Kamionka, Thierry; Lacroix, Guy

Working Paper

Assessing the External Validity of an Experimental Wage Subsidy

IZA Discussion Papers, No. 1508

Provided in Cooperation with:
IZA Network @ LISER, Luxembourg Institute of Socio-Economic Research (LISER)

Suggested Citation: Kamionka, Thierry; Lacroix, Guy (2005) : Assessing the External Validity of an Experimental Wage Subsidy, IZA Discussion Papers, No. 1508, Institute for the Study of Labor (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/20807

DISCUSSION PAPER SERIES

IZA DP No. 1508

Assessing the External Validity of an Experimental Wage Subsidy

Thierry Kamionka
Guy Lacroix

March 2005

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Assessing the External Validity of an Experimental Wage Subsidy

Thierry Kamionka
CNRS and CREST

Guy Lacroix
Université Laval, CIRPEE,
CIRANO and IZA Bonn

Discussion Paper No. 1508
March 2005

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
Email: iza@iza.org

IZA Discussion Paper No. 1508
March 2005

ABSTRACT

Assessing the External Validity of an Experimental Wage Subsidy*

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 4,134 single parents who had applied for welfare between January 1994 and March 1995. It turned out only 3,315 took part in 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 within a year. The purpose of this paper is to determine whether a non-response rate of 20% is likely to harm the external validity of the experiment. We compare the estimated impact of the program using experimental data only to that obtained using additional data on individuals not taking part in the experiment. We find strong evidence of non-response 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.

JEL Classification: I38, C41, C93

Keywords: social experiments, external validity, duration analysis

Corresponding author:

Guy Lacroix
Department of Economics
Pavillon de Sève
Université Laval
Ste-Foy, Québec G1K 7P4
Canada
Email: Guy.Lacroix@ecn.ulaval.ca

* The authors gratefully acknowledge financial support from le Fonds de recherche sur la société et la culture du Québec (FRSCQ). The paper was partly written while Lacroix was visiting the Instituto de Análisis Económico whose hospitality and financial support is gratefully acknowledged. The paper benefited from comments from The Social Research Demonstration Corporation (SRDC). We are also grateful to Joshua Angrist, Pierre Dubois, Rob Euwals, Jean-Pierre Florens, Denis Fougère, Stephen Gordon, Arnaud Lefranc, Thierry Magnac, Jean-Marc Robin, Gerard J. van den Berg and seminar participants at GREMAQ (Toulouse), at McGill University, at Université UQAM, at the annual conference of the ESPE (Athens), at the AFSE conference “Économie des ressources humaines” (Lyon), at the NASM of the Econometric Society (UCLA), at the CEPR-IZA Workshop "Improving Labor Market Performance: The Need for Evaluation" (Bonn), at the ASSET Meeting (Paphos), at the "Malinvaud" seminar (Paris), at Université Paris I - Sorbonne, at the JMA Conference (Montpellier), at the ESEM of the Econometric Society (Stockholm), at the EALE Conference (Seville) and at the LOWER conference (London).

1 Introduction

In seeking to alleviate the employment problems that plague particularly disadvantaged groups, governments have traditionally turned to skill-enhancing training programs. By enhancing skills it was hoped individuals would 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). Many governments have responded to such deceptive results by shying away from traditional training programs and by turning to 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.

The Canadian Self-Sufficiency Project (SSP) is a research and demonstration project that tests a program designed to help single parents receiving income assistance (IA) become self-reliant. It provides a generous, time-limited, earnings supplement to those who find a full-time job and leave IA. As it currently stands, the program requires that welfare recipients remain on welfare for at least one year to qualify for the supplement. Behavioural response to the program is investigated through two major studies: the SSP Recipients Demonstration (SSP-RD) and the SSP Entry Effects Demonstration (SSP-EED). The former focuses on welfare recipients who have already been on the rolls for one year while the latter is concerned with newly enrolled recipients and is the object of the paper.

The SSP-EED aims at documenting the so-called delayed exit effect. Because eligibility for the supplement is conditional on duration, some have expressed the fear this particular feature might induce recipients to postpone their exit from the rolls. The SSP-EED thus randomly selected a sample of single parents who applied for welfare between January 1994 and March 1995 and offered half of them the supplement. By virtue of the experimental design, a simple comparison between treatment and controls would normally suffice to establish whether the delayed exit effect is found in the data and whether the treatment has a significant impact on spell duration. Such a comparison provides appropriate estimates only under a number of relatively stringent assumptions.[1] Assuming they hold, one may also still wonder if the experimental estimate has any external validity. In other words how confident should we be in such an estimate were the program to become official policy ?

