Assessing the Extent of Randomization Bias in the Canadian Self-Sufficiency Experiment[†]

Thierry Kamionka[‡] Guy Lacroix[§]

Very Preliminary – Not to be quoted

March 2002

Abstract

In Canada, a policy aiming at helping single parents on social assistance become self-reliant was implemented on an experimental basis. The Self-Sufficiency Entry Effects Demonstration randomly selected a sample of 4,142 single parents who had applied for welfare between January 1994 and March 1995. It turned out only 3,315 agreed to be part of the experiment despite a 50% chance of receiving a generous, time-limited, earnings supplement conditional on finding a full-time jobs and leaving income assistance.

The purpose of this paper is to determine whether a refusal rate as high as 20% is likely to bias the measurement of the treatment effect. We compare the estimated impact of the program using experimental data only to those obtained using additional data on individuals not taking part in the experiment. 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 randomization bias in the data. When we correct for the bias, we find that estimates that rely on experimental data only significantly underestimate the true impact of the program.

[†] 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). They are also grateful for comments from Pierre Dubois, Jean-Pierre Florens and seminar participants at GREMAQ, Toulouse, and at the 15th annual conference of the European Society for Population Economics, Athens.

[‡] CNRS, CREST and IDEI (kamionka@ensae.fr).

[§] Département d’économique, Université Laval, CIRPÉE and CIRANO (guy.lacroix@ecn.ulaval.ca).

1 Introduction

In seeking to alleviate the problems that plague particularly disadvantaged groups when integrating 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 twenty years, the evaluation literature has generally found training programs to have had limited success in achieving these goals (see Heckman, LaLonde and Smith (1999) for a recent and detailed survey and Gilbert, Kamionka and 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 toward support programs. Yet, decrease in reliance has not generally translated into significant reductions in poverty rates. One may infer from such poor performance that training programs that were implemented over that period simply did not manage to increase productivity to a level that would make work a better alternative to social assistance.

Many governments have responded to such deceptive results by shying away from traditional training programs only 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 skills and attitudes that are necessary for self-reliance.

In Canada, a policy aiming at helping 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 found full-time jobs and left income assistance. SSP consists of two main studies: the SSP Recipients Demonstration (RD) and the SSP Entry Effects Demonstration (EED). The former focuses on welfare recipients who have been on welfare for at least a year. The latter focuses on newly enrolled recipients.

The RD began in 1992 and enrolled over 9,000 volunteers. About half were randomly offered the SSP program. The other half were not offered the supplement and constitute the experimental control group. The EED, 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 may entice some to remain longer on the rolls. The EED 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 the SSP are based on the Recipients Demonstration. Nearly all of them conclude that the program has had sizable impacts on exits from welfare (Michalopoulos, Card, Gennetian, Harknett and Robins (2000), Quets, Robins, Paan, Michalopoulos and Card (1999)). Others have found the program beneficial to children (Morris and Michalopoulos (2000)) and to have had ambiguous results on marital behaviour (Harknett and Gennetian (2001)).

There is little doubt the program has had significant impacts on individual behaviour. Because both the RD and the EED 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 the "treatment effects on the treated" only under a number of relatively stringent assumptions. One of those states that individuals taking part in the experiment constitute a true random sample of the population of interest. In other words, randomization bias is assumed away. There is little discussion of randomization bias in the literature partly because the data obtained from social experiments simply can not 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 behavioral disruptions have contaminated the estimated impacts.

The purpose of this paper is to document the extent of randomization bias in the SSP experiment and to propose a measure of the impact of such bias, if any. Our analysis focuses on the EED because refusal to participate was much higher in the EED that in the RD (20% *vs* 5%).[2] Our strategy is thus to compare the estimated impact of the program using experimental data only to those obtained using additional data on individuals not taking part in the experiment. Reasons for not participating are threefold. First, some recipients simply were not selected at baseline. This sample can be thought of as a legitimate control group for the purpose of the experiment. 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 Berlin, Bancroft, Card, Lin and Robins (1998) in finding evidence of delayed exits. Furthermore, we find strong evidence of randomization bias in the data. When we properly correct for the bias, we find that the estimates that rely on experimental data alone underestimate the true impact of the program.

The remainder of the paper is organized as follows. Section 2 provides a detailed description of the Entry Effects Demonstration. Section 3 describes the data on both participants and

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

[2] As many as 4,142 individuals were contacted for the EED. Yet, only 3,326 completed the baseline survey, and an additional 9 asked to be removed from the experiment after completing the survey. Thus the response rate is 80%.

non-participants in the EED. Non-parametric evidence on delayed exits is presented as well. Section 4 discusses the statistical model. In particular, the treatment of unobserved individual heterogeneity is discussed in details and its role in identifying treatment effects is highlighted. Section 5 reports our main findings. Finally, Section 6 concludes the paper.

2 The Entry Effects Demonstration

Economists have long recognized that policies that provide a conditional earnings supplement may have the unintended consequence of inducing some to modify their behaviour in order 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 nevertheless been shown to perform relatively well at predicting inflows and outflows from welfare caseloads (Garasky and Barnow (1992)).

The Self-Sufficiency Project was introduced in Canada in 1992. It aimed at measuring 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 potentially 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 currently on the rolls to delay their exit from welfare in order 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 Entry Effects Demonstration thus utilized 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 form 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 administrative records on income assistance from the British Columbia Ministry of Social Services. Immediately after the baseline interview, individuals were ran-

[3] To be considered as new entrants, applicants had not to have received IA in the six previous months. A significant minority (31%) had nevertheless received IA at some time in the two years prior to their current application (Berlin et al. (1998)).

domly assigned to either the program or the control group. Program members were sent a letter and brochure explaining their potential eligibility to an earnings supplement. They were reminded that they had to remain on welfare for at least 12 months to qualify for the supplement and that upon qualification, they had to find a full-time job within the next 12 months. They were also mailed a “reminder” six to seven months after their baseline interview.

2.1 Randomization Bias

Under ideal conditions, the classical randomization scheme used in the EED arguably is the best means by which to measure delayed exit effects and perhaps net program impact on welfare spells durations. According to Statistics Canada, though, as many as 20% of individuals who were originally selected did not complete the baseline interview. Non-response was partly due to the fact that some individuals had already left IA by the time they were contacted. Among those who were still on IA, many felt they would be off the rolls shortly and were reluctant to take part in an experiment designed for welfare participants. Still others might have refused to participate due to the intrusiveness of the experiment.[4]

The problems that arise due to randomization biases are best understood through a formal analysis. Consider three separate levels of selection as in Heckman (1992). First, the observed experimental sample is:

$$\begin{aligned} A &= 1, && \text{if an individual belong to the treatment group,} \\ A &= 0, && \text{otherwise.} \end{aligned}$$

This sample results from random assignment. Prior to assignment, an individual must decide whether to take part in the experiment. Let D^* be such that:

$$\begin{aligned} D^* &= 1, && \text{if an individual agrees to be subjected to randomization,} \\ D^* &= 0, && \text{otherwise.} \end{aligned}$$

Hence, all those for whom $D^* = 1$ will be assigned an experimental status $A = 1$ or $A = 0$. The purpose of randomizing at this stage is to avoid selection into $A = 1$ or $A = 0$ on the basis of unobservable characteristics. Most studies that use non-experimental data must tackle

[4] Fortin, Garneau, Lacroix, Lemieux and Montmarquette (1996) report that as many as a third of lone parents who received social assistance in Québec in 1994 were working in the underground economy, mainly as daycare workers. To the extent the same situation prevailed in British Columbia, it is conceivable that a number of refusals may have arisen because of the intrusiveness of the experiment.

this issue in one way or another (see LaLonde (1986) and Heckman, Ichimura, Smith and Todd (1998)). Yet, as stressed by Heckman and Smith (1995), Burtless (1995) and Heckman (1992), if refusal to take part in an experiment is correlated to unobservable characteristics, then there is no guarantee the data is void of any systematic biases.

