PARIS SCHOOL OF ECONOMICS
ÉCOLE D'ÉCONOMIE DE PARIS

WORKING PAPER N° 2009 - 15

# Sample attrition bias in randomized experiments:

# A tale of two surveys

Luc Behaghel

Bruno Crépon

Marc Gurgand

Thomas Le Barbanchon

JEL Codes: C14, C21, C24, J64, J68
Keywords: unemployment, job search, counselling, attrition, sample selection

PARIS-JOURDAN SCIENCES ECONOMIQUES
LABORATOIRE D'ECONOMIE APPLIQUÉE - INRA

48, BD JOURDAN – E.N.S. – 75014 PARIS
TÉL. : 33(0) 1 43 13 63 00 – FAX : 33 (0) 1 43 13 63 10
www.pse.ens.fr

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE – ÉCOLE DES HAUTES ÉTUDES EN SCIENCES SOCIALES
ÉCOLE NATIONALE DES PONTS ET CHAUSSÉES – ÉCOLE NORMALE SUPÉRIEURE

# Sample attrition bias in randomized experiments: A tale of two surveys *

L. Behaghel†, B. Crépon‡, M. Gurgand§, and T. Le Barbanchon¶

May 11, 2009

**Abstract**

The randomized trial literature has helped to renew the fields of microeconometric policy evaluation by emphasizing identification issues raised by endogenous program participation. Measurement and attrition issues have perhaps received less attention. This paper analyzes the dramatic impact of sample attrition in a large job search experiment. We take advantage of two independent surveys on the same initial sample of 8,000 persons. The first one is a long telephone survey that had a strikingly low and unbalanced response rate of about 50%. The second one is a combination of administrative data and a short telephone survey targeted at those leaving the unemployment registers; this enriched data source has a balanced and much higher response rate (about 80%). With naive estimates that neglect non responses, these two sources yield puzzlingly different results.

Using the enriched administrative data as benchmark, we find evidence that estimates from the long telephone survey lack external and internal validity. We turn to existing methods to bound the effects in the presence of sample selection; we extend them to the context of randomization with imperfect compliance. The bounds obtained from the two surveys are compatible but those from the long telephone survey are somewhat uninformative. We conclude on the consequences for data collection strategies.

## 1 Introduction

Sample attrition that occurs between the initial randomization and the measurement of outcomes is a major threat to the validity of randomized field experiments. Random assignment

*We would like to thank ANPE, Unédic and DARES for their involvement in the experiment from which the data are drawn, and for their financial support to the evaluation. Lucie Gadenne provided outstanding research assistance.

†Paris School of Economics (Inra), Crest and J-PAL

‡Crest and J-PAL

§Paris School of Economics (CNRS), Crest and J-PAL

¶Crest and Dares

to treatment creates a treatment group and a control group that are at the same time comparable and representative of the initial population. In the presence of sample attrition, however, the observed treatment and control groups may not be comparable anymore, threatening the internal validity of the experiment; and they may not be representative of the initial population, threatening its external validity.

In this paper, we analyze the consequences of sample attrition, in the context of a job search experiment in which the attrition and the measurement of outcomes posed specific challenges. The experiment took place in France over 12 months in 2007, in 10 different regions and involved the randomization of more than 200,000 job seekers[1]. The treatment groups were offered intensive counseling and monitoring services. Two distinct programs were actually tested: the first one was provided directly by the caseworkers of the French public employment agency (ANPE); the second one was supplied by private firms mandated by the French unemployment benefit provider (Unédic). Each program aimed at serving about 40,000 people over a year; however, unemployed workers randomized into treatment were free to participate or not, and the actual take-up rates (around 50%) implied that the programs had to be offered to about twice as many people. This extremely realistic setting – the "experimental" policy was actually a full-scale trial on a significant part of the French territory – gives the experiment high external validity. However, it also complicates the measurement of outcomes. In this paper, we focus on the methodological lessons learned from measuring the impact on transitions from unemployment to employment and on the type of job held.

Our approach can be briefly summarized: we take advantage of two independent sources available for part of the experimental sample. The first one is a long telephone survey designed for more than 8,000 persons that had a strikingly low response rate of about 50% (thereafter: "the long telephone survey"). The second one is a combination of administrative data and a short telephone survey targeted at those leaving the unemployment registers (thereafter: "the enriched administrative data"); this enriched data has a much higher response rate (about 80%). With naive estimates that neglect non responses, these two sources yield puzzlingly different results. We then turn to existing methods providing bounds on the effects in the presence of sample selection; we extend them to the context of randomization with imperfect compliance. The bounds obtained from the two surveys are compatible but those from the long telephone survey are somewhat uninformative. We conclude on the consequences for data collection strategies.

This methodological paper relates to two strands of literature. The first one, following Horowitz and Manski (2000), develops approaches to bound effects in the presence of sample selection. In particular, Lee (2008) proposes sharp bounds under the assumption that selection is a monotonous function of treatment assignment. Our contribution is to extend Lee's framework to the case of imperfect compliance and to estimate bounds associated with local average treatment effects (LATE). Second, our paper relates to the empirical literature on unemploy-

[1] L. Behaghel, B. Crépon, M. Gurgand and J. Guitard were in charge of designing the experiment and conducting the evaluation.

ment duration. Given its use of experimental data, it of course relates to the few existing job search experiments such as those reviewed by Meyer (1995). But given its focus on the measurement of exit from unemployment to employment, it also closely relates to the literature on the impact of unemployment benefit exhaustion. In particular, Card, Chetty and Weber (2007a and 2007b) show the importance of considering exit to re-employment rather than simply exit from registered unemployment as the outcome. In their review of the literature (in the working paper version), they also illustrate the role played by different data sources (administrative, large standard surveys like CPS, and ad hoc telephone surveys). What you measure and how you measure it matters a lot when it gets to exit from unemployment. We add a slightly different but complementary warning: our example shows (in a way that somehow surprised us) how non responses suffices to generate estimates that look economically sizeable and statistically significant – but are also spurious. This is a serious concern in view of the fact that detailed outcomes, such as the quality of jobs or even re-employment, are, in some cases, only available through surveys that can be subject to high attrition rates.

The next section presents the experiment and the data in more details. Section 3 shows "naive" results without taking into account sample selection. Section 4 provides evidence of sample selection. Section 5 discusses, extends and estimates bounds on the effects. Section 6 discusses alternative data collection strategies.

# 2 The programs, the experiment and the data

## 2.1 The programs

Since the 1980s, the French labor market has been characterized by high unemployment rates and persistent long-term unemployment. At the beginning of the two programs, unemployment was decreasing but 8.4 % of the labor force was still unemployed; among them, about 30% had been unemployed for a year at least.

In the mid 2000s, the French employment benefit provider (Unédic) had locally experimented with supplying counseling to unemployment benefit claimants who had a high statistical risk of long-term unemployment. Private firms (temporary agencies or firms specialized in the placement of job seekers) provided the counseling. Unédic decided to scale up this program, targeting 41,000 job seekers (exclusively among those eligible for unemployment benefits) in 2007. Each job seeker was to receive intensive counseling during up to six months. The payment to private firms was partly conditioned on re-employment of the job seeker. It typically had three equal shares: 1/3 at the beginning of the counseling; 1/3 if the job seeker found a job in the first six months (indefinite duration contract or fixed-term contract for at least 6 months); 1/3 if the worker was still employed after 6 months. In what follows, we call Unédic-mandated counseling the "private scheme".