Concern about the external validity of experimental estimates goes back to the work Campbell and Stanley (1966). Among the many "threats to validity", the threat to external validity

[1] See Heckman et al. (1999) for a detailed analysis. See also Manski (1995) for a critical assessment of random assignment of social programs and Moffitt (2003) for a similar analysis focusing on welfare programs.

concerns the extent to which the effects found in an experiment can be generalized to different individuals, contexts and outcomes.[2] The issue we address in this paper concerns the inference that can be drawn from the experimental setup for the *population* of welfare claimants.[3] In the SSP-EED demonstration, as in most experiments, the response rate was well below 100%. In fact as many as 20% of sampled individuals are not included in the experiment. Non-response in our context occurred for two reasons. First, a large fraction of individuals who were drawn from the population of welfare recipients at baseline could not be contacted to be asked to take part in the experiment. Second, among those who were contacted at baseline and told about the SSP treatment, a number nevertheless refused to be in the experiment. These two groups of individuals are part of the *population* of welfare claimants but not part of the *experimental sample*. It is thus legitimate to investigate whether non-response impinges upon the external validity of the experiment.

From a logistical point of view, the SSP-EED experiment is well designed.[4] Because assignment to the control and treatment groups was made after the claimant agreed to be participate in the experiment, there is no need to correct for participation as in Dubin and Rivers (1993)[5]. Likewise, individuals who could not be contacted at baseline are unaware of the SSP treatment and are thus not comparable to the "no-shows" considered by Heckman, Ichimura and Todd (1997). Finally, because the refusals were not exposed to the treatment, they do not quite correspond to the "dropouts" considered by *inter alia* Heckman and Smith (2000). The problem we seek to investigate is thus circumscribed to the external validity issue. Our strategy consists in comparing the estimated impact of the program on spell duration using the experimental data only to those obtained using additional information on welfare claimant not taking part in the experiment. Under the null assumption of external validity, the treatment effect should be robust to the information sets. Our results are consistent with those of Berlin,

[2]The issue of external validity has received a lot of interest amongst economists recently. See the work of Meyer (1995) and Manski (1995). Angrist (2004) offers a recent discussion of external validity in the context of the instrumental variables approach. A related topic concerns the extrapolation of experimental results to other sites. See Hotz, Imbens and Mortimer (2003).

[3]In our context, an additional external threat would arise if the new program changed the rate of inflow into welfare. Such "entry-effects" have been discussed in detail by Moffitt (1996, 2003). Naturally, entry-effects can not be measured by demonstration projects. Furthermore, all small-scale experiments are cast within a partial-equilibrium framework. General equilibrium effects are potential and important external threats. Lise, Seitz and Smith (2004) have recently investigated the impact of the SSP program within a general equilibrium framework. Their results show that in all likelihood the SSP program would have little impact on welfare recipients once wage effects and labour markets adjustments are taken into account.

[4]See Hotz (1992) for an appraisal of the JTPA design and the potential problems associated with it.

[5]Dolton, Lindeboom and van den Berg (1999) study the impact of non-response to a survey on the unemployment duration distribution using experimental data. As in our case, assignment to the control and treatment groups occurred prior to the survey. Although they do not model non-response explicitly, their data allows them to identify as many as four reasons not to respond to the survey and account for these by including dummy variables in the regression analysis.

Bancroft, Card, Lin and Robins (1998) in finding little evidence of delayed exits. Furthermore, we find strong evidence of non-response 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.

Nearly all the evaluation studies find that the SSP 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)). The results of this paper suggest these might not be representative of their potential impact on the population of welfare claimants.

The remainder of the paper is organized as follows. Section 2 provides a detailed description of the Entry Effects Demonstration. It also discusses the data at our disposal and presents preliminary evidence of non-response bias using non-parametric tests. Section 3 discusses the statistical model and the treatment of unobserved individual heterogeneity. Section 4 reports our main findings. Finally, Section 5 concludes the paper.

2 The Entry Effects Demonstration

The Self-Sufficiency Project was introduced in Canada in 1992. It comprised two demonstrations that aimed at measuring the response of IA recipients to a financial incentive that made work pay better. In both cases SSP offered a generous, time-limited (3 years), monthly cash payment to single parents who found a full-time job and left IA. In the SSP-EED demonstration, a random sample of welfare entrants were offered the supplement conditional on remaining on the rolls for a minimum of 12 months. This qualifying period would in all likelihood be an important parameter of the program were it to become official policy. Yet 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 IA rolls to become eligible. The second, "conditional" effect, is to induce those currently on the rolls to delay their exit from welfare in order to qualify for the supplement.

