\textstyle\int_0^u h(s \mid x; a; r; t; \nu; \theta)\, d\, s\right\}, \\ &= \alpha\, u^{\alpha-1} \varphi(x; a; r; t; \beta_d) \ \exp(-\nu) \exp\left\{-\varphi(x; a; r; t; \beta_d) \ \exp(-\nu) u^{\alpha}\right\}, \end{aligned}$$

where $u < 64$ months.

The contribution of censored spells is given by the conditional survival function

$$\begin{aligned} f(u \mid x; a; r; t; \nu; \theta) &= \exp\left\{-\textstyle\int_0^u h(s \mid x; a; r; t; \nu; \theta)\, d\, s\right\}, \\ &= \exp\left\{-\varphi(x; a; r; t; \beta_d) \ \exp(-\nu) u^{\alpha}\right\}, \end{aligned}$$

if $u \geq 64$ months.

ESTIMATION

We consider two alternative specifications for the unobserved heterogeneity distribution.

Discrete Distribution

The log-likelihood is

$$\log(L(\theta)) = \sum_{i=1}^{N} \log(l_i(\theta)), \tag{13}$$

where $l_i(\theta)$ is obtained by substituting the sequence $y_i = (e_i, a_i, r_i, t_i, u_i)$ and the observed vector of covariates x_i in (3), and where N is the sample size.

In equation (3), π_j is set equal to[22]

$$\pi_j = \begin{cases} p_0, & \text{if } j = 1, \\ (1 - p_0), & \text{if } j = 2, \end{cases}$$

where $\pi_1 = \text{Prob}[V = \nu_0]$, $\pi_2 = \text{Prob}[V = -\nu_0]$, and $\nu_0 \in I\!R$ is a parameter. The log-likelihood is then maximized with respect to θ ($\theta \in \Theta$). The number of mass points, J, is set to 2;[23] π_1 represents the probability the unobserved term V takes the value ν_0 ($\pi_2 = 1 - \pi_1$).

[21] Note that the hazard function of the Weibull model with parametric unobserved heterogeneity need not be monotonic in duration. In fact, if the distribution function of the unobserved heterogeneity is gamma, the hazard function is non-monotonic and is known as the Singh-Maddala.

[22] See the "Likelihood Function" section.

[23] The data support only two mass points. This is due to the fact that the individuals in our sample are relatively homogeneous as shown in Table 1.

Continuous Distribution

The model includes unobserved heterogeneity terms ν ($\nu > 0$). We assume these terms to be independently and identically distributed. Let $g(\nu;\gamma)$ be the pdf of ν.

The contribution of a given realization to the likelihood function is given by equation (4), where $S = I\!R^+$. The log-likelihood is given by equation (13), where $l_i(\theta)$ is the contribution to the likelihood of the sequence y_i.[24] Since the integral in $l(\theta)$ generally cannot be analytically computed, it must be numerically simulated.

Let $\hat{l}(\theta)$ denote the estimator of the individual contribution to the likelihood function. We assume

$$\hat{l}(\theta) = \frac{1}{H}\sum_{h=1}^{H} f(y \mid x;\nu_h;\theta),$$

where ν_h are drawn independently according to the pdf $g(\nu;\gamma)$. The drawings ν_h ($h = 1,\ldots,H$) are assumed to be specific to the individual. The parameter estimates are obtained by maximizing the simulated log-likelihood:

$$\log(L(\theta)) = \sum_{i=1}^{N} \log(\hat{l}_i(\theta)),$$

where $\hat{l}_i(\theta)$ is the simulated contribution of the sequence y_i to the likelihood function.

The maximization of this simulated likelihood yields consistent and efficient parameter estimates if $\frac{\sqrt{N}}{H} \to 0$ when $H \to +\infty$ and $N \to +\infty$ (see Gouriéroux & Monfort, 1991, and and Gouriéroux & Monfort, 1996). Under these conditions, this estimator has the same asymptotic distribution as the standard maximum likelihood (ML) estimator. We have used 1,000 draws from the random distributions when estimating the models. Using as few as 100 draws yielded essentially the same parameter estimates. Usually, fewer draws are considered adequate (see Kamionka, 1998; Gilbert, Kamionka, & Lacroix, 2001).

INCOMPLETE INFORMATION SCHEMES

It is possible to examine the impact of the non-response biases on the treatment effect by considering various estimates obtained using more or less complete information schemes. For instance, we can estimate the treatment effect using only the control and the treatment groups A and B.

Let f define the conditional density of the welfare durations, given the conditioning variables and the value of the vector of the parameters.

Treatment and Control Groups

Each individual contributes a sequence $y = (t, u)$ to the likelihood function. They all agreed to participate and all could be contacted at baseline (see Figure 8).

[24] In what follows, θ includes γ, the parameters of $q(\cdot)$.

The conditional contribution of a given realization to the likelihood function is

$$\ell_\nu(\theta) = 0.5\ f(u \mid x; t = 1; \nu; \theta),$$

if the individual belongs to A, and

$$\ell_\nu(\theta) = 0.5\ f(u \mid x; t = 0; \nu; \theta),$$

if the individual belongs to B.

The conditional distribution of the welfare durations corresponds to the hazard function (12), where $\delta_a = \delta_r = 0$. (Here, a and r are set equal to arbitrary values in the conditional distribution of the welfare duration.)

Participants in the Experiment

Each individual contributes a sequence $y = (r, t, u)$ to the likelihood function. All were selected for the experiment, some could be contacted, but others could not be reached (see Figure 9). Those who were contacted were offered the treatment with probability $p = 0.5$.

The conditional contribution of a given realization to the likelihood function is

$$\ell_\nu(\theta) = (1 - \phi(x, \nu))\ 0.5\ f(u \mid x; r = 1; t = 1; \nu; \theta),$$

if the individual belongs to A;

$$\ell_\nu(\theta) = (1 - \phi(x, \nu))\ 0.5\ f(u \mid x; r = 1; t = 0; \nu; \theta),$$

if the individual belongs to B; and

$$\ell_\nu(\theta) = \phi(x, \nu)\ f(u \mid x; r = 0; t = 0; \nu; \theta),$$

if the individual belongs to C.