Simultaneously but separately, the French employment public agency (ANPE) decided to launch its own in-house program of intensive counseling. In addition to the population targeted by Unédic, the program also targeted job seekers who were not eligible for unemployment benefits. Each job seeker was also to receive intensive counseling during up to six months and the target was to start the program with 40,000 job seekers in 2007. In what follows, ANPE in-house program is called the "public scheme".

The exact content of the programs varied locally. However the basic structure was the same everywhere and across the two programs: a more intensive follow-up, with at least a weekly contact (email, phone) and a monthly face-to-face meeting between the job seeker and his personal counselor. Compared to the usual track, where a contact is supposed to take place every month and where ANPE agents follow on average 120 job seekers, this is a significant increase in human resources dedicated to follow the job seeker. In contrast to comparable job search experiments in the US (Meyer, 1995; Ashenfelter, Ashmore and Deschênes, 2005), the treatments did not directly include stricter enforcement of search requirements, even though the more frequent interactions with counselors may be viewed as increased monitoring.

### 2.2 Experimental design

The populations concerned by the two programs partially overlapped. Benefit recipients were eligible for ANPE-provided counseling and Unédic-mandated counseling in regions where the two programs coexisted. In some regions, only one program was experimented. Moreover, job seekers who were not eligible for benefits over a sufficient period (1 year) were not eligible for Unédic-mandated counseling. Setting apart job seekers that had already been unemployed for three months at the beginning of the experiment, this defines four populations: (i) benefit recipients in regions with the two programs; (ii) benefit recipients in regions with ANPE-provided counseling only; (iii) benefit recipients in regions with Unédic-mandated counseling only; (iv) non recipients (eligible only for the public scheme). For the sake of simplicity, in the analysis below, we estimate average treatment effects on these heterogenous populations by pooling the different sub-samples together and using fixed effects to control for differences in populations. The effects of the private and public schemes are thus computed on different populations and are not directly comparable. This comparison is not the topic of interest here; we consider it in separate, on-going work.

The randomization took place during the first interview at the local ANPE branch (that is, at registration as unemployed). After the job seeker had been diagnosed eligible for the program(s), a software installed on the computer of the employment service agent randomly assigned him to treatment 1 (public scheme), treatment 2 (private scheme) or to the control group (usual track). The probabilities of assignment to each group varied locally and across the four populations so as to maximize the power of the statistical evaluation while complying with the quantitative objectives of each program (the objectives of 40,000 and 41,000 job seekers had been subdivided

locally). This often implied very high probabilities of assignment to treatment 2 (up to 85%) and much lower probabilities of assignment to treatment 1 (down to 6%) and to control (down to 9%).

Upon randomization, the job seeker was told by the employment service agent which track he was offered. The job seeker was free to refuse the more intensive tracks. Depending on the assignment and on his early decision during the first interview, he was afterwards contacted either by ANPE services for the usual track, or by a dedicated counselor from the ANPE-provided intensive scheme, or by one of the private firms mandated by Unédic. Job seekers of the two treatment groups effectively entered the program they were assigned to by signing a specific agreement; if they refused to sign, did not show up, or were eventually found not to meet the criteria of the intensive scheme, they went back to the usual track. Thus, a significant share of each treatment group (about 55%) did not actually enter the scheme they were assigned to. Following the usual terminology, they are non-compliers. The high rates of non compliance were expected and had been taken into account in the statistical power calculations; along with the unbalanced assignment probabilities, they are a factor limiting the precision of the evaluation. This appears as a price to pay for this large-scale, realistic setting; it is fortunately counteracted by the large samples.

### 2.3 The data

The data gathered for the evaluation comes from a variety of sources. We focus here on the two sources that are used to measure *employment*.

The first source is a telephone survey (the "long telephone survey") that took place in March and April 2008. The initial sample included 9011 job seekers of the four populations who had entered the experiment between April and June 2007. Job seekers had therefore been assigned for about 10 months when they were surveyed. The sample was stratified according to the job seekers' random assignment and to whether they had signed or not for an intensive scheme. The interviews were conducted by an independent pollster mandated by the French Ministry of Labor (DARES). The questionnaire was long (a maximum of 103 questions, for an estimated average time of 20 minutes). Detailed questions were asked upon the track followed when unemployed (what they were proposed, whether they accepted or not, why, what they did,...) and on the current employment situation. The first question of interest here is: "Question 73. What is your current situation today? 1. Employment (wage earner or not, including the creation of a firm, if effective). 2. Project of creating a firm. 3. Internship, studies. 4. Unemployment. 5. Retirement, early retirement (...). 6. Out of the labor force (...)." From this, we create the binary outcome "employment" equal to 1 if the person chose response 1, 0 otherwise. When employed the individual was also asked whether his job was under an indefinite duration contract, and whether it was part-time or full-time. We create the binary outcome "permanent employment" equal to 1 if the person was employed in a full-time, indefinite duration job.

The next two sources are the administrative records held by ANPE on all registered unemployment spells and a short telephone survey. The administrative file held by ANPE is used here to measure the time spent in registered unemployment. As is well known and well illustrated by Card et al. (2007a), the end of a registered spell may be due to quite different events. In our data, it is in some cases clear whether the job seeker has found a job or decided to exit the labor force (for about 50% of ending spells, the job seeker returned a form stating his motives or the information was coded directly by the employment services during an interview or a phone contact); but in about 50% of cases, there is no way to know from the administrative records whether the job seeker had found a job or not. As part of the experimental design, a very short phone survey was therefore mandated to an independent pollster, on a subsample of workers whose destination at the exit from unemployment was unknown from the administrative source. The questionnaire was extremely focused so as to actually mimic the form that other job seekers had filled out upon exiting registered unemployment. It had a maximum of 4 questions. We are using the first one: "Question 1. During the month of ..., you stopped being registered at ANPE. What was the reason?" The sampling probabilities for this survey were optimized to partly correct for the imbalance between treatment and control groups. Moreover, to avoid recall error, the survey was implemented monthly on those who had recently left the unemployment registers. Thus combined, the administrative records of registered unemployment spells and the short phone survey allow us to measure transitions from registered unemployment to employment at various horizons. We call the corresponding binary outcome variable "exit to employment".

It is worth noticing that the "employment" variable (from the long telephone survey) and the "exit to employment" variable (from the enriched administrative data) measure conceptually and practically distinct outcomes. This is first due to the fact that not all job seekers who have a job exit registered unemployment. It is indeed possible for a worker who has a part-time or unstable job to remain registered as a job seeker (his unemployment benefits being adjusted according to his earnings). He will not, in this case, be recorded as transiting from unemployment to employment; he will, however, be reported as employed in the long telephone survey. Second, this is due to the fact that the date of the information is not strictly identical in the two sources. For example, job seekers may have exited registered unemployment to employment, but may have lost their jobs by the time of the long telephone survey. We come back to this measurement difference when comparing the results from the two sources.

Another difference between the two sources lies in their very different response rates. Table 1 shows the (weighted and unweighted) size of the three samples we use: the full sample, for which only administrative records are available, the enriched administrative data, and the long telephone survey data. The response rate of the long telephone survey is 50%, much lower than the response rate in the enriched administrative data (81.2%). The full sample has 8137 observations: indeed, from the initial 9011 job seekers, we disregard those areas or periods that

did not yield individuals in the control group as well as in at least one of the treatment groups, as well as two regions where the enforcement of the experiment was problematic.

## 3 Naive results without correcting for sample selection

Table 2 shows naive estimates of treatment effects. More precisely, for each data source, we restrict the analysis to job seekers for whom the outcome variable is available. The approach is naive in the sense that we do not attempt to make any correction for sample selectivity at that stage.