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 EED thus randomly sampled single parents who had applied for and received IA in British

Columbia.[6] 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 IA 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 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 Data

The original EED sample was fielded between January 1994 and March 1995. Each month, an independent random sample from the population of welfare applicants was selected. 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 us individual IA histories on participants and non-participants alike using administrative files. It used the same algorithm as above to generate the sample of non-participants.[7] For confidentiality reasons, the data was restricted in two ways. First, only information on the first welfare spell was made available. Second, those who had refused to take part in the experiment were pooled with those who were not sampled at baseline.[8]

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 leaves a total of 4,134 individuals. Of these, 3,315 agreed to sign the informed consent form and complete the baseline survey. The response rate

[6]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)).

[7]Randomization occurred during the first month following application for benefits in most cases. Indeed, over 2,464 individuals had either received no or a single IA payment at randomization. Another 653 individuals had received two monthly payments. Finally, 92 individuals had received as many as three or four payments prior to 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 will tend to decrease the average duration of the experimental sample.

[8]Statistics Canada estimates that 8% 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.

is thus approximately equal to 80%. 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.[9] Finally, 122 individuals refused to take part in the experiment.[10] The randomization procedure yielded the experimental treatment and control groups (henceforth samples *A* and *B*, respectively). Statistics Canada provided us a sample of 3,073 individuals drawn among

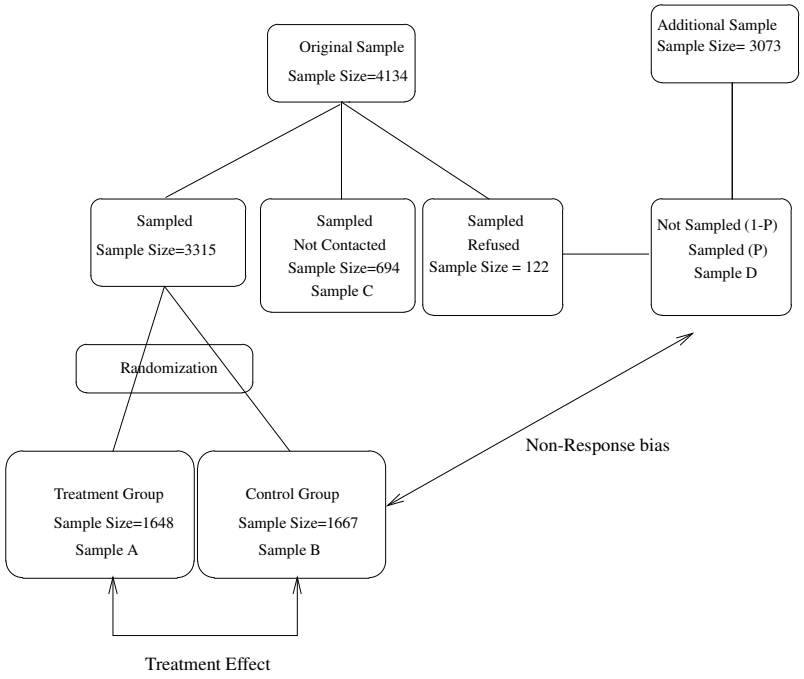

Figure 1: Randomization Scheme

those who were not sampled at baseline and those who refused to be in the experiment. We refer to this group as sample *D*.[11] Note that the refusals are not identifi able 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 or followed-up for different reasons (*C*), and still others would have refused to take part into the experiment. Thus under the null assumption that the data is void of non-response bias, groups *B* and *D* should behave similarly.

[9] 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.

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

[11] The total population of welfare applicants over the period covered by the EED is 7,390. Thus, samples *A*,*B*,*C* and *D* represent over 95% of the total population.

2.2 Descriptive Statistics

Table 1 provides descriptive statistics for each sample separately.[12] 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 similar 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 only have data for the first IA spell, over 9.6% of all spells are censored at 65 months.

To better ascertain the extent to which observable characteristics differ between samples *A*, *B*, *C* and *D*, we report simple logit regressions of belonging to a given sample in Table 3. 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 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.[13]

2.3 Non-Parametric Evidence

Recall from above that the EED aimed at determining whether IA applicants might be induced to delay their exit from welfare in order to qualify for the (relatively) generous earnings supplement. In order to qualify for the supplement, IA recipients had to remain on welfare for at least 12 months. Once qualified, those in sample *A* had to find a full-time job within 12 months

[12] 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.

[13] 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 3.