Here, $\phi(\nu, x)$ denotes the conditional probability the individual could not be contacted and is defined as in the context of a complete information scheme (see equation (6)), where $\xi_a = 0$. (In this equation and in the expression of the conditional hazard function, a is fixed to an arbitrary value.)

The expression of the conditional hazard function of the welfare durations is given by equation (12), where $\delta_a = 0$.

Selected and Non-selected Welfare Applicants

Each individual contributes a sequence $y = (e, a, t, u)$ to the likelihood function. Those who were selected at baseline have agreed to participate. Those who were not selected may or may not have agreed (see Figure 10).

The conditional contribution of a given realization to the likelihood function is

$$\ell_\nu(\theta) = p\ z(x, \nu)\ \ 0.5\ f(u \mid x; a = 1; t = 1; \nu; \theta),$$

if the individual belongs to A;

$$\ell_\nu(\theta) = p\, z(x,\nu)\, 0.5\, f(u \mid x; a=1; t=0; \nu; \theta),$$

if the individual belongs to B; and

$$\begin{aligned}
\ell_\nu(\theta) &= p\,(1-z(x,\nu))\; f(u \mid x; a=0; t=0; \nu; \theta), \\
&+ (1-p)\, z(x,\nu)\, f(u \mid x; a=1; t=0; \nu; \theta), \\
&+ (1-p)\,(1-z(x,\nu))\, f(u \mid x; a=0; t=0; \nu; \theta),
\end{aligned}$$

if the individual belongs to D.

Here, $z(x,\nu)$ is the conditional probability the individual agrees to participate in the experiment. The definition of $z(x,\nu)$ is similar to the one given for the complete information scheme (see equation (5)).

The expression of the conditional hazard function of the welfare durations is given by equation (12), where $\delta_r = 0$. (For convenience, r is fixed to an arbitrary value in the expression of the conditional hazard.)

Results

SINGLE TREATMENT EFFECT

The estimation results presented in Table 4 investigate the overall impact of the treatment on the average spell duration. Since the experiment's set-up is expected to delay exit before the qualifying period and to hasten it in the following months, using a single treatment effect provides a measure of the program's net impact. The first four columns of Table 4 provide estimates based on non-parametric unobserved heterogeneity (see equation (11)).[25]

The estimates of the first column are obtained from the experimental samples only. This specification is the only one in which we omit unobserved heterogeneity. This is done for two reasons. First, given that individuals were randomly assigned to control and treatment groups, unobserved characteristics should be distributed similarly across groups. Second, the maximum likelihood estimator of the treatment effect that neglects unobserved heterogeneity should be relatively close to a simple difference in mean durations between the two groups.

The estimate of α indicates the hazard function is decreasing with duration. The slope parameters show that duration increases with the number of children and decreases with age. Both parameter estimates are highly statistically significant. Women are also found to have longer mean spell durations than men. Finally, the treatment effect is found to reduce spell duration by approximately 7.5 per cent. This estimate is quite similar to that reported in the "Non-parametric Evidence" section, where it is found the treatment group had a 7.4 per cent shorter mean duration.

Column (2) of the table reports the results using groups A, B, and C (see Figure 9). The baseline hazard function decreases with duration. As previously, spell duration decreases with age and increases with the number of children. Likewise, women have longer spell durations than men. The impact of the treatment is very similar to that of column (1), although it is not statistically significant. Note that the parameter estimate of the contact binary variable is positive and significantly different from 0. This is consistent with the observation that individuals in sample C have significantly shorter spells (see Table 1). Hence, once we include those who could not be contacted at baseline, the treatment effect vanishes. The third panel of Table 4 reports the parameter estimates of the probability of not being contacted at baseline. The probability decreases with age and the number of children. Women are also less likely not to be contacted than men. These results are consistent with those obtained for descriptive statistics on sample C (see Table 1).

Column (3) of Table 4 reports the results using groups A, B, and D (see Figure 10). Contrary to the previous cases, the conditional hazard function increases with duration. Inclusion of this group allows us to model explicitly the participation decision. Omission of the latter thus induces a spurious negative duration dependence. This phenomenon is well known in duration models. The marginal duration model is the mixture of conditional duration models with respect of the acceptance decision. The sign of the slope parameters are similar to those obtained using groups A, B, and C. The parameter of the acceptance binary

[25] We only report results based on the restricted version (i.e. $p = 0.5$). Except for a few specifications, p could be estimated freely. The parameter estimates are relatively robust for the estimation of p.

variable is positive and statistically significant. Thus, among the individuals who could be contacted a priori, those who decided to participate have longer mean spell durations. The treatment effect is now nearly four times greater than the one obtained using samples A and B. Consequently, omission of the participation decision significantly biases the effect of the earning supplement on exits from welfare. The second panel of Table 4 reports the parameters of the conditional probability of agreeing to participate in the experiment. Not a single parameter is statistically significant in this specification.

Column (4) of Table 4 reports the results using groups A, B, C, and D (see Figure 6). The parameter estimates show the conditional hazard function is increasing with duration. The sign of the slope parameters are similar to those of the previous specifications. The impact of the treatment is again nearly four times greater than the one obtained using the experimental groups only. Spell duration is also longer for participants and for those who could be contacted. Both parameter estimates are statistically significant.

The next two panels indicate the probability of not being contacted decreases with age and the number of children, and is higher for women than for men. The parameters are very similar to those obtained using groups A, B, and C. Furthermore, the probability is significantly lower for those who are willing to participate *ex ante*. Finally, note that the probability of agreeing to participate increases with age, and the parameter estimate is statistically significant at five per cent.

The estimates in columns (1) to (4) of Table 4 are based on a rather restrictive specification for the unobserved heterogeneity component. Previous research has shown that the slope parameters of duration models are usually rather insensitive to particular distributional assumptions (see Heckman & Borgas, 1980; Bonnal, Fougère, & Sérandon, 1997; Gilbert et al., 2001). It is thus worth investigating whether our results are also robust to various assumptions pertaining to the distribution of the unobserved heterogeneity.