As discussed above, compliance to the randomized assignment was imperfect. We therefore display intention-to-treat estimates (ITT, panel A) as well as local average treatment effects (LATE, panel B). Moreover, the probability of assignment to the different experimental groups varied locally. We control for this by introducing a set of dummy variables, one for each area in which a given set of assignment probabilities was used.[2]

Using only unemployment registers (column 1) yields the largest sample: 8137 job seekers. Indeed, there is no issue of non response: the unemployment register is an exhaustive source. Job seekers assigned to the public scheme exit unemployment registers significantly more than those assigned to the control group. The impact of the private scheme is not statistically significant. Of course, the economic interpretation of these results is unclear: we do not know whether the public scheme helped job seekers to find a job, of if their exit from registered unemployment is due to discouragement or tighter monitoring. The rest of the table therefore uses more meaningful outcomes, based on our two main data sources. From the enriched administrative data (column 2), we find that the public scheme actually increased exit from unemployment to employment, whereas the private scheme had virtually no effect. By contrast, the long telephone survey tells a very different story, suggesting a positive and sizeable impact of the private scheme. Job seekers of the two treatment groups tend to be more frequently employed (in any job, including those who remain registered as job seekers while holding a job), even though the difference with the control group is not statistically significant (column 3). Once the quality of the job is taken into account, it appears that the private scheme increased the odds of finding a full-time job under an indefinite duration contract (column 4). The effect is sizeable and highly significant. The corresponding effect of the public scheme is smaller and marginally significant.

The contrast between these results could be rationalized. After all, the two data sources measure different outcomes. Moreover, the incentives of private firms hired by Unédic were precisely rewarding full-time, long-term jobs. A quick conclusion based on table 2 would be that the private scheme was only effective at helping job seekers to find high-quality jobs, whereas the public scheme was effective at helping them find any type of jobs.

[2] Alternatively, one can use design weights; the results are similar.

This, however, does not take into account the other difference between the two data sources: they have very different response rates. Results in column 5 suggests that this matters a lot. In order to control for differences in measured outcomes and to focus on non-response issues, we artificially generate an outcome variable that has the same meaning as in the enriched administrative source, but has the same response distribution as in the long telephone survey. Namely, we create a second variable measuring exit from unemployment to employment by combining part of the information of the unemployment registers with the information from the long telephone survey. This variable is set to missing if the job seeker did not respond to the survey, to 1 if the job seeker declared to be employed in the long telephone survey *and had left the unemployment registers*, to 0 otherwise. In other words, we simply correct for the fact that some of the employed people in column 3 are still on the unemployment registers, as long as they are still looking for a better job. Columns 2 and 5 therefore measure the same outcome (exit from registered unemployment towards employment), but with very different response rates.[3] Their comparison therefore tends to isolate the impact of the difference in response rates. This difference proves to matter a lot: with the enriched administrative data, the private scheme has no significant effect (the point estimate is 3.2 percentage points and the upper bound of the confidence interval for the LATE is 15.5); with the long telephone survey, the estimated effect is large and statistically significant (20.7 with a standard error of 6.4).

Our interpretation of these naive estimations is that, though the two schemes may well have differentiated impacts on different outcomes, there is a strong suspicion that part of the apparent positive impact of the private scheme on the probability to find "high-quality" jobs is spuriously driven by a sample selection bias. The next section provides further evidence of this.

# 4 Evidence on sample selectivity

We start with some notations using the potential outcome framework. For the sake of simplicity and without any loss of generality, let us consider two experimental groups only. $Z \in \{0,1\}$ denotes the random assignment variable while $T \in \{0,1\}$ denotes the treatment variable. We then define potential outcomes for the different assignments ($Z = 0$ and $Z = 1$). In general, the outcome may depend on the realized treatment $T$ as well as on the assignment $Z$. However, for each individual, the realized treatment itself is a function of the assignment. This can be

[3]Note however that the two "exit to employment" variables in columns 2 and 5 may still differ because they are not measured exactly at the same date. In particular, a job seeker who exited the unemployment registers, found a job, but lost it by the time of the long telephone survey, will be recorded as exiting to employment in column 2, but not in column 5. When assessing the impact of sample selection in the next section, we overcome this remaining difficulty by measuring exit to employment in differently selected subsamples, but always only using the measure of "exit to employment" provided by the enriched administrative dataset.

summarized in the function $\widetilde{y}(Z)$:[4]

$$\widetilde{y}(Z) \equiv y(Z, T(Z)).$$

The observed outcome is $y = \widetilde{y}(1)Z + \widetilde{y}(0)(1 - Z)$.

In this section, we focus on intention-to-treat effects, in particular on the impact of $Z$ on exit to employment, denoted by the indicator variable $y$. The naive ITT estimators of table 2 come from the OLS estimation of:

$$y = \beta_0 + \beta_1 Z + \epsilon. \tag{1}$$

If there was no sample selection issue, we would have $\beta_1 = E[\widetilde{y}(1) - \widetilde{y}(0)]$. This does not need be the case when the model is estimated on a selected sample. The goal of this section is therefore to clarify, and if possible to test, to what extent the puzzling discrepancy between results based on the two surveys may be due to sample selection.

Sample selection may create two sorts of problem. The first one is a problem of *external validity* of the naive estimates: as they are estimated on different samples representing different subpopulations, the naive ITT estimators may be internally valid (they consistently estimate the effect of the treatment for a population of respondents to a given survey) but lack external validity (the effect differs from the effect in the complete population). The second problem is a problem of *internal validity*: respondents in the control group differ from respondents in the treatment group, making the comparison irrelevant. To formalize this distinction, we need to introduce some further notations on response behavior. Response behavior itself may depend on assignment. In the same way as for outcomes, we define two potential response variables, $\widetilde{R}(0)$ and $\widetilde{R}(1)$, one for each assignment ($Z = 0$ and $Z = 1$, respectively). The observed response behavior is $R = \widetilde{R}(1)Z + \widetilde{R}(0)(1 - Z)$. With these notations, we can define different parameters of interest. The first one is the usual intention-to-treat parameter:

$$ITT = E(\widetilde{y}(1) - \widetilde{y}(0)).$$

The next two are ITT parameters for two populations of respondents:

$$ITT_{R1} = E(\widetilde{y}(1) - \widetilde{y}(0)|\widetilde{R}(1) = 1)$$

and

$$ITT_{R0} = E(\widetilde{y}(1) - \widetilde{y}(0)|\widetilde{R}(0) = 1).$$

As will become clear in section 5, $ITT_{R1}$ is a more interesting parameter in our problem. We therefore focus on $ITT_{R1}$ and $ITT$. With these notations, we will say that the OLS estimator

[4] As long as we focus on ITT effects, we do not need to introduce the usual potential outcomes $y(T)$ that depend on the *treatment* status $T$. Also, we do not need to make exclusion restrictions such as $y(Z, T(Z)) = y(T(Z))$. We introduce these additional notations and assumptions in the next section, where they become necessary.

for $\beta_1$ in (1) on a given sample is internally valid if it is consistent for $ITT_{R1}$, and externally (and internally) valid if it is consistent for $ITT$.

The conflicting results in table 2 are a clear indication that sample selection creates a problem of external validity for the naive estimation in at least one of the two samples. Indeed, if the naive estimators based on each of the two surveys are consistent for $ITT$, we expect them to be very close (note that their estimation samples partly overlap): this is not the case for the impact of the private scheme. The main question that remains is whether there is also a problem of internal validity. Can we provide evidence that the naive estimators lack internal validity, in particular in the long telephone survey where response rates are so low ? For this, contrasting the estimates based on the two surveys is not sufficient: if the effect of the treatment is heterogenous, it might well be the case that $E(\widetilde{y}(1) - \widetilde{y}(0)|\widetilde{R}_{tel}(1) = 1) \neq E(\widetilde{y}(1) - \widetilde{y}(0)|\widetilde{R}_{adm}(1) = 1)$ (the subscript *adm* and *tel* characterize response in the administrative data and in the long telephone survey, respectively); therefore, the estimators in columns 2 and 5 of table 2 need not converge to the same value.