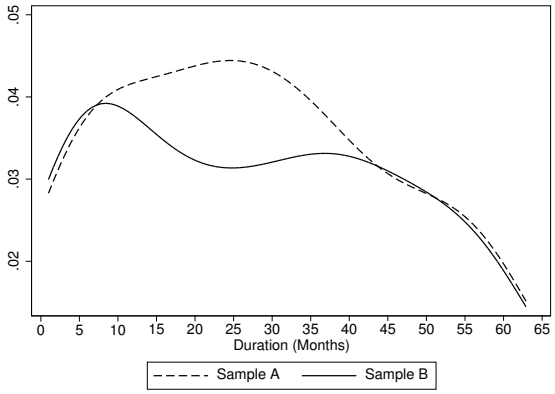

Figure 2: Kernel Smoothed Hazard Functions – Experimental Groups

in order to receive the supplement. Those in sample *B* continued to receive the standard IA benefi ts.

Behavioural response to the EED is best investigated through the use of hazard and survival functions.[14] Figure 2 plots smoothed hazard rates of IA spells for the experimental samples *A* and *B*.[15] The fi rst noteworthy feature of the fi gure is that the treatment sample *A* appears to be sensitive to the parameters of the EED. Indeed, the hazard rates increases in the fi rst 8 months for both groups upon entry into IA. The hazard rates of the treatment group keep increasing up until the 25^{th} month while those of the control decrease steadily.[16]

Weak delayed exit behaviour is evidenced by the difference between the hazard functions during the fi rst 7 months. Indeed, the hazard function of sample *A* lies below that of sample *B* during the fi rst 7 months, then crosses it and remains above for the next 30 months or so. The underlying survival functions are plotted below in Figure 3. Not surprisingly, the survival function of sample *A* lies above that of sample *B* up until month sixteen. This is consistent with the fi ndings of Michalopoulos and Hoy (2001) who have found that the individuals in sample *A* were proportionately more numerous to receive IA than those in sample *B* up until the 5^{th} quarter of the experiment. Based on Figure 3, it seems reasonable to claim that the earnings supplement fi rst induces individuals to delay their exit in the beginning months and then provides a relatively strong incentive to leave IA. It is worth investigating whether these

[14]This section only presents brief non-parametric evidence on non-response bias in the Applicant Study. More extensive analyzes using non-parametric permutation tests can be found in Lacroix and Royer (2001).

[15]Recall that approximately 20% of the sample had been on welfare for at least 2 months prior to randomization. If we use first month on IA instead of randomization date as the start of the spell, the figure is basically unchanged. We use the Epanechnikov kernel with optimal bandwidth to smooth the hazard functions.

[16]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).

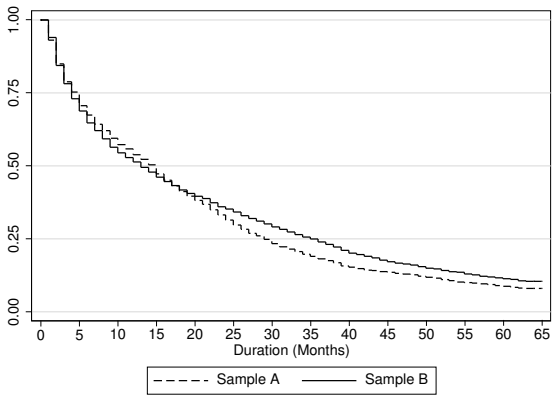

Figure 3: Survival Functions – Experimental Groups

differences are statistically signifi cant. 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[17]

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

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 fi rst 12 months are found to be 8.69 and 8.48, respectively, a difference approximately equal to 2.5% in favour of sample *A*. A simple $\chi^2(1)$ test can not reject the null assumption that both durations are equal. This fi nding is similar to that of Berlin et al. (1998) who report an average impact of approximately 3.0% that is hardly signifi cant. On the other hand, mean durations computed over $[0, 64]$ are equal 20.3 and 21.8, respectively. This time, the $\chi^2(1)$ test (=4.38) does reject the null assumption that mean durations are equal.

One could thus conclude that the treatment reduces mean duration by approximately 7.4%. Although 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 Tables 3. The more

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

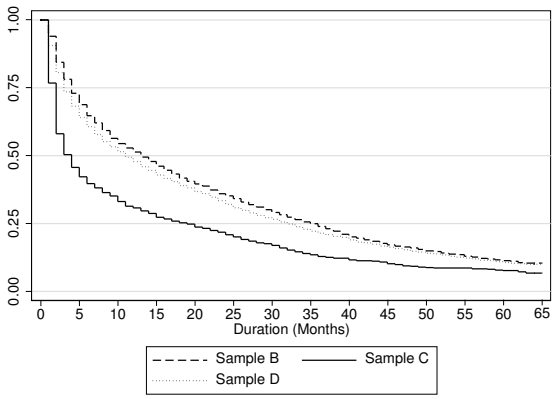

Figure 4: Survival Functions – Samples *B,C,D*

interesting question that must be addressed is whether our estimates are plagued with non-response biases. Before we address this question formally, we will present informal evidence that such biases may be present in the data.