The last four columns of Table 4 report results based on a particular parametric distribution using samples A, B, C, and D. The parameter estimates are comparable to those of column (4). The treatment effect is still sizable, although slightly smaller than that of column (4), except for the specification based on the student distribution (with five degrees of freedom). As with column (4), the mean spell duration of those who could be contacted or agreed to participate in the experiment is considerably longer. Furthermore, the parameter estimates of the two latent equations are very similar to those of column (4). Thus, the estimates of the treatment effect appear relatively robust with respect to the distribution of the unobserved heterogeneity.

MULTIPLE TREATMENT EFFECTS

The parameter estimates of the treatment effect presented in Table 4 make no distinction between the qualifying period and the ensuing months. Yet, the experiment is set up to measure potential delayed exit effects that may arise with a full-scale program. The non-parametric evidence provided in previous sections suggested that such effects are likely rather small, if at all significant. Our model can easily be modified to account for potential time-varying treatment effects. Using the experiment's design, we have re-estimated the

model by allowing the treatment to have a differentiated impact on the duration at discrete intervals ([0,11], [12,23], [24,35], [36 and more].).

The estimation results are reported in Table 5. The table has the same set-up as Table 4. The specification in the first column uses samples A and B. According to the parameter estimates, the treatment group does not appear to delay exit any more than the control group since the parameter estimate of the treatment effect is not statistically different from zero. The treatment effects for subsequent intervals are all highly significant. The results indicate the treatment effect reduces durations considerably over the [12,23] and [24,35] intervals. On the other hand, the treatment group appears to have longer spells over the [36 and more] interval. The parameter α indicates there is negative duration dependence in the data.

The second column reports the estimation results using samples A, B, and C. This specification yields rather strange results. Indeed, the parameter estimates suggest the treatment group has a much longer mean spell duration that the control group. There are no appealing reasons that may justify such a result, but further investigation certainly seems warranted.

Columns (3) and (4) yield essentially similar results. Contrary to the first two specifications, there now appears to be positive duration dependence in the data. Furthermore, the parameter estimates suggest there is no evidence of exit delayed behaviour. If anything, the treatment group has a shorter conditional duration over the [0,11] interval. Likewise, the treatment effect over the [12,23] and [24,35] intervals reduces duration considerably. In both cases, it is found the treatment has no impact on the mean duration over the [36 and more] interval.

The specifications in columns (5) to (8) are identical to that of column (4), but use parametric distributions for the unobserved heterogeneity. The parameter estimates of the treatment effect are qualitatively similar to those of columns (3) and (4), except they are much smaller in magnitude. Furthermore, only in column (5) is the treatment found to have an impact on the duration over the [36 and more] interval.

MEAN DURATIONS

The slope parameters cannot be directly interpreted as marginal impacts, since the expected duration is highly non-linear with respect to the covariates.[26] We thus report the conditional (on treatment) expected durations for various model specifications in Table 6. The top panel reports the expected durations based on the parameters of the first column of Table 4. This specification allows only one treatment effect and is based on the experimental samples only. The expected durations are computed by bootstrapping the samples 500 times and averaging the mean durations across individuals. This allows us to integrate over the distribution of the covariates in the experimental population. The table shows that men have somewhat shorter durations than women. Likewise, the treatment effect reduces duration by approximately 6.9 per cent for women, and 7.7 per cent for men.

[26] Indeed, it can be shown that $E(U|X,\nu,\theta) = \lambda^{-\frac{1}{\alpha}}\Gamma(1+1/\alpha)$, where $\lambda = \exp(-X'\beta - \nu)$.

The middle panel uses the same parameter estimates as the top panel, except the drawing is made within sample D. This allows us to measure the impact of differing distributions of the covariates between the experimental samples and the population of welfare recipients. The results show that the mean durations are very similar to those of the top panel. This is not surprising, given the results reported in Table 2. If anything, the durations are slightly shorter when using data from sample D as opposed to the experimental samples.

The bottom panel of Table 6 uses the parameter estimates of the fourth column of Table 5. The treatment effect is allowed to vary with duration, and data from all samples are used to estimate the parameters. To compute mean durations, only data from sample D are used since this sample best mimics the population of welfare recipients. The table shows that the treatment is much larger when using the complete model. Indeed, the treatment effect is found to reduce mean spell duration by as much as 25 per cent for both men and women.

To the extent that our model properly accounts for the non-response bias in the data, it must be concluded that the expected durations of experimental data void of any bias would be considerably shorter. We conjectured previously that such bias did not necessarily imply that the impact of the treatment itself would be biased. However, according to our parameter estimates and simulations, it does seem the estimate is biased.

Conclusion

Over the past 20 years, experimental designs have become the preferred means of evaluating employment and training programs. This is not surprising given that, in an ideal setting, social experimentation is able to solve the so-called "evaluation problem." In practice, implementation of a demonstration project is likely to be hampered by many logistical and behavioural problems that may prove detrimental to the quality of the data generated (see Hotz, 1992). Although the literature has singled out non-response or randomization bias as the main culprit, we know surprisingly little about the extent to which demonstrations are contaminated by these potential problems. The evidence is almost always indirect or inferential at best.

In Canada a policy to help single parents on social assistance become self-reliant was implemented on an experimental basis. The Self-Sufficiency Project (SSP) focused on newly enrolled recipients. The applicant study randomly selected a sample of 4,134 single parents who had applied for welfare between January 1994 and March 1995. Only 3,315 agreed to be part of the experiment despite a 50 per cent chance of receiving a generous, time-limited, earnings supplement conditional on finding a full-time job and leaving income assistance.

The purpose of this paper is to determine whether a non-response rate as high as 20 per cent is likely to bias the measurement of the treatment effect. Our empirical strategy is to compare the estimated impact of the program using experimental data with those results obtained using additional data on individuals not taking part in the experiment and drawn from the same population. We identify three reasons for not participating in the experiment. First, some recipients simply were not selected at baseline. Second, some were selected but refused to participate. Third, some were selected, but could not be reached at baseline. We write the likelihood of various sets of information and obtain relevant estimates of program impact on welfare spell durations.