Let Z be a set of observable individual characteristics and Ψ be such that $Z \in \Psi \Leftrightarrow D = 1$ and $Z \notin \Psi \Leftrightarrow D = 0$. Here, D is a selection variable such that whenever $D = 1$ the individual would have taken part in the program in the absence of randomization. Non-response bias occurs whenever $[Z \in \Psi \text{ or } D = 1] \nRightarrow D^* = 1$. As stressed above, many unobservable factors may lead someone to refuse to be subjected to an experiment. When using experimental data, it is customarily assumed that:

$$\Pr(D = 1|c) = \Pr(D^* = 1|c, p), \tag{1}$$

where c is the realization of the random variable C, $p = \Pr(A = 1)$, and Z is a subset of C, *i.e.* $Z \subseteq C$. This assumption simply states that randomization *per se* has no impact on the decision to participate in a given program.[5] Heckman (1992) has shown that this assumption implies the following:

$$\begin{aligned} F(y_1, c|A = 1) &= F(y_1, c|D^* = 1) = F(y_1, c|D = 1), & (2)\\ F(y_0, c|A = 0) &= F(y_0, c|D^* = 1) = F(y_0, c|D = 1), & (3) \end{aligned}$$

where $y(\cdot)$ is the realization of the random variable $Y(\cdot)$, the length of a welfare spell, say. Thus, Y_1 and Y_0 refer to the length of the spell of a member of the treatment and control groups, respectively, and $F(\cdot)$ represents the joint distribution of $Y(\cdot)$ and C. Equations (2) and (3) state that the distributions of welfare spells are void of selection biases and are not affected by randomization. Naturally, if equations (2) and (3) are true the following must hold:

$$E(Y_1|A = 1) - E(Y_0|A = 0) = E(\Delta|D = 1) \tag{4}$$

Thus a simple comparison of expected spell durations provides an unbiased estimator of the treatment effect on the treated. Our task in this paper is twofold. First, we wish to investigate if equation (3) holds true in the EED. Second, if the latter does not hold, we will seek to provide an estimator of the treatment effect that is void of randomization biases.[6]

[5] In the context of EED, it is probably fair to argue that randomization *per se* is fairly innocuous since being assigned to the control group is equivalent to not participating in the experiment. On the other hand, participation does involve significant intrusiveness costs alluded to earlier.

[6] Note that Assumption (1) is not necessary for the estimator in (4) to be unbiased. Thus even if we find evidence of randomization bias, it does not follow that the estimator of the treatment effect will change much if we correct for such bias.

The assumption is not generally verified when the distribution of the output on the labor market depends on unobserved characteristics and the distribution of these unobservable factors depends on the decision to participate (namely D).

3 Data

Statistics Canada, the data collection contractor, provided us individual IA histories on all those participating into the experiment. The histories start at random assignment and last over 65 months. Such long histories allows us to investigate both entry effects as well as the impact of the income supplement on exits from welfare.

As mentioned earlier, experimental data is intrinsically incapable of detecting randomization biases of any kind. Statistics Canada thus agreed to provide us two complementary sets of data. The first is a random sample of individuals who were not sampled at baseline. The second is the complete sample of those who were selected but who could not be reached by the time interviewers tried to contact them.[7]

The sampling scheme and the data at our disposal are illustrated in Figure 1. Those in the original sample were asked to be part of the experiment. As many as 3,315 individuals agreed to be subjected to randomization ($D^* = 1$). The randomization procedure yielded the experimental treatment and control groups (henceforth groups *A* and *B*, respectively). The group referred to as *C* includes those who were selected at baseline but could not be contacted for various reasons. These individuals have not been subjected to randomization. It is not known *a priori* what fraction of the group would have accepted the invitation to be part of the experiment. Finally, approximately 20% of those originally selected refused to be part of the experiment. Unfortunately, we have no information on this particular group. To the extent these individuals have unobserved characteristics that are distinct from those of groups *A* and *B*, it is very likely the treatment effect obtained from a simple comparison between the experimental groups will be somewhat biased. To ascertain this possibility, we were provided a sample of over 3,073 individuals who were either not selected at baseline or who refused to participate in the experiment. We refer to this sample as group *D*. Those who have refused are not identifiable in the data. As such Group *D* is a complex mix of groups *A*, *B* and *C*. Indeed, among those who were not selected, 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 into the experiment. Under the null assumption of no randomization bias, groups *B* and *D* should behave in a similar manner. If it is found

[7] As we will show in Section 3, as many as a third of those who could not be contacted at baseline would have qualified for the supplement had they been contacted.

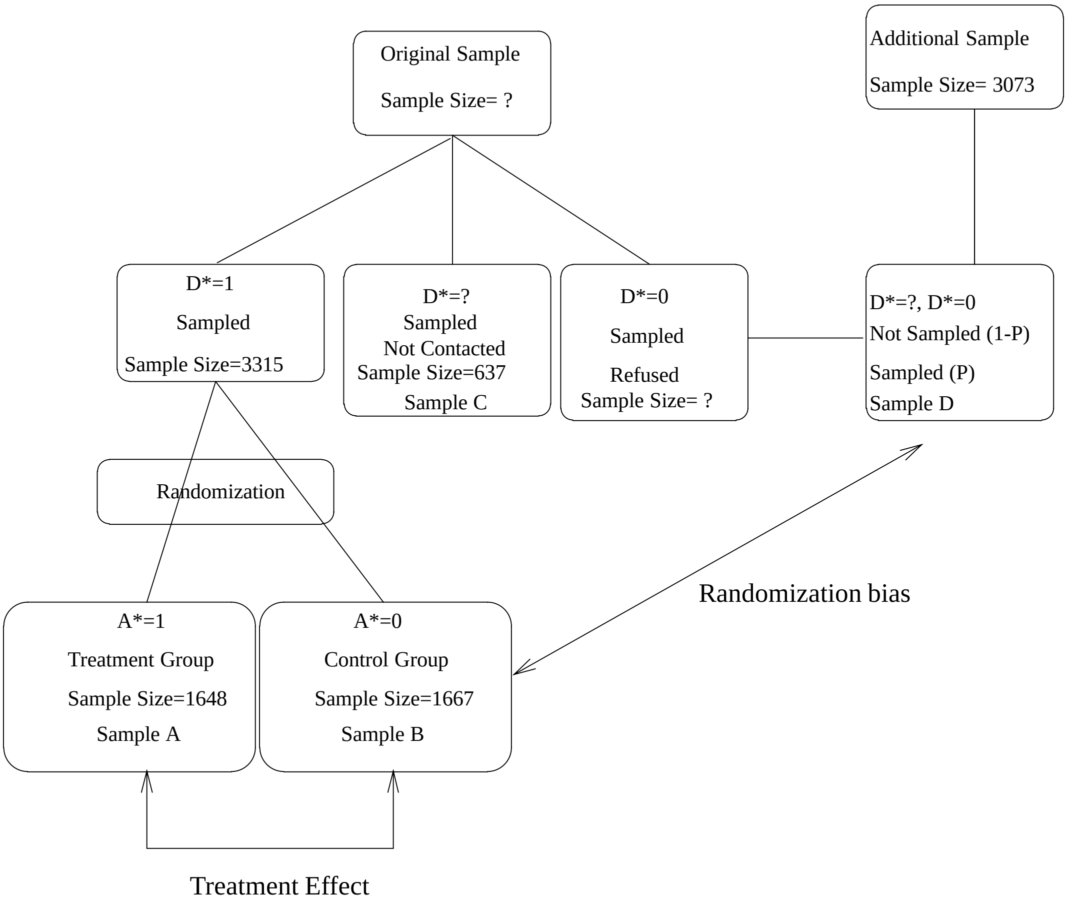


Figure 1: Randomization Scheme

that there are systematic differences, it will be necessary to investigate whether the treatment effect is biased.