Figure 4 plots the survival functions of samples *B*, *C* and *D*. Notice first 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 non-response bias, sample *D* would normally constitute a proper control group since the two differ only insofar as the individuals in the former (*D*) were not sampled while those in the latter (*B*) were sampled and agreed to participate in the experiment.[18] 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 would not have been contacted had they been 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 supplement had they been contacted at baseline.

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

[18]One can not rule out the possibility that a number of the 122 refusals are in set *D* and contribute to the difference between the survival curves. Under the null assumption of no bias they should not make any difference.

3 Modeling Individual Spell Durations

In order to derive an appropriate estimator of the treatment effect, non-response bias must be explicitly taken into account. The sampling frame within which the experiment took place was illustrated in Figure 1. Our task is to model all the available information. In order to do this, we first need to determine the probability of belonging to the experimental samples. According to Statistics Canada, individuals in samples *A* and *B* represent 45% of all claimants over the enrolment period.[19] 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%. We will thus consider that each applicant faces a probability $p = 0.65$ of being sampled.[20]

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

$$
\begin{aligned}
E &= \begin{cases} 1, \text{ if the individual was sampled at baseline,} \\ 0, \text{ otherwise.} \end{cases} \\
A &= \begin{cases} 1, \text{ if the individual is willing to participate in the experiment,} \\ 0, \text{ otherwise.} \end{cases} \\
R &= \begin{cases} 1, \text{ if the individual could be contacted at baseline,} \\ 0, \text{ otherwise.} \end{cases} \\
T &= \begin{cases} 1, \text{ if the individual belongs to the treatment group,} \\ 0, \text{ otherwise.} \end{cases}
\end{aligned}
$$

Finally, let y be a realization of the experiment:

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

where u is the duration of an IA spell.[21]

[19] See footnote 11.

[20] The indeterminacy of the probability of being sampled arises due to some confusion related to sample *C*. According to private communications with Statistics Canada officials, our sample *C* only includes individuals that could not be contacted at baseline. In such a case, the probability of being sampled is roughly equal to 65%. 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%. The model was estimated with $p = 0.60$ and $p = 0.65$. The main results are very robust to the choice of p.

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

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 2: Realizations of random variables

Which arguments of $y(\cdot)$ are observable depend on which set an individual belongs to. Only T and U are observable for all individuals.[22] Thus for sample *A* we know individuals have been sampled ($e = 1$), that they have agreed to participate ($a = 1$), that they could be contacted ($r = 1$) and are eligible for the supplement ($t = 1$). Table 2 summarizes the realizations of the random variables according to group membership.

3.1 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. 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 $\theta \in \Theta \subset I\!R^P$ is a vector of parameters. When the IA 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 fi nite 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.

[22] The IA spells are right censored at 65 months.

3.2 Modeling Non-Response

In this section we focus on the conditional distributions of variables 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. Clearly these two probabilities are independent of individual characteristics.

In the context of the EED, two types of non-response were observed in the data. The first concerns individuals who refused to take part in the experiment. The second arises due to logistic problems: sampled individuals could not be contacted at baseline because they had moved between the time they applied for IA and the time they were sampled. Given the nature of the data at our disposal, it is not possible to model non-response explicitly. Instead we assume that the acceptation and mobility decisions are linked to observable and unobservable characteristics. Hence define $z(x, \nu)$ as the conditional probability that 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

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

and 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

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

where

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

where a is the realization of the participation decision, β_r is a vector of parameters and $\xi_a \in \mathbb{R}$ is a scalar parameter. 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.[23]

[23] It should be noted that assuming there is no correlation between the latent variables does not imply that they are independent. Indeed, the conditional expectation of the recontact variable depends on the acceptation decision. Consequently, whereas the errors term 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 ξ_a (see equation (6)).

3.3 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 support 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 term.

   Assume that

$$\text{Prob}[V = v] = \begin{cases} p_0, & \text{if } v = \nu_0, \\ (1 - p_0), & \text{if } v = -\nu_0, \end{cases} \tag{7}$$

   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. $E(V) = 0$).

   The log likelihood is

$$\log(L(\theta)) = \sum_{i=1}^{N} \log(l_i(\theta)), \tag{8}$$

   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[24]

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

   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 support points J is set to 2.[25] π_1 represents the probability that the unobserved term V takes the value ν_0 ($\pi_2 = 1 - \pi_1$).