We find strong evidence of non-response bias in the data. When we correct for the bias, we find the estimates of the treatment effect that rely solely on experimental data underestimate the true impact of the program. We conjecture this is because those who agreed to participate have longer mean spell durations and are likely less responsive to financial incentives than others. Furthermore, we find no evidence of the so-called "delayed exit effect" that may arise due to the program set-up.

Finally, the sensitivity of the parameter estimates to distributional assumptions pertaining to the unobserved heterogeneity is also investigated. We find many parametric distributions yield similar results to those obtained from a simple non-parametric model.

Tables and Figures

Table 1: Descriptive Statistics

Variable	Sample			
	A	B	C	D
Sex (Women $= 1$)	0.89	0.91	0.86	0.90
	(0.31)	(0.28)	(0.34)	(0.30)
Age	32.65	32.37	31.79	32.42
	(7.88)	(7.41)	(7.85)	(7.73)
Children	1.65	1.68	1.57	1.65
	(0.80)	(0.82)	(0.77)	(0.81)
Mean spell length†	20.28	21.75	13.76	20.34
	(0.47)	(0.51)	(0.75)	(0.38)
Median spell length	15	13	4	11
Proportion of censured spells	7.83	10.20	6.59	9.63
No. observations	1,648	1,667	637	3,073

Note: †Estimated from Kaplan-Meir survival rates and tail corrections proposed by Brown, Hollander, and Korwar, 1974.

Table 2: Logit Regressions

	Sample			
Variable	A vs. B	A vs. D	B vs. D	C vs. D
Intercept	0.151	-0.700^{*}	-0.851^{*}	-0.650^{*}
	(0.215)	(0.184)	(0.186)	(0.253)
Sex (Women $= 1$)	-0.193	-0.021	0.173	-0.378^{*}
	(0.122)	(0.103)	(0.108)	(0.135)
Children	-0.065	-0.018	0.047	-0.102^{**}
	(0.044)	(0.034)	(0.038)	(0.057)
Age	0.003	0.004	0.001	-0.013^{*}
	(0.005)	(0.184)	(0.004)	(0.006)
Observations	3,315	4,721	4,740	3,710
Log-likelihood	$-2,294.5$	$-3,053.3$	$-3,071.5$	$-1,693.6$

Notes: * Statistically significant at five per cent or better.
** Statistically significant at 10 per cent or better.

Table 3: Realizations of Random Variables

Group	E	A	R	T
A	1	1	1	1
B	1	1	1	0
C	1	0,1	0	0
D	0,1	0,1	0,1	0

Table 4: Maximum Likelihood Estimates: Single Treatment Effect

Parameter Estimates	Non-Parametric Heterogeneity				Parametric Heterogeneity			
	$A+B$	$A+B+C$	$A+B+D$	$A+B$ $+C+D$	$A+B$ $+C+D$	$A+B$ $+C+D$	$A+B$ $+C+D$	$A+B$ $+C+D$
					Expo-nential	Gamma	Log-Normal	Student (5)
Duration								
α	0.873	0.896	1.506	1.382	1.048	1.035	0.983	0.993
	(0.013)	(0.015)	(0.026)	(0.024)	(0.020)	(0.020)	(0.016)	(0.019)
ν		0.460	−1.326	−1.246	−0.424	−0.497	−1.499	−1.236
		(0.036)	(0.039)	(0.041)	(0.073)	(0.074)	(0.107)	(0.217)
Intercept	2.753	2.027	3.820	2.552	1.493	1.458	1.293	1.109
	(0.120)	(0.121)	(0.149)	(0.133)	(0.137)	(0.134)	(0.135)	(0.130)
Women	0.198	0.209	0.161	0.213	0.272	0.277	0.222	0.215
	(0.064)	(0.064)	(0.065)	(0.062)	(0.053)	(0.052)	(0.047)	(0.057)
Age/100	−0.697	−0.776	−1.063	−0.579	−0.988	−0.900	−0.716	−0.605
	(0.240)	(0.249)	(0.251)	(0.242)	(0.213)	(0.207)	(0.190)	(0.213)
Children	0.203	0.203	0.239	0.269	0.202	0.196	0.187	0.189
	(0.052)	(0.055)	(0.058)	(0.058)	(0.047)	(0.046)	(0.043)	(0.046)
Treatment	−0.075	−0.059	−0.288	−0.294	−0.176	−0.187	−0.186	−0.259
	(0.037)	(0.042)	(0.044)	(0.048)	(0.037)	(0.037)	(0.033)	(0.036)
Accept			1.148	1.167	1.495	1.560	1.727	1.620
			(0.112)	(0.086)	(0.125)	(0.115)	(0.115)	(0.136)
Contacted		0.810		0.242	0.431	0.336	0.196	0.208
		(0.066)		(0.077)	(0.160)	(0.141)	(0.160)	(0.125)
Acceptance								
Intercept			2.026	1.461	1.043	1.046	0.978	0.785
			(0.245)	(0.201)	(0.187)	(0.184)	(0.182)	(0.180)
Women			0.130	0.112	0.180	0.166	0.202	0.232
			(0.124)	(0.107)	(0.100)	(0.098)	(0.094)	(0.098)
Age/100			−0.419	0.402	−0.049	−0.087	−0.162	−0.066
			(0.546)	(0.443)	(0.419)	(0.413)	(0.407)	(0.395)
Children			−0.011	0.021	0.031	0.029	0.026	0.024
			(0.114)	(0.093)	(0.090)	(0.089)	(0.087)	(0.085)
Not Contacted								
Intercept		−0.493		1.860	1.328	1.288	1.039	0.576
		(0.154)		(0.212)	(0.245)	(0.243)	(0.226)	(0.220)
Women		−0.288		−0.276	−0.284	−0.297	−0.234	−0.192
		(0.085)		(0.111)	(0.122)	(0.118)	(0.109)	(0.108)
Age/100		−0.988		−0.880	−1.540	−1.463	−1.475	−1.114
		(0.085)		(0.433)	(0.510)	(0.512)	(0.466)	(0.437)
Children		−0.140		−0.165	−0.177	−0.176	−0.170	−0.148
		(0.078)		(0.094)	(0.120)	(0.115)	(0.107)	(0.096)
Accepted				−3.732	−2.346	−2.279	−1.899	−1.593
				(0.122)	(0.134)	(0.133)	(0.132)	(0.150)
Likelihood	−12,391	−18,522	−33,553	−34,310	−34,427	−34,453	−34,470	34,491