3.1 Descriptive Statistics

Table 1 provides sample descriptive statistics.[8] The table is divided into four different columns, each corresponding to a particular sample. The first two columns, denoted group *A* and group

[8] The administrative files contain more information on individual characteristics than those reported in the table. To insure confidentiality of IA claimants, we were only provided information on characteristics reported in the table.

B, refer to the original experimental samples. Individuals in *A* were offered the SSP treatment whereas those in *B* were not. Group *C* refers to individuals who were selected for the experiment but who could not be contacted at baseline. Finally, Group *D* refers essentially to individuals who were not selected at baseline or who refused to take part in the experiment.

The first two columns show that the experimental treatment and control groups are very similar in terms of observable characteristics. This is not surprising since treatment is randomly assigned among those who agree 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 that the mean IA spell duration is relatively the same for individuals in samples *A*, *B* and *D*. Those in sample *C* have a significantly shorter mean and median durations. Finally, note that although we observe individual IA histories for over 65 months, more than 9.6% of all spells are censored.

To better ascertain the extent to which observable characteristics differ between samples *A*, *B*, *C* and *D*, we report simple probit 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 merged 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 only 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.[9]

3.2 Non-Parametric Evidence

Recall from Section 2 that the Entry Effects Demonstration aimed at determining whether IA applicants might be induced to delay their exits from welfare in order to qualify for the (relatively) generous earnings supplement. In order 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 in order to receive the supplement. Those in sample *B* continued to receive the standard IA benefit.

[9] 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.

Behavioural response to the EED is best investigated through the use of hazard and survival curves. Figure 7 in appendix plots smoothed hazard rates of IA spells for the experimental samples. The first noteworthy feature of the figure is that recipients appear sensitive to the parameters of the EED. Indeed, the hazard rates increase slightly in the first few months upon entry into IA and then decrease significantly up until the 11^{th} month.[10] Upon qualifying for the supplement, the hazard function of sample *A* increases significantly above that of sample *B*, as expected. Those who have not found a job by month 24 are automatically disqualified. Hence the hazard function peaks at month 24 and declines rapidly afterward.[11]

Delayed exit behaviour is evidenced by the difference between the hazard functions during the first 12 months. Indeed, the hazard function of sample *A* lies below that of sample *B* during the first 12 months, then crosses it and remains above for the next 24 months or so. The underlying survival functions are plotted in Figure 8. Not surprisingly, the survival function of sample *A* lies above that of sample *B* up until month sixteen.

Based on these figures, it seems reasonable to claim that the earnings supplement first induces individuals to delay their exits in the first 12 months and then provides a relatively strong incentive to leave IA. It is worth investigating though whether these differences are statistically significant. In order to do this, we turn to standard statistical tests. It can be shown that the surface below the survival function between $[0, \infty[$ is equal to the mean duration of IA spells. Likewise, it can be show that the estimated mean duration restricted to the interval $[0, \tau]$ is[12]

$$\hat{\mu}_\tau = \int_0^\tau \hat{S}(t)dt, \tag{5}$$

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

$$\hat{V}\left[\hat{\mu}_\tau\right] = \sum_{i=1}^{T}\left[\int_{t_i}^{\tau} \hat{S}(t)dt\right]^2 \frac{d_i}{Y_i(Y_i - d_i)} \tag{6}$$

where T is the number of distinct intervals over $[0, \tau]$, d_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. This estimator allows us to compare the mean durations of IA spells of samples *A* and B over the interval $[0, 12]$. Estimators can also be computed for other intervals. Table 3 reports test results for three such intervals. The first column reports statistics for the first 12 months.

[10] 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 and Lacroix (2002) and Fougère, Fortin and Lacroix (2002).

[11] The hazard function of sample *B* increases slightly between months 34 and 38. It is not clear what causes this. Since there are more than 500 observations left in sample *B* at month 31, it cannot be a statistical artefact. Further investigation into that matter certainly seems warranted.

[12] See Klein and Moeschberger (1997) for a formal derivation.

The mean duration of sample *A* is found to be approximately 2.5% greater than that of sample *B* although the difference is not statistically different. This is similar to the findings of Berlin et al. (1998) who report an average impact of approximately 3.0%. Column (2) of the table reports the mean durations between months 12 and 65. This time, sample *A* has a much shorter mean spell duration. It is estimated to be approximately 12.6% shorter and the difference is highly significant. Finally, according to column (3) sample *A* has a 6.78% shorter mean spell duration than sample *B* over the whole 65 month period.

The estimators reported in Table 3 are equivalent to standard difference estimators used in most studies that are based on experimental data. Even though our estimates do not account for individual characteristics, it is very unlikely the program impact will be affected by such variables given the results of Tables 2. The more interesting question that must be addressed is whether our estimates are plagued by randomization biases. Before we address this question formally, we will present informal evidence that such biases are be present in the data.

Figure 9 plots the smoothed hazards of samples *B* and *D*. Notice that the hazard function of sample *D* almost always lies either above or is very close to that of sample *B*. The underlying survival curves are plotted in Figure 10 along with that of sample *C*. The figure shows that sample *B*'s survival function lies above that of sample *D*. 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*. This is not very surprising given that *D* includes individuals would have been in sample *C* had they been contacted at baseline. These individuals have very short spells. Yet, according to their survival curve as many as a third would have qualified for the earnings supplement had they been contacted.

The above discussion has shown that the experimental control group suffers from non-response bias. It does not necessarily follow that the comparison between samples *A* and *B* necessarily lead to biased program impact. 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*. In order to measure the program impact correctly, non-response must be modeled explicitly and accounted for in a regression framework.