[24] See section 3.1.

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

2. *Continuous distributions*

The unobserved heterogeneity terms ν can also be assumed to be independently and identically distributed across individuals. We will write $g(\nu;\gamma)$ as any well-behaved probability density function 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 (8), where $l_i(\theta)$ is the contribution to the likelihood of the sequence y_i.[26] 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 specifi c 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 effi cient parameter estimates if $\frac{\sqrt{N}}{H} \to 0$ when $H \to +\infty$ and $N \to +\infty$ (see Gourriéroux and Monfort (1991, 1996)). Under these conditions, this estimator has the same asymptotic distribution as the standard 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) and Gilbert et al. (2001)).

3.4 Specification of conditional hazard function

The conditional hazard function for welfare durations is given by

$$h(u \mid x;a;r;t;\nu;\theta) = h_0(u;\alpha)\ \varphi(x;a;r;t;\beta_d)\ \exp(-\nu), \tag{9}$$

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:

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

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

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.[27]

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

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

where $u \leq 64$ months.

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

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

if $u > 64$ months.

3.5 Likelihood Functions

It is possible to examine the impact of the non-response biases on the estimated treatment effect by considering estimates based on various information sets. In what follows we derive the likelihood function of four different estimators. Each is based on data that are likely to be available at little cost when using demonstration projects.

3.5.1 *Standard Experimental Estimator (A,B)*

For comparative purposes we start we the standard experimental estimator that omits non-response. The treatment effect is based solely only the experimental control and treatment groups *A* and *B*. Each individual contributes a sequence $y = (t, u)$ to the likelihood function.

[27] 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.

Since they all agreed to participate and could all be contacted at baseline 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*.

The conditional distribution of the welfare durations corresponds to the hazard function (9), where $\delta_a = \delta_r = 0$ (here a and r are set equal to arbitrary values in the conditional distribution of the welfare duration).

3.5.2 *Selected Samples at Baseline (A,B,C)*

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. 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 (see equation (6)), where $\xi_a = 0$ (a is fi xed to an arbitrary value both in this equation and in the conditional hazard function). The conditional hazard function of the welfare durations is given by the equation (9) where $\delta_a = 0$.

3.5.3 *Selected and Contacted with Non-Selected Sample (A,B,D)*

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

selected may or may not have agreed. 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 expression of the conditional hazard function of the welfare durations is given by equation (9), where $\delta_r = 0$ (r, for convenience, is fi xed to an arbitrary value in the expression of the conditional hazard).

3.5.4 *All Samples (A,B,C,D)*

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

$$\ell_\nu(\theta) = p\ z(x,\nu)\ (1-\phi(x,a,\nu))\ 0.5\ f(u \mid x; a=1; r=1; t=1; \nu; \theta), \tag{10}$$

if the individual belongs to group *A*;

$$\ell_\nu(\theta) = p\ z(x,\nu)\ (1-\phi(x,a,\nu))\ 0.5\ f(u \mid x; a=1; r=1; t=0; \nu; \theta), \tag{11}$$

if the individual is in group *B*;

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

if the individual is in group *C*;

and

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

if the individual belongs to group *D*.[28]

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

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

Let S_a and S_r denote the sets of possible values of A and R. Both are defi ned in a similar fashion to S_e. Finally, the contribution to the likelihood function can be written[29]

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

[28]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 *D* was the complement of samples *A*, *B* and *C*. In principles, 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 11, sample *D* comprises over 95% of that population. Further, selection into the sample was made using a random procedure. We have thus chosen not to weigh the function so as to avoid making an already involved function overly complicated.

[29]One may question whether there is a unique mapping between these reduced form equations and the structural model. Note that we have 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 that their variances be normalized to unity. Furthermore, there are no correlations between the latent variables and the duration variable. It is then then possible to show that a generalized order condition holds for each latent equation in the conditional model (see Fomby, Hill and Johnson (1984)).

4 Results

4.1 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 setup is expected to delay exit prior to 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 the table provide estimates based on non-parametric unobserved heterogeneity (see equation (7)).[30]

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 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 2.3 where it was found that the treatment group had a 7.3% shorter mean duration.

Column 2 of the table reports the results using samples *A*, *B*, and *C*. The baseline hazard function is decreasing with duration. As previously, spell duration decreases with age and increases with the number of children. Likewise, women are found to 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 zero. This is consistent with the fact that those who could be contacted have longer spell durations (see Table 1). Hence, once we include those that could not be contacted at baseline, the treatment effect vanishes. The third panel of the table reports the parameter estimates of the probability of not being contacted at baseline. It is found that the probability is decreasing 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 the table reports the results using samples *A*, *B*, and *D*. Contrary to the previous cases, the conditional hazard function is increasing with duration. Inclusion of this group

[30] We only report results based on the restricted version, *i.e.* $p_0 = 0.5$. Except for a few specifications, p could be estimated freely. The parameter estimates are relatively robust to the estimation of p_0.