**Sufficient condition for internal validity: non response ignorability**

However, what we can do is to test one of the natural conditions for internal validity, which we call "non-response ignorability" (by analogy with the notion of "treatment ignorability"):

$$\widetilde{y}(0) \perp (\widetilde{R}(1), \widetilde{R}(0)).$$

Under non response ignorability, the naive estimators are internally valid, in the sense that they consistently estimate $ITT_{R1}$. [5]

Non response ignorability is usually not testable. Indeed, the distribution of $\widetilde{y}(0)$ is typically only observed for those with $\widetilde{R}(0) = 1$. The advantage of having two surveys is to make a test of non response ignorability possible. Indeed, if non response is ignorable in the two surveys,

$$\widetilde{y}(0) \perp (\widetilde{R}_{adm}(1), \widetilde{R}_{adm}(0), \widetilde{R}_{tel}(1), \widetilde{R}_{tel}(0)),$$

then we have

$$E(\widetilde{y}(0)|\widetilde{R}_{adm}(0) = 1, \widetilde{R}_{tel}(0) = 1) = E(\widetilde{y}(0)|\widetilde{R}_{adm}(0) = 1, \widetilde{R}_{tel}(0) = 0).$$

These two quantities are identified, as $\widetilde{y}(0)$ is observed whenever $\widetilde{R}_{adm}(0) = 1$. Consequently, we can test this equation by restricting the sample to the part of the control group for which we have the administrative information and by testing if

$$E(y|R_{tel} = 1, Z = 0) = E(y|R_{tel} = 0, Z = 0).$$

[5] The naive estimators are consistent for $E(y|Z = 1, R = 1) - E(y|Z = 0, R = 1)$. Randomization and non response ignorability then imply $E(y|Z = 1, R = 1) - E(y|Z = 0, R = 1) = E(\widetilde{y}(1)|\widetilde{R}(1) = 1) - E(\widetilde{y}(0)|\widetilde{R}(0) = 1) = ITT_{R1}$.

The results are displayed on the first line of table 3. With p-value below 1%, we clearly reject non response ignorability in the long telephone survey, based on the information from the administrative dataset. More specifically, those who do not respond to the long telephone survey have a much higher probability to exit unemployment to employment (in the absence of treatment) than those who respond: 39.6% compared to 31.2%. As a consequence, estimations among respondents to the long telephone survey strongly *underestimate* the rate of exit to employment when there is no intensive counseling.

**Other conditions for internal validity**

It should be noted, however, that non response ignorability is only a sufficient, but not a necessary condition for internal validity. Formally, the necessary and sufficient condition for internal validity of the naive estimator is

$$E(\widetilde{y}(0)|\widetilde{R}(1)=1) = E(\widetilde{y}(0)|\widetilde{R}(0)=1).$$

Unfortunately, even with two surveys, testing this condition directly is not possible, as it involves a moment from the joint distribution of $\widetilde{y}(0)$ and $\widetilde{R}(1)$, two potential outcomes that are never simultaneously realized. The condition is therefore impossible to test directly, and also hard to interpret.

Instead, we can review plausible cases where internal validity holds despite the fact that non response is not ignorable, and check whether they are compatible with the evidence. The first obvious one is when $\widetilde{R}(1) = \widetilde{R}(0)$ for everyone: those who respond when they are in the control group are exactly the same as those who respond when they are in the treatment group. Then $E(\widetilde{y}(0)|\widetilde{R}(1)=1) = E(\widetilde{y}(0)|\widetilde{R}(0)=1)$ is trivial, and the naive estimator consistently estimates the ITT among respondents. However, $\widetilde{R}(1) = \widetilde{R}(0)$ also implies $E(\widetilde{R}(1)=1) = E(\widetilde{R}(0)=1)$ which can be tested as $E(R|Z=1) = E(R|Z=0)$. Table 4 shows that this is rejected by the data for the long telephone survey: job seekers respond more when they are assigned to the private scheme than when assigned to the control group (the response rate increases by 5.6 percentage points, with a standard error of 2.1). The difference is also positive, though smaller and not statistically significant, for the public scheme.

A second natural case to consider is the case where the effect of the treatment is homogeneous and the sample selection has the same impact on the control group as on the treatment group. In other words, though non response is not ignorable ($E(\widetilde{y}(0)|\widetilde{R}(0)=1) = E(\widetilde{y}(0)) + \delta$ with $\delta \neq 0$), it has no impact on the estimation of the effect of treatment as it impacts the treatment group in the same way: $E(\widetilde{y}(1)|\widetilde{R}(1)=1) = E(\widetilde{y}(1)) + \delta$. When comparing the control and the treatment group, the impact of selection disappears (the two $\delta$'s cancel out). We can use the two surveys to test if the data supports this story. Taking the administrative survey as a benchmark, we can test whether

$$
\begin{aligned}
&E(\widetilde{y}(0)|\widetilde{R}_{adm}(0)=1,\widetilde{R}_{tel}(0)=1)-E(\widetilde{y}(0)|\widetilde{R}_{adm}(0)=1,\widetilde{R}_{tel}(0)=0)\\
&=E(\widetilde{y}(1)|\widetilde{R}_{adm}(1)=1,\widetilde{R}_{tel}(1)=1)-E(\widetilde{y}(1)|\widetilde{R}_{adm}(1)=1,\widetilde{R}_{tel}(1)=0)\\
&=\delta
\end{aligned}
$$

This is done by comparing the different lines of table 3. The fifth column gives an estimate of $\delta$ for each of the three experimental groups. While the difference between respondents and the complete population is similar in the control group and in the treatment group assigned to the public scheme (-3.8 compared with -2.9 percentage points), it is much smaller and not statistically different from 0 for job seekers assigned to the private scheme (-0.2 percentage points). In other words, for the three experimental groups, looking at respondents only leads to under estimate the rate of exit to employment; but the degree of underestimation is significantly lower in the group assigned to the private scheme. This is evidence against the case of homogeneous treatment effects and similar impacts of sample selection across experimental groups. Note that sample selection does not appear to differ in the control group and in the group assigned to the public scheme. This echoes the finding of table 2 where estimates of the impact of the public scheme do not differ significantly across the two surveys. The problem of sample selection seems to specifically affect the job seekers assigned to the private scheme.

Of course, if we allow for heterogenous treatment effects, we may also imagine cases where the results of table 4 and 3 are compatible with the internal validity of the naive estimators. However, these cases would arguably be somewhat ad hoc. To sum up, there is a strong suspicion against the internal validity of the naive estimators in the long telephone survey: the rates of response differ markedly between the experimental groups, non response is correlated with the exit rate to employment in the absence of treatment, and the effect of sample selection seems to differ across experimental groups. If we were still to believe in the internal validity of the naive estimators to measure an effect on respondents, the lack of external validity would remain problematic: the results of table 2 would imply that the effect of the private scheme is much higher for respondents to the long telephone survey than for the average job seeker. Any conclusion on the efficiency of the program itself would therefore be misguided.