4 Modeling Individual Spell Durations

In order to derive an appropriate estimator of the treatment effect, non-response and randomization biases must be explicitly taken into account. The framework within which the experiment took place is illustrated in Figure 2, 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 who in the population are not willing *a priori* to partici-

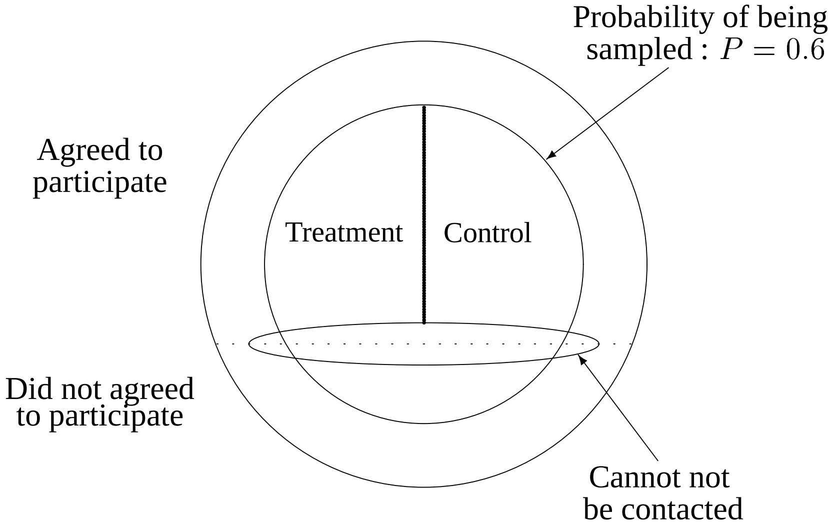

Probability of being sampled : $P = 0.6$
Agreed to participate
Treatment
Control
Did not agreed to participate
Cannot not be contacted

Figure 2: Welfare Applicants.

pate in such an experiment are located below the dashed line. Likewise, those who could not be contacted are located in the ellipse. Among the latter, a 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 all have accepted 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 surface between the inner and outer circles is the set of applicants who were not selected at baseline. This set can be broken down in sets similar to those of the experimental samples: acceptance, refusal, contacted, non-contacted, *etc.*

Our task is to model all the information that is available in Figure 2. In order to do this, we need to determine precisely the probability of belonging to the experimental samples. The original experimental samples comprised 3,383 individuals (1677 in the treatment group and 1706 in the control group). It was later discovered that 59 individuals did not meet one of the three criteria to be included in the study. Furthermore, five control group members and four treatment group members withdrew from the study and requested that none of their data be used for research purposes.[13] According to Statistics Canada, the experimental samples represented 45% of all claimants over the enrollment period.[14] If we consider those who could not be contacted as well as those who refused to participate in the experiment, then we can easily establish that the average probability of being sampled each month was 62.5%. We will thus consider that each applicant faces a probability $p = 0.6$ of being sampled.

[13] See Michalopoulos and T.Hoy (2001) for the details.

[14] This information was provided to us through private communications.

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

In order to model individual contributions to the likelihood function, we need to define a number of dummy variables. Thus let:

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

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

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

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

Finally, let y be a realization of the experiment:

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

where u is the duration of a welfare spell.[15]

Which arguments of $y(\cdot)$ are observable depend on which set an individual belongs to. Only T and U are observable for all individuals.[16] Thus, for those in *A* we know that they have been sampled in the experiment ($e = 1$), that they have agreed to participate ($a = 1$), that they could be contacted ($r = 1$) and are eligible for the supplement ($t = 1$). The table above summarizes the realizations of the random variables according to group membership.

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

[16] The welfare duration are right censored at 65 months.

4.1 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. In order to simplify the presentation, we assume that 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

$$l(\theta) = \sum_{j=1}^{J} f(y \mid x; \nu_j; \theta)\, \pi_j, \tag{7}$$

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

$$l(\theta) = \int_S f(y \mid x; \nu; \theta)\; g(\nu; \gamma)\; d\,\nu, \tag{8}$$

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.

4.2 Modeling Individual Contributions

In this section we focus on the conditional distributions of variable A, R and U. Recall that the probability of being sampled in the experiment is p and that 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 that the individual agrees to participate in the experiment. We will assume that

$$z(x, \nu) = \text{Prob}[A^* \geq 0 \mid x; \nu], \tag{9}$$

where

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

where ϵ_a is a normal random variable with mean zero 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 that the individual cannot be contacted. We assume

$$\phi(x, \nu, a) = \text{Prob}[R^* \geq 0 \mid x; a; \nu], \tag{10}$$

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 that ϵ_r is a normal random variable with mean zero and variance equal to 1. For simplicity, we further assume that ϵ_a, ϵ_r and ν are independent.

Finally, let $q(e, a, r)$ denote the conditional probability that 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 that:

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

Hence, an individual can be assigned to the treatment group if and only if he/she has been sampled in the experiment, has 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

$$\ell_\nu(\theta) = p \; z(x, \nu) \; (1 - \phi(x, a, \nu)) \; 0.5 \; f(u \mid x; a = 1; r = 1; t = 1; \nu; \theta), \tag{11}$$

if the individual belongs to group *A*;

$$\ell_\nu(\theta) = p \; z(x, \nu) \; (1 - \phi(x, a, \nu)) \; 0.5 \; f(u \mid x; a = 1; r = 1; t = 0; \nu; \theta), \tag{12}$$

if the individual is in group *B*;

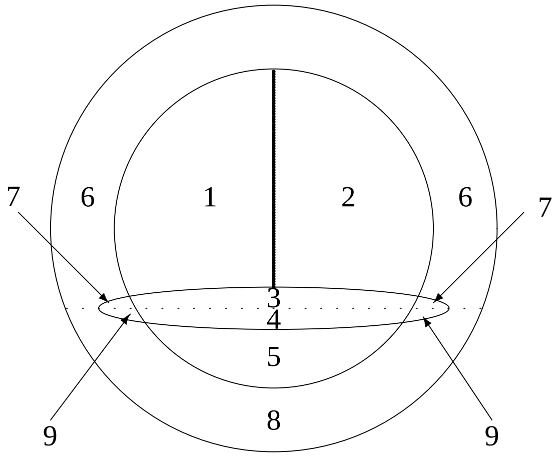


Figure 3: Welfare Applicants.

$$\begin{aligned}\ell_\nu(\theta) &= p\; z(x,\nu)\; \phi(x,a,\nu)\; f(u \mid x; a=1; r=0; t=0; \nu; \theta), \\ &+ p\; (1{-}z(x,\nu))\; \phi(x,a,\nu)\; f(u \mid x; a=0; r=0; t=0; \nu; \theta),\end{aligned} \tag{13}$$

if the individual is in group *C*;

and

$$\begin{aligned}\ell_\nu(\theta) &= p\; (1{-}z(x,\nu))\; (1{-}\phi(x,a,\nu))\; f(u \mid x; a=0; r=1; t=0; \nu; \theta), \\ &+ (1{-}p)\; z(x,\nu)\; (1{-}\phi(x,a,\nu))\; f(u \mid x; a=1; r=1; t=0; \nu; \theta), \\ &+ (1{-}p)\; z(x,\nu)\; \phi(x,a,\nu)\; f(u \mid x; a=1; r=0; t=0; \nu; \theta), \\ &+ (1{-}p)\; (1{-}z(x,\nu))\; (1{-}\phi(x,a,\nu))\; f(u \mid x; a=0; r=1; t=0; \nu; \theta), \\ &+ (1{-}p)\; (1{-}z(x,\nu))\; \phi(x,a,\nu)\; f(u \mid x; a=0; r=0; t=0; \nu; \theta),\end{aligned} \tag{14}$$

if the individual belongs to group *D*.

The contribution of each group to the likelihood function is indicated in Figure 3. Thus groups *A* and *B* contribute sections 1 and 2 (equations (11) and (12), respectively). Likewise, group *C* (equation (13)) corresponds to sections 3 and 4. Group *D* (equation (14)) 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

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

4.3 Unobserved heterogeneity

Estimation of the parameters by means of maximum likelihood requires that we specify the distribution of the unobserved heterogeneity terms. We will first approximate arbitrary continuous distributions using a finite number of mass points (see Heckman and Singer (1984)). Next we will investigate the robustness of the slope parameters using various continuous distributions.

1. *Discrete distributions*

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

   $$\text{Prob}[V=v] = \begin{cases} p_0, & \text{if } v=\nu_0, \\ (1-p_0), & \text{if } v=-\nu_0, \end{cases} \tag{15}$$

   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 zero and variance 1.