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 samples *A*, *B* and *C*. The parameter of the acceptance binary variable is positive and statistically signifi cant. 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 samples *A* and *B*. Consequently, omission of the participation decision signifi cantly 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 signifi cant in this specifi cation.

Column 4 of the table reports the results using groups *A*, *B*, *C* and *D*. The parameter estimates show that the conditional hazard function is increasing with duration. The sign of the slope parameters are similar to those of the previous specifi cations. 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 signifi cant.

The next two panels indicate that the probability of not being contacted is decreasing 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 signifi cantly lower 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 signifi cant at 5%.

The estimates in columns (1)–(4) of Table 4 are based on a rather restrictive specifi cation for the unobserved heterogeneity component. Previous research has nevertheless 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.

The last four columns of Table 4 report results based on particular parametric distribution and using samples *A*, *B*, *C* and *D*. The parameter estimates are thus comparable to those of column 4. The treatment effect is still sizable although slightly smaller than that of column (4), except for the specifi cation based on the student distribution (with 5 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 treat-

ment effect appears to be relatively robust with respect to the distribution of the unobserved heterogeneity.

4.2 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 setup so as 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 ([0,12[, [12,24[, [24,36[, [36 and more].).

The estimation results are reported in Table 5. The table has the same setup 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 interval are all highly significant. The results indicate that the treatment effect reduces durations considerably over the [12,24[and [24,36[intervals. On the other hand, the treatment group appears to have longer spells over the [36 and more] interval. The parameter α indicates that 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 that 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 [0,12[interval. Likewise, the treatment effect over the [12,24[and [24,36[intervals reduces duration considerably. In both cases, it is found that the treatment has no impact on the mean duration over the [36 and more] interval.

The specifications in columns (5)–(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 $T \geq 36$ interval.

4.3 Mean Durations

The slope parameters can not directly be interpreted as marginal impacts since the expected duration is highly non-linear with respect to the covariates.[31] We thus report the conditional (on treatment) expected durations for various model specifi cations in Table 6. The top panel of the table reports the expected durations based on the parameters of the fi rst column of Table 4. This specifi cation 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 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% for women, and 7.7% for men.

The middle panel uses the same parameter estimates as the top panel except that the drawing is made within sample D. This allows 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 the table 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 estimates the parameters. To compute mean durations, only data from sample D is 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% for both men and women.

Our results show that the experimental samples are composed of self-selected individuals with longer mean spell duration than the population of IA claimants. We conjectured earlier that such bias did not necessarily imply that the estimates of the treatment effect needed be biased. According to our parameter estimates and to our simulations, though, it does seem that the estimates are biased.

5 Conclusion

Over the past twenty years demonstration projects have become the preferred means 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,

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

implementation of a such projects is likely to be hampered by many logistical and behavioural problems that may prove detrimental to the so-called external validity of the experiment (see Hotz (1992), Moffi tt (1992, 1996, 2003), Manski (1995)). Although the literature has singled out randomization bias as the main culprit, we know surprisingly little about the extent to which non-response harms the validity of demonstrations projects. The evidence brought to bear is almost always indirect at best.

In Canada, a policy aiming at helping single parents on income assistance become self-reliant was implemented on an experimental basis. The Self-Suffi ciency Entry Effects Demonstration (EED) focused on newly enrolled recipients. The EED randomly selected a sample of 4,134 single parents who had applied for welfare between January 1994 and March 1995. It turned out only 3,315 actually took part in the experiment despite a 50% chance of receiving a generous, time-limited, earnings supplement conditional on fi nding a full-time job and leaving income assistance. A large fraction of those who did not participate were left out of the experiment because they could not be contacted at baseline. Because the non-respondents are part of the target population, and because they behave somewhat differently from the experimental sample, their omission raises concern about the external validity of the experiment. We thus investigate whether a non-response rate of 20% is likely to bias the measurement of the treatment effect. Our empirical strategy consists in comparing the estimated impact of the program using experimental data only to those obtained using additional data on individuals not taking part in the experiment and drawn from the same population.

We write the likelihood of various sets of information and obtain relevant estimates of program impact on welfare spell durations. We fi nd strong evidence of non-response bias in the data. When we correct for the bias, we fi nd that 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 fi nancial incentives than others. Furthermore, we fi nd no evidence of the so-called "delayed exit effect" that may arise due to the program setup. Finally, the sensitivity of the parameter estimates to distributional assumptions pertaining to the unobserved heterogeneity is also investigated. We fi nd that many parametric distributions yield similar results to those obtained from a simple non-parametric model.