Though this is speculative, tables 4 and 3 suggest a story on the response behavior that may have created a bias in the long telephone survey. Start from table 3 where we see that job seekers with better chances to find a job are less likely to respond to the phone survey. One plausible explanation is that some job seekers have unobserved characteristics that make them more efficient in their job search, even without any help from the public employment system. They might therefore not value the help of the public employment system too much and not wish to respond to a survey it mandates. However, when they receive counseling from a private company, they may consider it as a useful help and become more willing to respond to surveys. This would explain why response rates raise with assignment to the private scheme in table 4. This change in response behavior induced by the treatment also modifies the composition of the

observable treatment group. Since the new respondents are more efficient in their job search, they drive the exit rate in the treatment group upward through a composition effect rather than through the causal effect of the intensive counseling. They create an upward bias. As for those assigned to the public scheme, they may be less sensitive to the change, as the service is still provided in house by the ANPE. The change in response behavior, hence the bias, would be lower (and not detected statistically in our data). Of course, this is just one story. However, it helps make the point that sample selectivity is a serious issue in the long telephone survey and that we need approaches that are robust to such selection. We turn to this in the next section.

# 5 Non response and bounds

## 5.1 Estimators

Manski and Horowitz (2000) propose to use the fact that a variable is bounded to derive bounds on its mean when there is non response. In our case, this yields a worst-case scenario where all missing outcomes are set to 0 in the treatment group and to 1 in the control group, and a best-case scenario where they are set to 1 in the treatment group and to 0 in the control group. The width of the resulting identifiable interval for the effect of the treatment is equal to the sum of the non-response rates in the two groups. With response rates slightly above 50% and 80%, these bounds are uninformative.

Lee (2008) proposes an alternative approach that provides sharp bounds on ITT effects in the case where selection is monotonic with regard to random assignment. He applies these bounds to an experiment where selection is due to the fact that the outcome, wages, is only observed for those who get a job. However, he also suggests that his approach can be used in the case where selection is due to survey non response. To follow this route, we need to take into account two features of our data. First, the outcome itself is bounded: the employment status is a binary variable. This allows us to consider a slightly different estimand than Lee, with two advantages: the population covered can be larger and the bounds tighter. Second, there is imperfect compliance to the initial randomized assignment; we therefore extend the approach to the estimation of intention-to-treat effects (ITT) and to local average treatment effects (LATE) on "responding compliers", *i.e.* those who are induced to receive the treatment by the random assignment *and* who respond to the survey.

### 5.1.1 Bounds under perfect compliance

For the sake of simplicity, we proceed in two steps, first considering the case where compliance is perfect, so as to make the comparison with Lee's approach transparent. We consider the standard potential outcome framework in a simpler version than in the previous section. We define a treatment variable $T \in \{0, 1\}$, two potential outcomes $y(1)$ and $y(0)$ and assume random

assignment to treatment. We consider the case where there is non response in the output variable and we introduce potential response $R(1)$ and $R(0)$. $R(j)$ is the response behavior when assigned to treatment $j$. The response behavior is therefore $R = R(0)(1-T) + R(1)T$. In that context, it is perfectly feasible to use Lee's approach to derive bounds for the parameter $E(y(1)-y(0)|R(1)=1, R(0)=1)$, *i.e.* the average treatment effect for the "always respondents". However, given the fact that $y(0)$ is bounded here, we can also provide bounds for the parameter $E(y(1)-y(0)|R(1)=1)$, *i.e.* the average treatment effect for the "respondents if treated".

Noting $\Delta_R = E(y(1)-y(0)|R(1)=1)$[6], we have the following proposition :

**Proposition 1** *Assuming*

1. *The output variables are bounded:*

$$y(k) \in [m, M], \ k \in \{0,1\}$$

2. *Monotonicity of response behavior:*[7]

$$R(1) \geq R(0)$$

*then in the presence of non response the parameter $\Delta_R$ cannot be identified from the data but it belongs to an identifiable interval whose lower and upper bounds $\underline{\Delta}_R$ and $\overline{\Delta}_R$ are:*

$$\underline{\Delta}_R = \big(E(yR|T=1) - E(yR|T=0)\big)/E(R|T=1) - M\big(E(R|T=1) - E(R|T=0)\big)/E(R|T=1)$$
$$\overline{\Delta}_R = \big(E(yR|T=1) - E(yR|T=0)\big)/E(R|T=1) - m\big(E(R|T=1) - E(R|T=0)\big)/E(R|T=1)$$

The length of the interval is

$$\overline{\Delta}_R - \underline{\Delta}_R = (M-m)\big(E(R|T=1) - E(R|T=0)\big)/E(R|T=1).$$

It depends on the difference in the response rates and not on their sum as it is the case for Horowitz-Manski bounds which are obtained without any assumptions on the response behavior.

The proof of the proposition is in the appendix. The proposition itself calls for two comments. First, let us compare it to the bounds based on Lee (2008). Here, given the fact that the outcome is binary, the quantiles that appear in Lee's formulas are replaced by 0 or 1. Note $\overline{\Gamma}_R$ and $\underline{\Gamma}_R$ Lee's bounds for $E(y(1)-y(0)|R(1)=1, R(0)=1)$. The length of the interval is

$$\overline{\Gamma}_R - \underline{\Gamma}_R = (M-m)\big(E(R|T=1) - E(R|T=0)\big)/E(R|T=0).$$

[6]?? It is the same parameter as $ITT_{R1}$ in the previous part.

[7]Alternatively, monotonicity could be defined by $R(1) \leq R(0)$. In our application, however, response rates are higher in the treatment groups.

Clearly, when treatment status does not affect response behavior (i.e. $R(1) = R(0)$), the two intervals are identical and empty. In general, however, there are observations such that $R(1) > R(0)$, so that the interval is not empty and $\overline{\Delta}_R - \underline{\Delta}_R < \overline{\Gamma}_R - \underline{\Gamma}_R$. Moreover, under the assumption that $R(1) \geq R(0)$, the population of the "respondents if treated" includes the population of the "always respondents". To summarize, compared to $E(y(1) - y(0)|R(1) = 1, R(0) = 1)$, $E(y(1) - y(0)|R(1) = 1)$ covers a larger population and yields a smaller identifiable interval. Given this advantage, we decide to focus on that parameter. It should be noted, however, that this is made possible by the fact that $y(0)$ is bounded, allowing us to bound $E(y(0)|R(1) = 1, R(0) = 0)$. In a more general case where $y(0)$ is not bounded, Lee's bounds are the only choice.

A second point of comparison with Lee (2008) is to ask whether the monotonicity assumption is plausible here. The assumption may at first seem innocuous. As noted by Lee (2008), this assumption is embedded in any standard model of sample selection that writes an additive latent selection model such as

$$R = 1 \textit{ if and only if } R^* \geq 0 \textit{ where } R^* = \alpha + \beta T + v.$$

Moreover, the appendix shows that the assumption of monotonicity can be stated in a way that more closely relates to our data collection process, that comprised multiple attempts to reach the job seeker. We call this the "extended monotonicity" assumption. The attempts to reach people involve purely random components (orthogonal to assignment). Once the job seeker is reached in a given attempt $j$, we make the behavioral assumption that his response behavior is monotonic with regard to assignment: $R_j(1) \geq R_j(0)$. In this slightly more general framework, the results of proposition 1 still hold. However, we have to assume monotonic response behavior for each attempt $j$ and that the probability to reach the job seeker is not affected by the assignment.

Nevertheless, this model of response behavior can still be problematic in our application. Assume for instance that the impact of treatment on response behavior is twofold: (i) treated job seekers are more satisfied with the public employment system, and therefore more prone to respond to surveys; (ii) treated job seekers are more likely to find a job, to move, and are therefore less likely to be reached for the interview. If there is heterogeneity in these effects, one can imagine that $R(1) > R(0)$ for job seekers for whom effect (i) dominates, and $R(1) < R(0)$ when effect (ii) dominates.