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

2. *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 (the pdf of normal or student distributions, for example).

4.4 Specification of conditional hazard function

The conditional hazard function for welfare durations is given by

$$h(u \mid x;a;r;t;\nu;\theta) = h_0(u;\alpha)\ \varphi(x;a;r;t;\beta_d)\ \exp(-\nu), \tag{16}$$

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 that:

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

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.

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.

4.5 Estimation

We consider two alternative specifications for the unobserved heterogeneity distribution.

1. *Discrete Distribution*

   The log likelihood is

$$\log(L(\theta)) = \sum_{i=1}^{N} \log(l_i(\theta)), \tag{17}$$

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 (7), and where N is the sample size.

In equation (7) π_j is set equal to[17]

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

where $\pi_1 = \text{Prob}[V = \nu_0]$, $\pi_2 = \text{Prob}[V = -\nu_0]$ and $\nu_0 \in \mathbb{R}$ 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. π_1 represents the probability that the unobserved term V takes the value ν_0 ($\pi_2 = 1 - \pi_1$).

2. *Continuous Distribution*

The model includes an 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 (8), where $S = \mathbb{R}^+$. The log-likelihood is given by equation (17), where $l_i(\theta)$ is the contribution to the likelihood of the sequence y_i.[18] 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 that

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

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 and Monfort (1991, 1996)). Under these conditions, this estimator has the same asymptotic distribution as the standard ML estimator. Following Kamionka (1998) and Gilbert et al. (2001)

[17] See section 4.1.

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

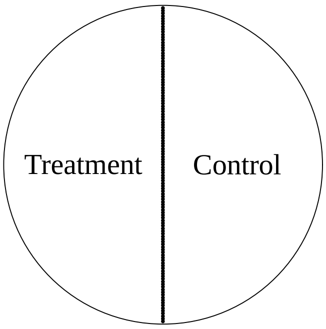


Figure 4: Participants to the experience that could be contacted.

we have used 40 draws from the random distributions when estimating the models. Using as few as 30 draws yielded essentially the same parameter estimates.

4.6 Incomplete Information Schemes

It is possible to examine the impact of the randomization and 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 parameters.

1. *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 4).

   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*;

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

   if the individual belongs to *B*.

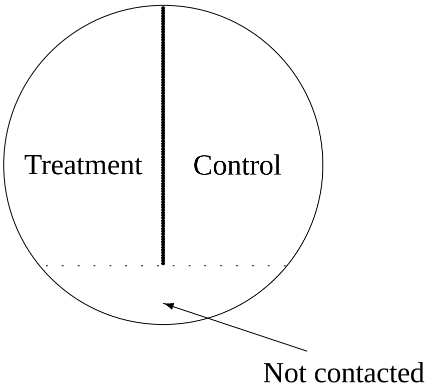


Figure 5: Participants to the experiment.

The conditional distribution of the welfare durations corresponds to the hazard function (16), where $\delta_a = \delta_r = 0$ (here a and r are set equal to arbitrary values in the conditional distribution of the welfare duration).

2. *Participants to 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 5). 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*;

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

if the individual belongs to *C*;

Here, $\phi(\nu, x)$ denotes the conditional probability that the individual could not be contacted and is defined as in the context of a complete information scheme (see equation (10)), where $\xi_a = 0$ (here a is fixed to an arbitrary value in this equation and in the expression of the conditional hazard function).

The expression of the conditional hazard function of the welfare durations is given by the equation (16) where $\delta_a = 0$.

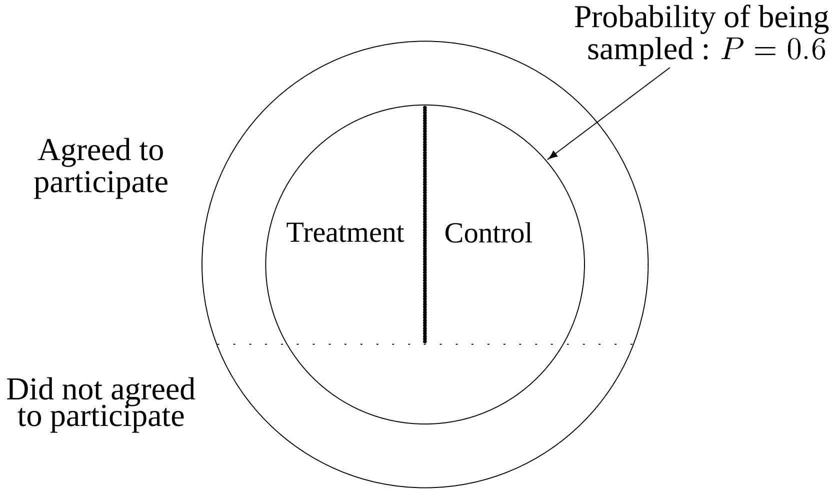

Probability of being sampled : $P = 0.6$
Agreed to participate
Treatment
Control
Did not agreed to participate

Figure 6: Selected and Non-Selected welfare applicants.

3. *Selected and non-selected welfare applicants*

   Here, each individual contributes a sequence $y = (e, a, t, u)$ to the likelihood function. Those that were selected at baseline have agreed to participate in the experiment. Those who were not selected may or may not have agreed (see figure 6).

   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 the *B*;

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

   if the individual belongs to *D*.

   Here, $z(x,\nu)$ is the conditional probability that 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 (9)).

   The expression of the conditional hazard function of the welfare durations is given by equation (16), where $\delta_r = 0$ (r, for convenience, is fixed to an arbitrary value in the expression of the conditional hazard).

5 Results

Estimation results using non-parametric heterogeneity are reported in Table 4 (see equation (17)). 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 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 that 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%. This estimate is quite similar to that reported in section 3.2 where it was found that the treatment group had a 6.78% shorter mean duration.

Column 2 of the table reports the results using groups *A*, *B*, and *C* (see Figure 5). The baseline hazard function is decreasing with duration. Spell duration increases with age and the number of children. Furthermore, women have longer spell durations than men. The impact of the treatment is very similar to that of column (1) despite the fact that the regression now accounts for the fact that individuals were either contacted or not at baseline. Incidentally, the parameter estimate of the contact binary variable is positive and significantly different from zero. This is consistent with the observation that individuals in sample *C* have significantly shorter spells (see Table 1).

The third panel of the table reports the parameter estimates of the probability of being contacted at baseline. It is found that the probability is increasing with age and the number of children. Women are also more likely to be contacted than men. These results are consistent with those obtained for descriptive statistics on sample *C* (see Table 1).

Column 3 of the table reports the results using groups *A*, *B*, and *D* (see Figure 6). Contrary to the previous cases, the conditional hazard function is increasing with duration. In fact, all regressions that include group *D* find an increasing conditional hazard function. 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* and *B* (column 1 *vs* column 3). The parameter of the acceptance binary variable is positive and statistically significant. Thus among the individuals that could be contacted *a priori*, those who decided to participate have longer mean spell duration. The treatment effect is now nearly four times greater than the one obtained using groups *A* and *B*. Consequently, omission of the participation decision significantly biases the effect of the earning supplement on the exits from welfare. The second panel of the table reports the parameters of the conditional probability of agreeing to participate in the experiment. Unfortunately, not a single parameter is statistically significant in this specification.