References

Angrist, J.D. (2004) 'Treatment effect heterogeneity in theory and practice.' *The Economic Journal* 114(March), C52–C83

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-suffi ciency 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

Campbell, D., and J. Stanley (1966) *Experimental and Quasi-Experimental Designs for Research* (Boston: Houghton Mifflin)

Dolton, P., M. Lindeboom, and G.J. van den Berg (1999) 'A taxonomy of survey non-response and its relation to the labor market.' In *The Creation of Employer-Employee Matched Data,* ed. J. Haltiwanger, J. Lane, J. Spletzer, J. Theeves, and K. Troske (Amsterdam: North-Holland)

Drolet, S., B. Fortin, and G. Lacroix (2002) 'Welfare benefi ts and the duration of welfare spells: Evidence from a natural experiment in Canada.' Forthcoming, *Journal of Public Economics*

Dubin, J.A., and D. Rivers (1993) 'Experimental estimates of the impact of wage subsidies.' *Journal of Econometrics* 56, 219–242

Fomby, T.B., R. C. Hill, and S.R. Johnson (1984) *Advanced Econometric Methods* (Springer-Verlag)

Fougère, D., B. Fortin, and G. Lacroix (2002) 'The effects of welfare benefi ts on the duration of welfare spells: Evidence from a natural experiment in Canada.' In *Institutional and Financial Incentives for Social Insurance,* ed. C. D'Aspremont and P. Pestieau (Kluwer Press) chapter 1, pp. 1–24

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

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

Harknett, K., and L. A. Gennetian (2001) 'How an earnings supplement can affect the marital behaviour of welfare recipients: Evidence from the self-suffi ciency 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., and G.E. Borjas (1980) 'Does unemplyment cause future unemployment ? defi nitions, questions and answers from a continuous time model of heterogeneity and state dependence.' *Economica* pp. 247–283

Heckman, J.J., H. Ichimura, and P.E. Todd (1997) 'Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme.' *Review of Economic Studies* 64(4), 605–654

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 Eds. D. Card (North-Holland) chapter

Heckman, J.J.and N. Hohmann, and J. Smith (2000) 'Substitution and dropout bias in social experiments: A study of an influential social experiment.' *QJE* pp. 651–690

Hotz, V. J. (1992) 'Designing an evaluation of the Job Training Partnership Act.' In *Evaluating Welfare and Traiing Programs,* ed. C.F. Manski and I. Garfi nkel (Harvard University Press) chapter 2, pp. 76–114

Hotz, V.J., G.W. Imbens, and J.H. Mortimer (2003) 'Predicting the effi cacy of future training programs using past experiences at other locations.' mimeo

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

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

Lacroix, G., and J. Royer (2001) 'Vérifi cation empirique de l'absence de biais de non-réponse à l'aide d'une procédure de test MC.' *mimeo*, Université Laval

Lise, J., S. Seitz, and J. Smith (2004) 'Equilibrium policy experiments and the evaluation of social programs.' Mimeo, University of Maryland

Manski, C.F. (1995) 'Learning about social programs from experiments with random assignment of treatments.' Institue for Research on Poverty, D.P. 1061-95

Meyer, B.D. (1995) 'Natural and quasi-experiments in economics.' *Journal of Business and Economic Statistics* 13(2), 151–161

Michalopoulos, C., and T. Hoy (2001) 'When fi nancial work incentives pay for themselves: Interim fi ndings from the self-suffi ciency project's applicant study.' *Working Paper*, SRDC

Michalopoulos, C., D. Card, L. A. Gennetian, K. Harknett, and P. K. Robins (2000) 'The self-suffi ciency project at 36 months: Effects of a fi nancial work incentive on employment and income.' *Working Paper*, SRDC

Moffi tt, R. A. (1992) 'Evaluation methods for program entry effects.' In *Evaluating Welfare and Training Programs,* ed. C. F. Manski and I. Garfi nkel (Harvard Univeristy Press) pp. 231–152

Moffi tt, 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

— (2003) 'The role of randomized fi led trials in social science research: A perspective from evaluations of reforms of social welfare programs.' Cemmap Working Paper CWP23/02

Morris, P., and C. Michalopoulos (2000) 'The self-suffi ciency 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-suffi ciency project's fi nancial incentives.' *Working Paper*, SRDC

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	1648	1667	637	3073

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

Table 3: 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	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 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≤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≤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≥ 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†	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†	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‡	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)

* 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.