The bottom line is that the monotonicity assumption cannot be taken for granted, even though it is clearly much less demanding than the assumption that non response is ignorable. We therefore consider results based on this assumption as tentative only. They can be seen from two perspectives: are the bounds useful to reconcile the contradictory results obtained when using the two data sources? More speculatively, if the monotonicity assumption holds, does it

enable us to recover interesting information from the long telephone survey despite its low and unbalanced response rates?

### 5.1.2 Bounds under imperfect compliance

As a second step, we now extend the framework to the case of imperfect compliance. This setting is frequently used in practice when conducting an experiment. It is for example the case with the so-called encouragement design (see Duflo, Glennerster, and Kremer, 2007). We consider the potential outcome framework with random assignment to treatment and imperfect compliance of Angrist, Imbens and Rubin (1996). $Z \in \{0,1\}$ is the variable related to assignment and $T \in \{0,1\}$ is the final treatment status. The potential treatment variables are $T(0)$ and $T(1)$ (corresponding to $Z=0$ or $Z=1$, respectively). Potential outcomes are $y(t,z)$, with $t \in \{0,1\}$ and $z \in \{0,1\}$. We consider the usual set of assumptions of the Angrist, Imbens and Rubin model:

**Assumption 1** *1. SUTVA*

*2. Monotonicity* $T(1) \geq T(0)$

*3. Exclusion* $y(t,z) = y(t)$

*4. Randomness* $y(1), y(0), T(1), T(0) \perp Z$

It is well known that under this set of assumptions, the usual Wald estimator identifies the Local Average Treatment Effect

$$\Delta_{Late} = \frac{E\left(y\left|Z=1\right.\right) - E\left(y\left|Z=0\right.\right)}{E\left(T\left|Z=1\right.\right) - E\left(T\left|Z=0\right.\right)} = E\left(y\left(1\right) - y\left(0\right)\left|T\left(1\right) - T\left(0\right) = 1\right.\right)$$

We consider the case where there is non response and we introduce potential response behavior $R\left(t,z\right)$. We make the following assumptions

**Assumption 2** *1. Exclusion* $R(t,z) = R(t)$

*2. Monotonicity of response behavior*

$$R(1) \geqslant R(0)$$

The response behavior only depends on the treatment and not on the assignment to treatment. The parameter of interest we consider is the Local Average Treatment Effect on "responding compliers": $\Delta_{Late,R} = E(y(1) - y(0)|(T(1) - T(0))R(1) = 1)$.

**Proposition 2** *Under Assumption 1 and Assumption 2, the parameter $\Delta_{Late,R}$ cannot be identified but it belongs to an identifiable interval which Lower and Upper bounds $\underline{\Delta}_{Late,R}$ and $\overline{\Delta}_{Late,R}$ defined by*

$$\underline{\Delta}_{Late,R} = \frac{E(yR|Z=1) - E(yR|Z=0)}{E(TR|Z=1) - E(TR|Z=0)} - M\frac{E(R|Z=1) - E(R|Z=0)}{E(TR|Z=1) - E(TR|Z=0)}$$

*and*

$$\overline{\Delta}_{Late,R} = \frac{E(yR|Z=1) - E(yR|Z=0)}{E(TR|Z=1) - E(TR|Z=0)} - m\frac{E(R|Z=1) - E(R|Z=0)}{E(TR|Z=1) - E(TR|Z=0)}$$

The proof is in the appendix. Notice that the size of the interval is

$$\overline{\Delta}_{Late,R} - \underline{\Delta}_{Late,R} = (M-m).\frac{E(R|Z=1) - E(R|Z=0)}{E(TR|Z=1) - E(TR|Z=0)}$$

As in the previous case, it is a function of the difference in the response rates but here in addition it depends on the compliance rate for the respondents. A low rate of compliance leads to a widening of the bounds.

For inference purposes, note that the bounds on the ATE among responding compliers have the form of Wald estimands. For instance, we can rewrite

$$\underline{\Delta}_{Late,R} = \frac{E((y-M)R|Z=1) - E((y-M)R|Z=0)}{E(TR|Z=1) - E(TR|Z=0)}$$

This quantity can be estimated by the Wald estimator of the impact of $TR$ on $(y-M)R$ using $Z$ as an instrument. Standard errors follow as usual.

To summarize, Lee's approach can be extended and applied to our problem of non response with imperfect compliance. Some gains can be made using the fact that the outcome is bounded. The generalization to LATE is fairly direct, although imperfect compliance leads to larger identifiable intervals. The main caveat is that the monotonicity assumption is arguably a strong assumption in that context.

### 5.2 Results

The resulting bounds are displayed on table 5, which parallels the naive estimations of table 2. Arguably, selection is less likely to be incidental with the enriched administrative data, and bounds are not needed. If we do use bounds, this does not substantially modify the previous findings: as the response rates are very close in the different experimental groups, the estimated identifiable intervals are narrow. Like in table 2, the private scheme has no detectable effect on

exit from unemployment registers to employment. The effect of the public scheme is positive, statistically significant and sizeable: the lower bound of the LATE estimate shows that exit has increased by 13.6 percentage points (standard error of 5.3), which corresponds to an increase of about 1/3 starting from an exit rate around 35%.

By contrast, using bounds considerably modifies the message from the long telephone survey. The wide estimated intervals are due to two things: first, the large difference in response rates between job seekers assigned to the private scheme (and, to a lesser extent, to the public scheme) and job seekers assigned to the control group; second, the low take-up of the treatment. The lower bound of the effect even becomes negative when considering the impact of the private scheme on permanent employment (full-time employment under indefinite duration contract). However, the corresponding upper bound is also very high (21.7 percentage points with a standard error of 8.7 percentage points). We can neither reject the hypothesis that the program has been highly effective, nor rule out that it has been detrimental. The same applies for other outcomes and to the impact of the public scheme. Clearly, in the case of the long telephone survey, using bounds to get estimates that are robust to (monotonous) non response leads to uninformative results.

The comparison between estimates in columns 1 and 4 is of particular interest: what becomes the puzzle of conflicting results between the two surveys when sample selectivity is taken into account? The estimated identifiable intervals for the two programs now overlap (even without taking into account sampling error): the contradiction disappears, even if we assume that the effects are homogeneous. This is reassuring, but it comes at a high cost: taking into account sample selection has reconciled the two surveys, but it has also recognized that the long telephone survey could hardly be used. This means that we cannot test the prediction that the private scheme was specifically effective in raising employment in stable and full-time jobs, as implied by the incentives set by Unédic.

# 6 Conclusion

The previous section has explored statistical solutions to the issue of non response in the context of a large job search experiment. The results are mixed: even though the bounding approach can reconcile initially puzzling results, it does not yield informative estimates on interesting outcomes such as transition to stable, full-time jobs. Moreover, we had to assume non response monotonicity, which may be a strong assumption. This concluding section explores the consequences for data collection strategies.

There is a strong tendency in the literature to rely on increasingly rich administrative data sources, as exemplified by Swedish and Norwegian administrative data. In our case, given the limitations of the unemployment registers, this has proven useful as long as we have been able to supplement insufficient information through a short phone survey. However, this is far from being a panacea, for three reasons. First, non response to the survey remains a potential issue,

although we found no evidence of sample selection bias in our enriched administrative data set. Second, the outcome measured (transition from registered unemployment to employment), though relevant, remains a limited one. A possibility would be to rely on other administrative sources: social security records, matched with our sample, could in principle provide some information on the type of job. However, and this is the third limitation, there are delays in obtaining and matching such data. In the context of our experiment, ANPE and Unédic needed results quickly. Though gathering more and better administrative data is a route we are pursuing in this research, it cannot be the solution for the (short-term) political demand.