Table 5: Maximum Likelihood Estimates: Multiple Treatment Effects

Parameter Estimates	Non-Parametric Heterogeneity				Parametric Heterogeneity			
	$A+B$	$A+B+C$	$A+B+D$	$A+B$ $+C+D$	$A+B$ $+C+D$	$A+B$ $+C+D$	$A+B$ $+C+D$	$A+B$ $+C+D$
					Expo-nential	Gamma	Log-Normal	Student (5)
Duration								
α	0.783	0.880	1.451	1.462	1.111	1.065	1.053	1.008
	(0.011)	(0.016)	(0.031)	(0.025)	(0.025)	(0.021)	(0.021)	(0.018)
ν		−0.622	1.330	−1.384	−0.214	−0.083	−1.124	−1.479
		(0.053)	(0.045)	(0.038)	(0.067)	(0.078)	(0.093)	(0.232)
Intercept	3.061	1.832	3.001	2.763	0.746	0.906	0.803	1.364
	(0.141)	(0.136)	(0.147)	(0.131)	(0.147)	(0.151)	(0.163)	(0.120)
Women	0.236	0.207	0.172	0.189	0.291	0.263	0.252	0.212
	(0.080)	(0.067)	(0.066)	(0.062)	(0.058)	(0.055)	(0.053)	(0.056)
Age/100	−0.817	−0.883	−0.765	−0.609	−1.244	−1.034	−0.956	−0.520
	(0.303)	(0.264)	(0.255)	(0.239)	(0.231)	(0.215)	(0.210)	(0.206)
Children	0.241	0.214	0.283	0.247	0.209	0.200	0.200	0.177
	(0.065)	(0.059)	(0.060)	(0.056)	(0.050)	(0.047)	(0.046)	(0.045)
Treatment								
$T < 12$	0.074	−0.053	−0.382	−0.329	−0.256	−0.284	−0.290	−0.327
	(0.059)	(0.046)	(0.075)	(0.075)	(0.048)	(0.047)	(0.046)	(0.049)
$12 \leq T < 24$	−0.254	1.107	−0.621	−0.634	−0.125	−0.143	−0.149	−0.290
	(0.074)	(0.101)	(0.074)	(0.070)	(0.062)	(0.059)	(0.058)	(0.055)
$24 \leq T < 36$	−0.444	1.041	−0.539	−0.529	−0.391	−0.342	−0.326	−0.288
	(0.094)	(0.089)	(0.073)	(0.073)	(0.078)	(0.073)	(0.072)	(0.073)
$T \geq 36$	0.444	0.763	0.103	0.119	−0.249	−0.118	−0.059	0.099
	(0.105)	(0.080)	(0.084)	(0.084)	(0.099)	(0.091)	(0.087)	(0.086)
Accept			1.240	1.133	1.293	1.517	1.574	1.821
			(0.108)	(0.094)	(0.118)	(0.112)	(0.111)	(0.115)
Contacted		0.642		0.269	0.869	0.695	0.633	0.095
		(0.078)		(0.078)	(0.162)	(0.171)	(0.184)	(0.103)
Acceptance								
Intercept			2.031	1.615	0.152	0.448	0.375	0.757
			(0.237)	(0.198)	(0.175)	(0.166)	(0.167)	(0.163)
Women			0.132	0.092	0.205	0.201	0.198	0.234
			(0.119)	(0.105)	(0.093)	(0.089)	(0.088)	(0.090)
Age/100			−0.426	0.405	−0.049	−0.115	−0.133	−0.074
			(0.518)	(0.440)	(0.400)	(0.383)	(0.382)	(0.358)
Children			−0.003	0.009	0.011	0.013	0.014	0.031
			(0.112)	(0.092)	(0.085)	(0.081)	(0.080)	(0.079)
Not Contacted								
Intercept		−0.541		2.021	0.236	0.346	0.220	0.525
		(0.164)		(0.209)	(0.227)	(0.213)	(0.211)	(0.202)
Women		−0.312		−0.281	−0.223	−0.226	−0.230	−0.185
		(0.091)		(0.109)	(0.117)	(0.108)	(0.106)	(0.102)
Age/100		−1.023		−0.875	−1.648	−1.531	−1.478	−1.118
		(0.376)		(0.428)	(0.511)	(0.472)	(0.460)	(0.417)
Children		−0.152		−0.169	−0.188	−0.169	−0.164	−0.142
		(0.083)		(0.093)	(0.116)	(0.107)	(0.105)	(0.092)
Accepted				−4.031	−2.259	−1.847	−1.726	−1.510
				(0.117)	(0.136)	(0.121)	(0.125)	(0.125)
Likelihood	−12,391	−18,499	−25,758	−34,253	−34,387	−34,409	−34,416	−34,457

Table 6: Mean Spell Duration*

Model		Women and Men	Women	Men
		Experimental Sample $(A+B)$		
Model $A+B^{\dagger}$	$T=0$	23.547	24.082	18.568
		(0.044)	(0.035)	(0.091)
	$T=1$	21.913	22.426	17.138
		(0.043)	(0.034)	(0.086)
		Sample D		
Model $A+B^{\dagger}$	$T=0$	23.490	24.040	18.698
		(0.046)	(0.034)	(0.089)
	$T=1$	21.857	22.385	17.260
		(0.044)	(0.036)	(0.084)
		Sample D		
Model $A+B+C+D^{\ddagger}$	$T=0$	26.130	26.417	23.644
		(0.019)	(0.012)	(0.057)
	$T=1$	19.309	19.594	16.836
		(0.020)	(0.015)	(0.054)

Notes: * Computed on the basis of 500 replications of the relevant samples. Empirical standard errors in parentheses.

† Based on the parameter estimates of column (1), Table 4.

‡ Based on the parameter estimates of column (4), Table 5.