Columns 4 and 5 of the table reports the results using groups *A*, *B*, *C* and *D* (see Figure 2). The model of column 5 corresponds to the restricted model alluded to in section 4.3. The probability of the discrete unobserved heterogeneity term is thus fixed to 0.5 and a parameter, γ, is interacted with the individual random effect in the participation equation.

The parameter estimates of column 4 show that the conditional hazard function is increasing with duration. The sign of the slope parameters are similar to those obtained using samples *A* and *B* only. 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 that the probability of being contacted is increasing with age, the number of children and is higher for women than for men. The parameters are very similar those obtained using groups *A*, *B* and *C*. Furthermore, the probability is significantly greater for those who are willing to participate *ex ante*. Finally, note that the probability of agreeing to participate increases with age and that the parameter estimate is statistically significant at 10%.

The parameter estimates of the restricted model (column 5) are very similar to those of column 4, with the exception that the parameter estimate of “Accept” increases significantly and that of “Contacted” is no longer significant. This is possibly due to the fact that parameter estimate of γ is significantly different from 1. Thus the impact of the individual random effect in the “acceptance” is probably not the same as in the “contacted” equations.

The results presented in Table 4 are based on a rather restrictive non-parametric specification for the unobserved heterogeneity. Previous research has shown that the slope parameters of duration models are usually rather insensitive to particular distributional assumptions (see Heckman and Borjas (1980), Bonnal, Fougère and 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.

Table 5 only reports results using groups *A*, *B*, *C* and *D*. The parameter estimates are thus comparable to those in column 4 of Table 4. Each column of in the table corresponds to a particular parametric distribution.[19] The table is split vertically to underline the fact that the results are relatively homogeneous within two separate categories. In the first, the results based on the exponential, the gamma and the log-normal distribution yield results that are

[19] The student-t distribution is based on 15 degrees of freedom.

very similar to those of the corresponding non-parametric distribution. Indeed, the treatment effect is still sizable and the mean spell duration of those who could be contacted or agreed to participate is considerably longer. Furthermore, the parameter estimates of the the two latent equations are very similar to those of Table 4.

The second category of results are based upon the logistic, the normal and the student distributions. The slope parameter using these distributions are relatively similar, but they differ significantly from those of the non-parametric specification. Indeed, in all three cases the treatment effect is now found to be statistically not significant. It is also found that those who would agree to participate in the experiment have, *ceteris paribus*, shorter mean spell durations. This is incompatible both with intuition and with simple descriptive statistics on spell durations. Turning to the latent equations, all three specifications find a positive relation between age and the likelihood of accepting to participate in the experiment, which runs counter to the results of the other specifications. Finally, the parameter estimates of the "contact" equation are qualitatively similar to those of the first category. But in most cases, the estimates are much larger in magnitude.

The results of the second category of specifications are quite at odds with all other results presented in Tables 4 and 5. This suggests that the distribution functions are probably incompatible with the data. Further investigation into that matter certainly is warranted.

6 Conclusion

Over the past twenty years experimental designs have become the preferred mean of many by which to evaluate 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 it generates (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 ongoing demonstrations are contaminated by these potential problems. The evidence brought to bear is almost always indirect or inferential at best.

In Canada, a policy aiming at helping single parents on social assistance become self-reliant was implemented on an experimental basis. The Self-Sufficiency Entry Effects Demonstration (EED) focused on newly enrolled recipients. The EED randomly selected a sample of 4,142 single parents who had applied for welfare between January 1994 and March 1995. It turned out only 3,315 agreed to be part of the experiment despite a 50% 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 refusal rate as high as 20% 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 only to those obtained using additional data on individuals not taking part in the experiment. 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. Thirdly, 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 randomization bias in the data. When we correct for the bias, we find that estimates that rely on experimental data only 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.

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

References

Berlin, G., W. Bancroft, D. Card, W. Lin, and P. K. Robins (1998) ‘Do work incentives have unintended consequences ? Measuring "entry effects" in the Self-sufficiency project.’ *Working Paper*, SRDC

Bonnal, L., D. Fougère, and A. Sérandon (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., W. Hollander, and R. M. Korwar (1974) ‘Nonparametric tests of independence for censured data, with applications to heart transplant studies.’ In *Reliability and Biometry: Statistical Analysis of Lifelength*, ed. F. Proschan and R. J. Serling (Philadelphia: SIAM) pp. 327–354

Burtless, G. (1995) ‘The case for randomized field trials in economic and policy research.’ *Journal of Economic Perspective* 9(2), 63–84

Drolet, S., B. Fortin, and G. Lacroix (2002) ‘Welfare benefits and the duration of welfare spells: Evidence from a natural experiment in Canada.’ mimeo, Department of Economics, Université Laval

Fortin, B., G. Garneau, G. Lacroix, T. Lemieux, and C. Montmarquette (1996) *L’économie souterraine au Québec: Mythes et réalités* (Les presses de l’Université Laval)

Fougère, D., B. Fortin, and G. Lacroix (2002) ‘The effects of welfare benefits on the duration of welfare spells.’ In *Institutional and Financial Incentives for Social Insurance*, ed. C. d’Aspremont, V. Ginsburg, H. Sneessens, and F. Spinnewyn (Kluwer Academic Press)

Garasky, S., and B. S. Barnow (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., T. Kamionka, and G. Lacroix (2001) ‘The impact of government-sponsored training programs on the labour market transitions of disavantaged men.’ *Working Paper 2001–15*, CREST, Paris

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

— (1996) *Simulation-Based Econometric Methods* Core Lectures (Oxford University Press)

Harknett, K., and L. A. Gennetian (2001) ‘How an earnings supplement can affect the marital behaviour of welfare recipients: Evidence from the Self-sufficiency project.’ *Working Paper*, SRDC

Heckman, J., and B. Singer (1984) ‘A method for minimizing the distributional assumptions in econometric models for duration data.’ *Econometrica* pp. 271–320

Heckman, J. J. (1992) ‘Randomization and social policy evaluation.’ In *Evaluating Welfare and Training Programs*, ed. F. C. Manski and I. Garfinkel (Harvard University Press) chapter 5

Heckman, J. J., H. Ichimura, J. Smith, and P. Todd (1998) ‘Characterizing selection bias using experimental data.’ *Econometrica* 66, 1017–1098

Heckman, J.J., and G.E. Borjas (1980) ‘Does unemplyment cause future unemployment ? definitions, questions and answers from a continuous time model of heterogeneity and state dependence.’ *Economica* pp. 247–283

Heckman, J.J., and J.A. Smith (1995) ‘Assessing the case for social experiments.’ *Journal of Economic Perspective* 9(2), 85–110

Heckman, J.J., R.J. LaLonde, and J.A. Smith (1999) ‘The economics and econometrics of active labor market programs.’ In *Handbook of Labor Economics*, ed. O. Ashenfelter and D. Card (North-Holland)

Hotz, V. J. (1992) ‘Designing an evaluation of the Job Training Partnership Act.’ In *Evaluating Welfare and Training Programs*, ed. F. C. Manski and I. Garfinkel (Harvard University Press) chapter 2

Kamionka, T. (1998) ‘Simulated maximum likelihood estimation in transition models.’ *Econometrics Journal* 1, C129–C153

Klein, J. P., and M. L. Moeschberger (1997) *Statistics for Biology and Health* (Springer)

LaLonde, R. J. (1986) ‘Evaluating the econometric evaluations of training programs with experimental data.’ *American Economic Review* 76(4), 604–620

Michalopoulos, C., and T.Hoy (2001) ‘When financial work incentives pay for themselves: Interim findings from the Self-sufficiency project’s applicant study.’ *Working Paper*, SRDC