The comparison of the enriched administrative data and the long telephone survey suggests that surveys may sometimes be irreplaceable to get information on detailed outcomes (quality of the job, type of contract, for instance). Surveys therefore still have a role to play in experimental evaluations. As we have learned that non response can make these surveys useless, the question is whether we can implement them in a way that minimizes the consequence of non response. We believe that it could be useful to design surveys so as to be able to identify sample selection models. This requires instruments that impact response behavior but are independent from treatment assignment. In on-going work, we show that these instruments can be generated following the same basic principle as for treatment assignment: randomization. The idea is to randomly vary the effort with which information is obtained. If this generates sufficient (random) differences in response rates, pointwise identification of treatment effects becomes possible again. Applying this idea in practice is the next challenge.

The randomized trial literature has helped to renew the field of microeconometric policy evaluation by emphasizing identification issues raised by endogenous program participation. Measurement and attrition issues have received less attention. The lesson of this job search experiment is that the measurement of the outcomes and the response rates can significantly impact the results. Attrition bias should therefore be as serious a concern as endogenous program participation. We believe that the same inventiveness shown in tackling the latter problem could be usefully directed toward designing surveys and developing estimators to address the former.

## Appendix: Proof of propositions 1-2

We prove results in the more general case where response monotonicity is replaced by *extended monotonicity*. Results under monotonicity hold *a fortiori*. Extended monotonicity is defined as follows:

**Assumption 3** ***Extended Monotonicity***
*The potential response behaviors for j in (0,1) are of the form*

$$R(j) = G\left(\underline{B}(j), \underline{\Delta}(j)\right)$$

*with* $\underline{B}(j)' = (B_1(j), \ldots, B_K(j))$ *and* $\underline{\Delta}(j)' = (\Delta_1(j), \ldots, \Delta_L(j))$ *for some K and L. Where :*

- $R(j) \in \{0,1\}$
- *G is non decreasing in each component* $B_k(j)$
- $B_k(j)$ *satisfy monotonicity :* $B_k(1) \geq B_k(0)$
- $\underline{B}(j)$ *are independent from T*
- $\underline{\Delta}(j)$ *are ignorable for* $j \in \{0,1\}$
- $\underline{\Delta}(0)$ *and* $\underline{\Delta}(1)$ *have the same distribution*

One example could be the following : there are K attempts to reach people and the $B_k(j) \in \{0,1\}$ could be the desired response behavior at the $k^{th}$ attempt, but there may be purely random departure from this desired response behavior, and the observed response is $\Delta_k(j)B_k(j)$ where the components $\Delta_k(j) \in \{0,1\}$ is purely random:

$$\begin{aligned} R(j) &= \Delta_1(j) B_1(j) + (1 - \Delta_1(j) B_1(j)) \Delta_2(j) B_2(j) + (1 - \Delta_1(j) B_1(j)) (1 - \Delta_2(j) B_2(j)) \Delta_3(j) B_3(j) \\ &+ \cdots + \Big( \prod_{k=1}^{K-1} (1 - \Delta_k(j) B_k(j)) \Big) \Delta_K(j) B_K(j) \end{aligned}$$

In this case we have $1 - R(j) = \prod_{k=1}^{K} (1 - \Delta_k(j) B_k(j))$, and it is easy to see that the response behavior satisfies non decreasing requirement.

Another example could be $R(j) = 1(B(j) \geq \Delta(j))$.

**Proof** of Proposition 1

The parameter $\Delta_R$ writes as

$$\begin{aligned}
\Delta_R &= E\left(y\left(1\right)-y\left(0\right)|R\left(1\right)=1\right) \\
&= E\Big(y\left(1\right)R\left(1\right)-y\left(0\right)R\left(1\right)\Big)\Big/E\left(R\left(1\right)\right) \\
&= E\Big(y\left(1\right)R\left(1\right)-y\left(0\right)\left(R\left(0\right)+R\left(1\right)-R\left(0\right)\right)\Big)\Big/E\left(R\left(1\right)\right) \\
&= \Big(E\left(yR|T=1\right)-E\left(yR|T=0\right)\Big)\Big/E\left(R|T=1\right)-E\Big(y\left(0\right)\left(R\left(1\right)-R\left(0\right)\right)\Big)\Big/E\left(R\left(1\right)\right)
\end{aligned}$$

Given the expression for the response rate,

$$\begin{aligned}
E\Big(y\left(0\right)R\left(j\right)\Big) &= E\Big(y\left(0\right)G\left(\underline{B}\left(j\right),\underline{\Delta}\left(j\right)\right)\Big)=E\Big(E\Big(y\left(0\right)G\left(\underline{B}\left(j\right),\underline{\Delta}\left(j\right)\right)\Big|y\left(0\right),\underline{B}\left(j\right)\Big)\Big) \\
&= E\Big(y\left(0\right)\int G\left(\underline{B}\left(j\right),x\right)f\left(x\right)\Big)
\end{aligned}$$

where $f\left(x\right)$ is the distribution of $\underline{\Delta}\left(j\right)$ and does not depend on $j$ by assumption. Let $\Psi\left(X\right)=\int G\left(X,x\right)f\left(x\right)$. As $G$ is non decreasing in each component of $X$ so is $\Psi$. We thus have

$$E\Big(y\left(0\right)R\left(j\right)\Big)=E\Big(y\left(0\right)\Psi\left(\underline{B}\left(j\right)\right)\Big)$$

and

$$E\Big(y\left(0\right)\left(R\left(1\right)-R\left(0\right)\right)\Big)=E\Big(y\left(0\right)\left(\Psi\left(\underline{B}\left(1\right)\right)-\Psi\left(\underline{B}\left(0\right)\right)\right)\Big)$$

Given monotonicity conditions $B_k\left(1\right)\geq B_k\left(0\right)\geq 0$, and $\Psi$ non decreasing in each component, we have $\Psi\left(\underline{B}\left(1\right)\right)-\Psi\left(\underline{B}\left(0\right)\right)\geq 0$. Thus as $m\leq y\left(0\right)\leq M$, we have

$$mE\big(\Psi\left(\underline{B}\left(1\right)\right)-\Psi\left(\underline{B}\left(0\right)\right)\big)\leq E\Big(y\left(0\right)\left(R\left(1\right)-R\left(0\right)\right)\Big)\leq ME\big(\Psi\left(\underline{B}\left(1\right)\right)-\Psi\left(\underline{B}\left(0\right)\right)\big)$$

and thus

$$mE\left(R\left(1\right)-R\left(0\right)\right)\leq E\Big(y\left(0\right)\left(R\left(1\right)-R\left(0\right)\right)\Big)\leq ME\left(R\left(1\right)-R\left(0\right)\right)$$

which yields the result.