Figure 1: Randomization Scheme

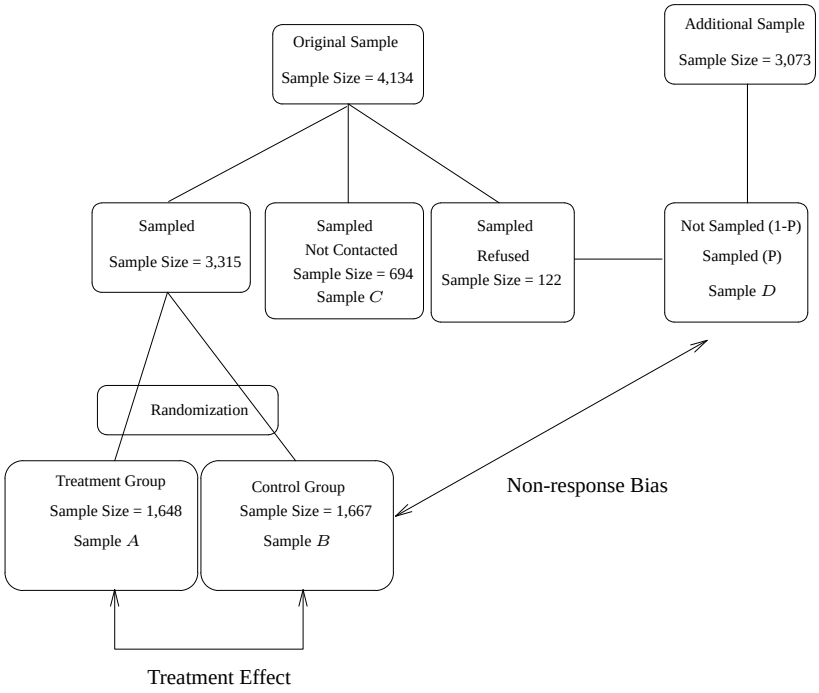


Figure 2: Kernel Smoothed Hazard Functions — Experimental Groups

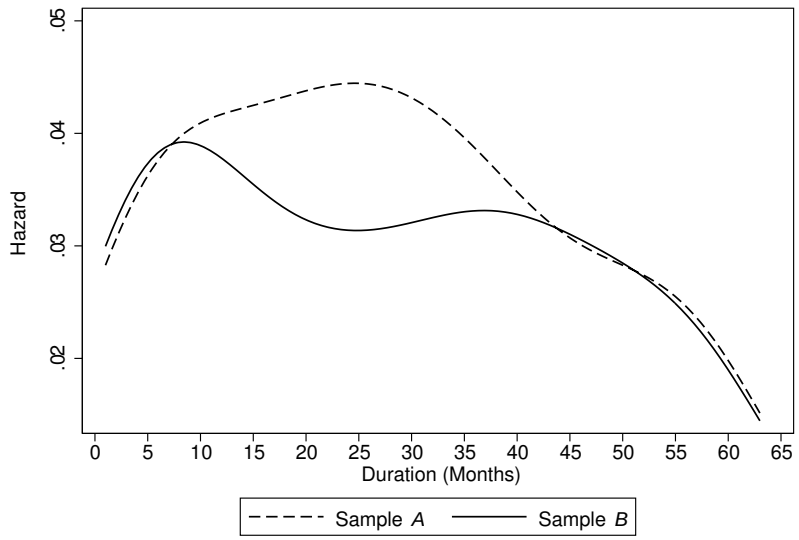


Figure 3: Survival Functions — Experimental Groups

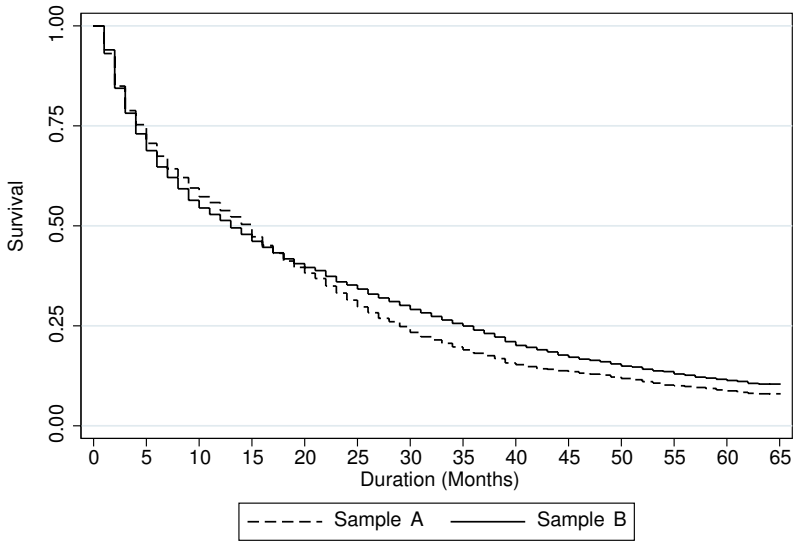


Figure 4: Confidence Intervals of Survival Functions — Experimental Groups

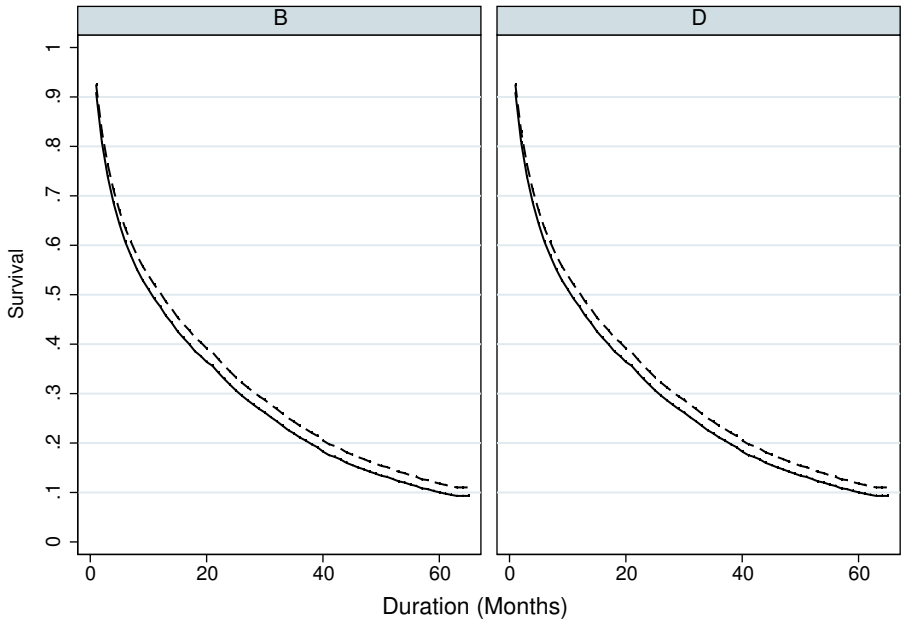


Figure 5: Survival Functions — Control, Not Contacted and Unsampled Groups