Michalopoulos, C., D. Card, L. A. Gennetian, K. Harknett, and P. K. Robins (2000) ‘The Self-sufficiency project at 36 months: Effects of a financial work incentive on employment and income.’ *Working Paper*, SRDC

Moffitt, R. A. (1992) ‘Evaluation methods for program entry effects.’ In *Evaluating Welfare and Training Programs*, ed. C. F. Manski and I. Garfinkel (Harvard University Press) chapter 6, pp. 231–152

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

Morris, P., and C. Michalopoulos (2000) ‘The Self-sufficiency project at 36 months: Effects on children of a program that increased parental employment and income.’ *Working Paper*, SRDC

Quets, G., P. K. Robins, E. C. Paan, C. Michalopoulos, and D. Card (1999) ‘Does SSP Plus increase employment ? The effect of adding services to the Self-sufficiency project’s financial incentives.’ *Working Paper*, SRDC

Table 1: Descriptive Statistics

Variable	*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	1648	1667	637	3073

[†] Estimated from Kaplan-Meir survival rates and tail corrections proposed by Brown, Hollander and Korwar (1974)

Table 2: Probit Regressions

Variable	*A vs B*	*A vs D*	*B vs D*	*C vs D*
Intercept	0.094	-0.435*	-0.523*	-0.423*
	(0.134)	(0.113)	(0.114)	(0.143)
Sex (Women=1)	-0.121	-0.013	0.106	-0.215*
	(0.077)	(0.063)	(0.066)	(0.077)
Children	-0.041	-0.011	0.029	-0.057**
	(0.027)	(0.024)	(0.023)	(0.031)
Age	0.002	0.002	0.001	-0.007*
	(0.003)	(0.003)	(0.003)	(0.003)
Observations	3315	4721	4740	3710
Log-Likelihood	-2294.5	-3053.3	-3071.5	-1693.6

* Statistically significant at 5% or better. ** Statistically significant at 10% or better.

Table 3: Asymptotic Means Tests

Interval	0–12		12–65		1–65	
	A	*B*	*A*	*B*	*A*	*B*
Mean duration	8.691	8.481	11.508	13.169	20.277	21.752
Variance	(0.005)	(0.026)	(0.081)	(0.109)	(0.227)	(0.270)
$H_0 : \hat{\mu}^A = \hat{\mu}^B (\chi^2_{0.05}(1))$	1.429		14.467		4.384	

Table 4: Maximum Likelihood Estimates: Non-Parametric Heterogeneity

Parameter	A + B	A+B+C	A+B+D	A+B+	A+B+
Estimates				C+D	C+D
Duration					
α	0.873	0.921	1.332	1.370	1.433
	(0.013)	(0.015)	(0.026)	(0.028)	(0.029)
ν		1.025	1.358	-1.378	-1.349
		(0.059)	(0.040)	(0.035)	(0.036)
Intercept	2.753	2.145	3.087	2.671	2.687
	(0.120)	(0.141)	(0.220)	(0.120)	(0.127)
Sex (Women=1)	0.198	0.200	0.184	0.207	0.199
	(0.064)	(0.059)	(0.062)	(0.049)	(0.054)
Age	-0.697	-1.044	-0.715	-0.532	-0.556
	(0.240)	(0.239)	(0.232)	(0.187)	(0.210)
Children	0.203	0.203	0.248	0.229	0.283
	(0.052)	(0.053)	(0.054)	(0.046)	(0.053)
Treatment	-0.075	-0.076	-0.247	-0.242	-0.254
	(0.037)	(0.037)	(0.042)	(0.037)	(0.039)
Accept			0.478	0.831	1.489
			(0.267)	(0.095)	(0.138)
Contacted		1.820		0.134	-0.138
		(0.143)		(0.068)	(0.134)
Acceptance					
Intercept			1.783	1.537	0.866
			(0.418)	(0.226)	(0.183)
Sex (Women=1)			0.214	0.086	0.186
			(0.171)	(0.119)	(0.096)
Age			0.330	0.851	0.276
			(0.977)	(0.505)	(0.404)
Children			-0.077	-0.009	0.038
			(0.160)	(0.109)	(0.090)
Not Contacted					
Intercept		0.427		2.461	2.447
		(0.188)		(0.229)	(0.282)
Sex (Women=1)		-0.277		-0.320	-0.258
		(0.098)		(0.111)	(0.151)
Age		-1.524		-0.740	-1.186
		(0.408)		(0.447)	(0.581)
Children		-0.123		-0.190	-0.196
		(0.091)		(0.101)	(0.136)
Accepted				-4.519	-4.567
				(0.146)	(0.169)
Probability		-1.229	0.295	-0.221	
		(0.055)	(0.033)	(0.025)	
γ					0.387
					(0.054)
Likelihood	-12,391.5	-18,444.5	-30,137.7	-34,236.0	-34,265.4

Table 5: Maximum Likelihood Estimates: Parametric Heterogeneity

Parameter	Exponential	Gamma	Log-	Logistic	Normal	Student
Estimates			Normal			(15)
Duration						
α	1.048	1.035	0.983	1.650	1.672	1.533
	(0.020)	(0.020)	(0.016)	(0.046)	(0.050)	(0.045)
σ	-0.424	-0.497	-1.499	0.407	0.903	0.704
	(0.073)	(0.074)	(0.107)	(0.043)	(0.044)	(0.047)
Intercept	1.493	1.458	1.293	5.388	5.024	4.736
	(0.137)	(0.134)	(0.135)	(0.302)	(0.270)	(0.278)
Sex (Women=1)	0.272	0.277	0.222	0.464	0.497	0.435
	(0.053)	(0.052)	(0.047)	(0.101)	(0.087)	(0.086)
Age	-0.988	-0.900	-0.716	-2.216	-1.774	-2.103
	(0.213)	(0.207)	(0.190)	(0.384)	(0.362)	(0.344)
Children	0.202	0.196	0.187	0.260	0.257	0.352
	(0.047)	(0.046)	(0.043)	(0.086)	(0.083)	(0.071)
Treatment	-0.176	-0.187	-0.186	-0.078	-0.081	-0.087
	(0.037)	(0.037)	(0.033)	(0.063)	(0.063)	(0.058)
Accept	1.495	1.560	1.727	-3.814	-2.672	-2.336
	(0.125)	(0.115)	(0.115)	(0.283)	(0.240)	(0.228)
Contacted	0.431	0.336	0.196	3.261	2.326	2.122
	(0.160)	(0.141)	(0.160)	(0.171)	(0.144)	(0.125)
Acceptance						
Intercept	1.043	1.046	0.978	2.253	1.860	1.977
	(0.187)	(0.184)	(0.182)	(0.422)	(0.357)	(0.366)
Sex (Women=1)	0.180	0.166	0.202	0.435	0.564	0.416
	(0.100)	(0.098)	(0.094)	(0.229)	(0.187)	(0.193)
Age	-0.049	-0.087	-0.162	1.683	2.226	1.144
	(0.419)	(0.413)	(0.407)	(0.894)	(0.806)	(0.812)
Children	0.031	0.029	0.026	0.061	-0.031	0.220
	(0.090)	(0.089)	(0.087)	(0.209)	(0.183)	(0.188)
Not Contacted						
Intercept	1.328	1.288	1.039	6.099	5.365	5.566
	(0.245)	(0.243)	(0.226)	(0.549)	(0.452)	(0.498)
Sex (Women=1)	-0.284	-0.297	-0.234	-0.215	-0.099	-0.298
	(0.122)	(0.118)	(0.109)	(0.261)	(0.198)	(0.217)
Age	-1.540	-1.463	-1.475	-2.594	-1.690	-2.856
	(0.510)	(0.512)	(0.466)	(1.066)	(0.881)	(0.940)
Children	-0.177	-0.176	-0.170	-0.252	-0.347	-0.003
	(0.120)	(0.115)	(0.107)	(0.244)	(0.207)	(0.191)
Accepted	-2.346	-2.279	-1.899	-9.011	-8.536	-7.891
	(0.134)	(0.133)	(0.132)	(0.399)	(0.375)	(0.383)
Likelihood	-34,427.2	-34,453.3	-34,470.6	-34,391.9	-34,380.5	-34,409.9