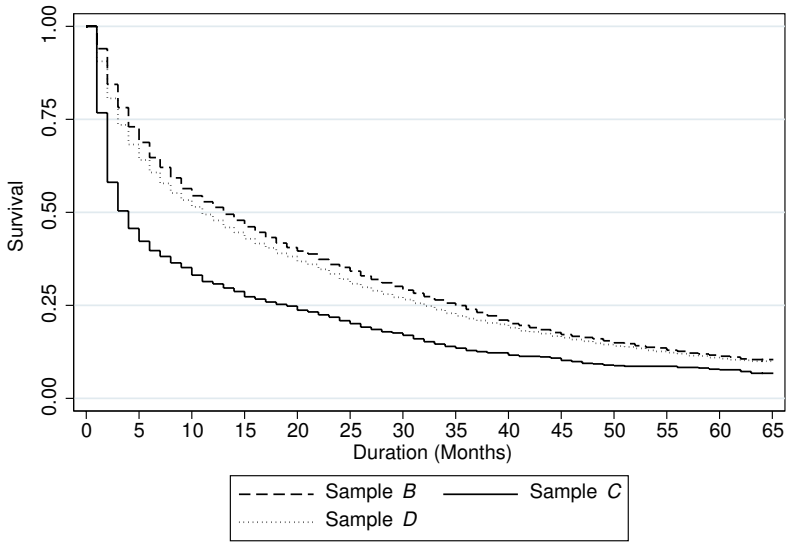


Figure 6: Welfare Applicants

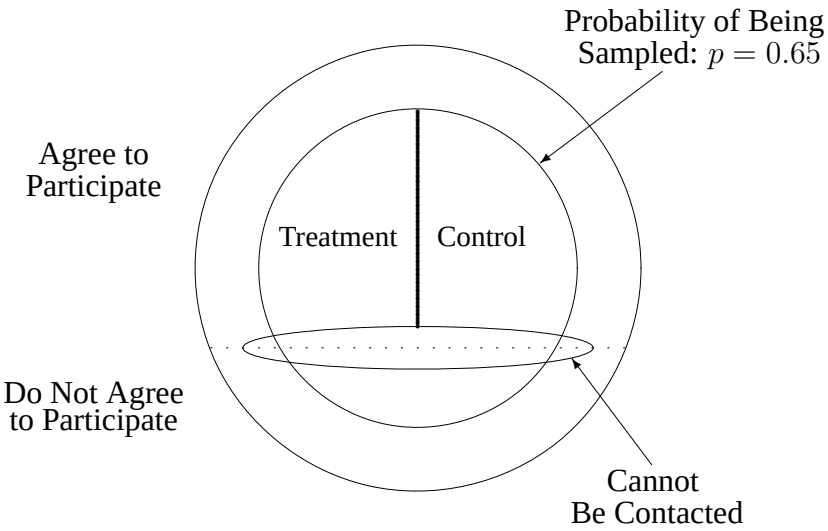


Figure 7: Welfare Applicants

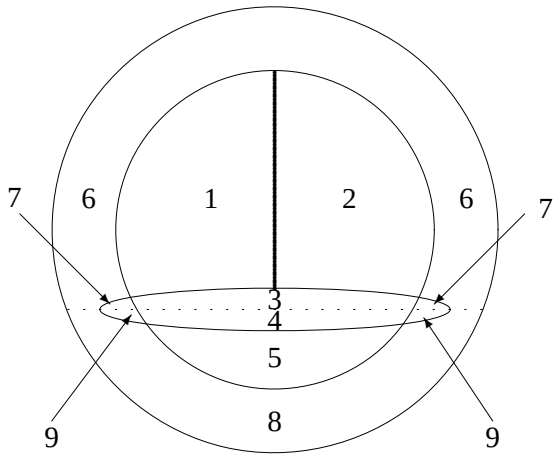


Figure 8: Participants in the Experiment Who Could Be Contacted

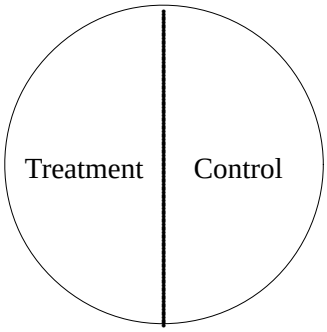


Figure 9: Participants in the Experiment

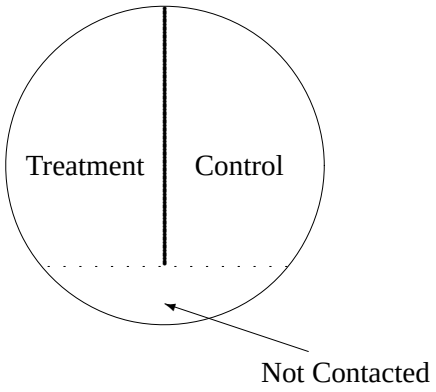
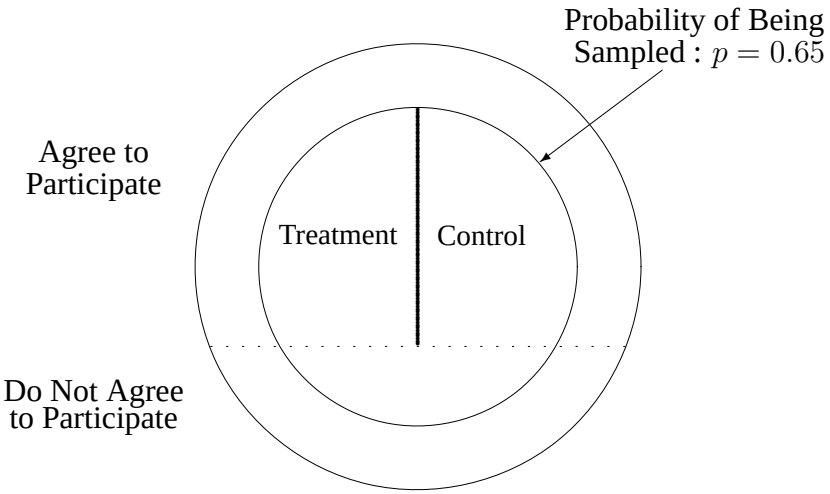