Figure 7: Kernel Smoothed Hazard Rates, Samples A and B

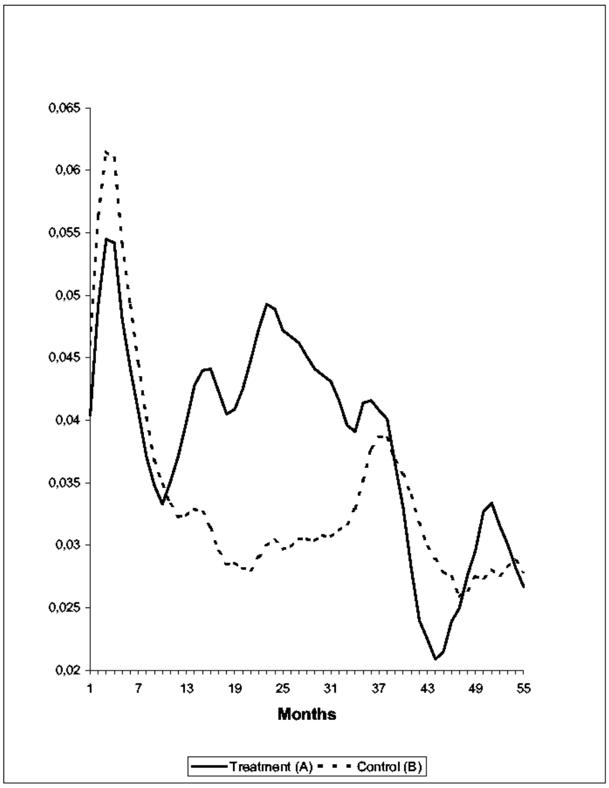


Figure 8: Survival Function, Samples A and B

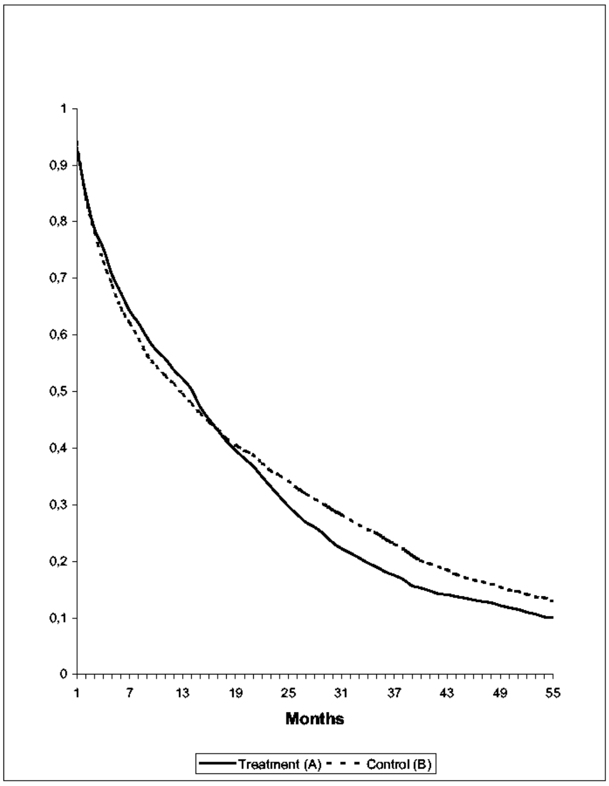


Figure 9: Kernel Smoothed Hazard Rates, Samples B and D

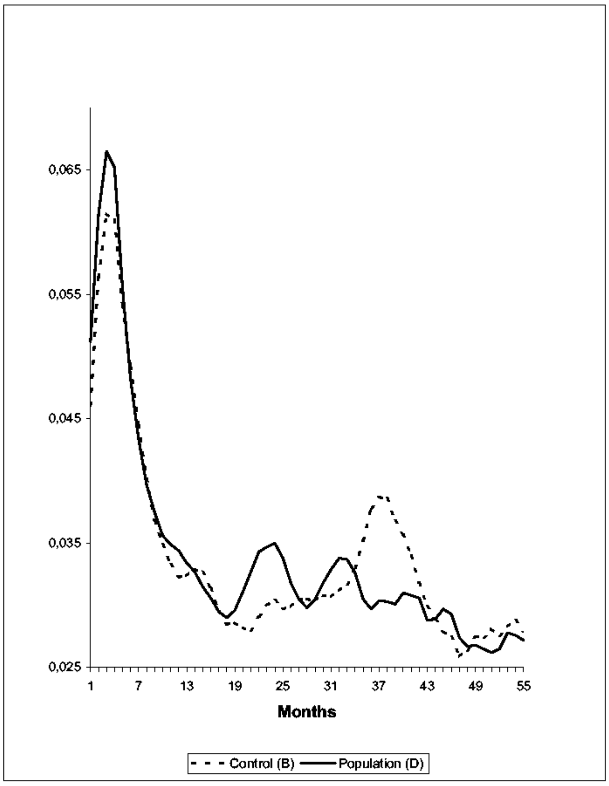
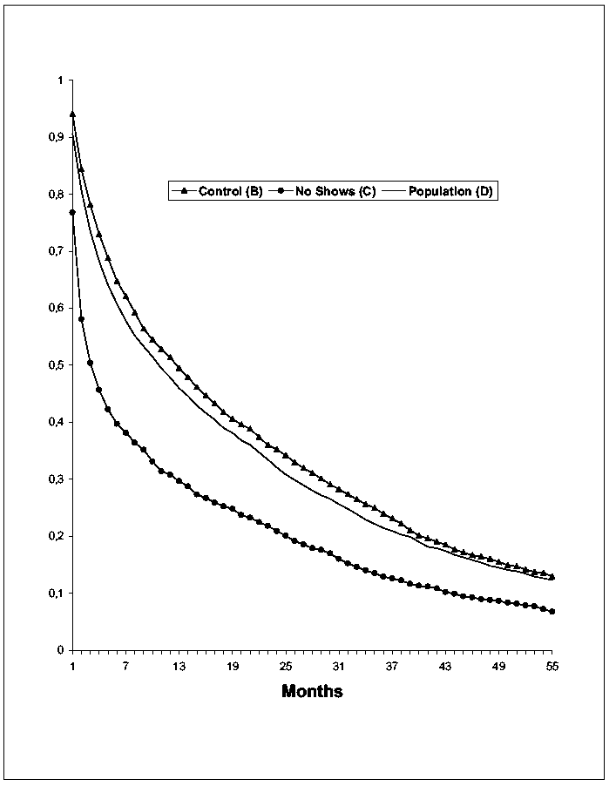


Figure 10: Survival Function, Samples B, C and D


Control (B)
No Shows (C)
Population (D)
Months