Figure 10: Selected and Non-selected Welfare Applicants



References

Berlin, G., Bancroft, W., Card, D., Lin, W., & Robins, P. K. (1998). *Do work incentives have unintended consequences? Measuring "entry effects" in the Self-Sufficiency Project.* Ottawa: Social Research and Demonstration Corporation.

Bonnal, L., Fougère, D., & Sérandon, A. (1997). Evaluating the impact of French employment policies on individual labour market histories. *The Review of Economics Studies, 64*(4), 683–718.

Brown, J. B., Hollander, W., & Korwar, R. M. (1974). Nonparametric tests of independence for censured data, with applications to heart transplant studies. In F. Proschan & R. J. Serling (Eds.), *Reliability and biometry: Statistical analysis of lifelength.* Philadelphia: Society for Industrial and Applied Mathematics, pp. 327–354.

Drolet, S., Fortin, B., & Lacroix, G. (Forthcoming). Welfare benefits and the duration of welfare spells: Evidence from a natural experiment in Canada. *Journal of Public Economics.*

Fomby, T. B., Hill, R. C., & Johnson, S. R. (1984). *Advanced econometric methods.* New York: Springer-Verlag.

Fougère, D., Fortin, B., & Lacroix, G. (2002). The effects of welfare benefits on the duration of welfare spells: Evidence from a natural experiment in Canada. In C. d'Aspremont & P. Pestieau (Eds.), *Institutional and financial incentives for social insurance.* Kluwer Press, pp. 1–24.

Garasky, S. & Barnow, B. S. (1992). Demonstration evaluations and cost neutrality: Using caseload models to determine the federal cost neutrality of New Jersey's REACH demonstration. *Journal of Policy Analysis and Management, 11*(3), 624–636.

Gilbert, L., Kamionka, T., & Lacroix, G. (2001). *The impact of government-sponsored training programs on the labour market transitions of disadvantaged men* (Working Paper 2001–15). Paris: Critical Research on Economic and Social Transformation.

Gouriéroux, C. & Monfort, A. (1991). Simulation based econometrics in models with heterogeneity. *Annales d'économie et de statistique, 20*(1), 69–107.

Gouriéroux, C. & Monfort, A. (1996). Simulation-based econometric methods. *Core Lectures.* Oxford: Oxford University Press.

Harknett, K. & Gennetian, L. A. (2001). *How an earnings supplement can affect the marital behaviour of welfare recipients: Evidence from the Self-Sufficiency Project* (SRDC Working Paper Series 01-01). Ottawa: Social Research and Demonstration Corporation.

Heckman, J. & Singer, B. (1984). A method for minimizing the distributional assumptions in econometric models for duration data. *Econometrica 52*(2), 271–320.

Heckman, J. J. (1992). Randomization and social policy evaluation. In F. C. Manski & I. Garfinkel (Eds.), *Evaluating welfare and training programs*. Cambridge: Harvard University Press.

Heckman, J. J. & Borjas, G. E. (1980). Does unemployment cause future unemployment? Definitions, questions and answers from a continuous time model of heterogeneity and state dependence. *Economica, 47*(187), 247–283.

Heckman, J. J., Lalonde, R. J., & Smith, J. A. (1999). The economics and econometrics of active labor market programs. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics.* Amsterdam: North-Holland.

Hotz, V. J. (1992). Designing an evaluation of the Job Training Partnership Act. In C. F. Manski & I. Garfinkel (Eds.), *Evaluating welfare and training programs*. Cambridge: Harvard University Press, chapter 2, pp. 76–114.

Kamionka, T. (1998). Simulated maximum likelihood estimation in transition models. *Econometrics Journal, 1,* C129–C153.

Klein, J. P. & Moeschberger, M. L. (1997). *Survival analysis.* New York: Springer-Verlag.

Lacroix, G. & Royer, J. (2001). *Vérification empirique de l'absence de biais de non-réponse à l'aide d'une procédure de test MC* [Mimeograph]. Université Laval.

Michalopoulos, C. & Hoy, T. (2001). *When financial work incentives pay for themselves: Interim findings from the Self-Sufficiency Project's applicant study.* Ottawa: Social Research and Demonstration Corporation.

Michalopoulos, C., Card, D., Gennetian, L. A., Harknett, K., & Robins, P. K. (2000). *The Self-Sufficiency Project at 36 months: Effects of a financial work incentive on employment and income.* Ottawa: Social Research and Demonstration Corporation.

Moffitt, R. A. (1992). Evaluation methods for program entry effects. In C. F. Manski & I. Garfinkel (Eds.), *Evaluating welfare and training programs.* Cambridge, MA: Harvard University Press, pp. 231–252.

Moffitt, R. A. (1996). The effect of employment and training programs on entry and exit from the welfare caseload. *Journal of Policy Analysis and Management, 15*(1), 32–50.

Morris, P. & Michalopoulos, C. (2000). *The Self-Sufficiency Project at 36 months: Effects on children of a program that increased parental employment and income.* Ottawa: Social Research and Demonstration Corporation.

Quets, G., Robins, P. K., Pan, E. C., Michalopoulos, C., & Card, D. (1999). *Does SSP Plus increase employment? The effect of adding services to the Self-Sufficiency Project's financial incentives.* Ottawa: Social Research and Demonstration Corporation.