**Proof** of Proposition 2

Consider $E(yR|Z=1) - E(yR|Z=0)$. Given the assumption made, we have

$$\begin{aligned} E(yR|Z=1) &- E(yR|Z=0) \\ &= E(y(1)R(1)T(1) + y(0)R(0)(1-T(1))|Z=1) \\ &- E(y(1)R(1)T(0) + y(0)R(0)(1-T(0))|Z=0) \\ &= E(y(1)R(1)T(1) + y(0)R(0)(1-T(1))) \\ &- E(y(1)R(1)T(0) + y(0)R(0)(1-T(0))). \end{aligned}$$

This can be easily rewritten as

$$\begin{aligned} E(yR|Z=1) &- E(yR|Z=0) \\ &= E((y(1)-y(0))(T(1)-T(0))R(1) \\ &+ y(0)(R(1)-R(0))(T(1)-T(0))). \end{aligned}$$

Considering $y(1) = y(0) = 1$ leads to

$$\begin{aligned} E(R|Z=1) &- E(R|Z=0) \\ &= E((R(1)-R(0))(T(1)-T(0))). \end{aligned}$$

Moreover,

$$\begin{aligned} E(TR|Z=1) &- E(TR|Z=0) \\ &= E(T(1)R(1) - T(0)R(1)) \\ &= E((T(1)-T(0))R(1)), \end{aligned}$$

thus

$$\begin{aligned} \frac{E(yR|Z=1) - E(yR|Z=0)}{E(TR|Z=1) - E(TR|Z=0)} &= E((y(1)-y(0))|(T(1)-T(0))R(1)=1) \\ &+ \frac{E(y(0)(R(1)-R(0))(T(1)-T(0)))}{E(TR|Z=1) - E(TR|Z=0)}. \end{aligned}$$

Notice also that given monotonicity of the treatment $T(1) \geq T(0)$, we have $E((T(1)-T(0))R(1)) \geq 0$ and therefore $E(TR|Z=1) - E(TR|Z=0) \geq 0$.
Like for proposition 1 we have

$$E\Big(y(0)(R(1)-R(0))(T(1)-T(0))\Big) = E\Big(y(0)(\Psi(\underline{B}(1)) - \Psi(\underline{B}(0)))(T(1)-T(0))\Big)$$

where $\Psi(X) = \int G(X,x) f(x)\, dx$ with $f$ being the distribution of $\underline{\Delta}$.

Therefore, given monotonicity when $y(0) \in [m, M]$, we have

$$E\Big(y\,(0)\,\big(R\,(1)-R\,(0)\,\big)\big(T\,(1)-T\,(0)\,\big)\Big) \in$$
$$\Big[E\Big(m\,(\Psi\,(\underline{B}\,(1))-\Psi\,(\underline{B}\,(0)))\,(T\,(1)-T\,(0))\Big), E\Big(M\,(\Psi\,(\underline{B}\,(1))-\Psi\,(\underline{B}\,(0)))\,(T\,(1)-T\,(0))\Big)\Big]$$

that is

$$E\Big(y\,(0)\,\big(R\,(1)-R\,(0)\,\big)\big(T\,(1)-T\,(0)\,\big)\Big) \in$$
$$\Big[mE\Big(\big(R\,(1)-R\,(0)\,\big)\big(T\,(1)-T\,(0)\,\big)\Big), ME\Big(\big(R\,(1)-R\,(0)\,\big)\big(T\,(1)-T\,(0)\,\big)\Big)\Big].$$

In the end we obtain

$$E\Big(y\,(0)\,\big(R\,(1)-R\,(0)\,\big)\big(T\,(1)-T\,(0)\,\big)\Big) \in$$
$$\Big[m\Big(E\,(R\,|Z=1)-E\,(R\,|Z=0)\,\Big), M\Big(E\,(R\,|Z=1)-E\,(R\,|Z=0)\,\Big)\Big],$$

which yields the result.

## Bibliography

Angrist, J., G. Imbens, and D. Rubin, (1996) : "Identification of Causal Effects Using Instrumental Variables", *Journal of the American Statistical Association*, Vol. 91, No. 434, pp. 444-472.

Ashenfelter, O., A. Ashmore and O. Deschênes (2005) : "Do unemployment insurance recipients actively seek work? Evidence from randomized trials in four U.S. States", *Journal of Econometrics*, vol. 125(1-2), pages 53-75.

Card, D., R. Chetty, and A Weber (2007a): "The Spike at Benefit Exhaustion: Leaving the Unemployment System or Starting a New Job", NBER Working Paper 12893.

Card, D., R. Chetty, and A Weber (2007b): "The Spike at Benefit Exhaustion: Leaving the Unemployment System or Starting a New Job", *American Economic Review Papers and Proceedings*, 97(2), 113-118.

Duflo E., R. Glennerster and M. Kremer (2007) : "Using Randomization in Development Economics Research: A Toolkit", in T. Paul Schults, and John Strauss (eds.) *Handbook of Development Economics*, Elsevier Science Ltd.: North Holland, 2007 Vol. 4, pp. 3895-62.

Horowitz J., and C. Manski (2000) : "Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data", *Journal of the American Statistical Association*, Vol. 95, 2000

Meyer B. (1995) : "Lessons from the U.S. Unemployment Insurance Experiments", *Journal of Economic Literature*, 33(1), 91-131.

Table 1: Population sizes

| | Unemployment registers | Enriched adm data | Long telephone survey |
|---|---|---|---|
| Sample | 8137 | 6176 | 4295 |
| Weigthed population | 37 564 | 30 579 | 18 787 |
| Implied response rate | 100% | 81.2% | 50% |

**Note** : Response rates are computed using the sampling weights of the long telephone survey. In the case of the enriched administrative data, the weights are multiplied by the sampling weights of the short telephone survey.

Table 2: Program effects without accounting for sample selection

| | Unemployment registers | Enriched adm data | Long telephone survey sample | | |
|---|---|---|---|---|---|
| | Exit | Exit to Employment | Employment | Permanent Employment | Exit to Employment |
| | | | A. Intention-To-Treat | | |
| Private scheme | 2.9 | 0.9 | 6.6 | 8.4 | 8.8 |
| | (2.1) | (2.4) | (3.0) | (2.2) | (2.8) |
| Public scheme | 4.2 | 5.4 | 3.5 | 3.0 | 6.4 |
| | (1.3) | (1.9) | (2.2) | (1.5) | (2.1) |
| | | | B. LATE | | |
| Private scheme | 8.3 | 3.2 | 15.3 | 19.3 | 20.7 |
| | (5.7) | (6.1) | (6.9) | (5.0) | (6.4) |
| Public scheme | 11.6 | 14.0 | 8.4 | 7.1 | 15.5 |
| | (3.8) | (4.9) | (5.4) | (3.8) | (5.1) |
| nb observations | 8137 | 6176 | 4279 | 4279 | 4279 |

**Note** : Estimation is performed for the sample of respondents to each data source (unemployment registers, enriched administrative data and long telephone survey: see text). The LATE estimator is obtained by 2SLS instrumenting treatment status by treatment assignment. Robust standard errors are in parenthesis. Fixed effects are introduced to control for variations of assignment probabilities by areas and by periods. The estimation uses the sampling weights of the long telephone survey. In the case of the enriched administrative data, the weights are multiplied by the sampling weights of the short telephone survey.

Table 3: Exit to employment from the enriched administrative source depending on the response status in the long phone survey

| | Status in the long phone survey | | | | | |
|---|---|---|---|---|---|---|
| | Non respondents (nr) | Respondents (r) | All (a) | Difference (r-nr) | Difference (r-a) | p-value of the difference |
| Without intensive counseling | 39.6 (1.7) | 31.2 (2.1) | 35.0 (1.3) | −8.5 (2.7) | −3.8 (1.2) | 0.002 |
| With private scheme | 33.9 (2.1) | 33.4 (2.4) | 33.6 (1.6) | −0.5 (3.2) | −0.2 (1.4) | 0.876 |
| With public scheme | 44.1 (1.7) | 37.9 (2.0) | 40.8 (1.3) | −6.2 (2.6) | −2.9 (1.2) | 0.016 |

**Note** : Exit rates from unemployment registers to employment are computed based on the administrative data, comparing two subgroups: those who respond and those who do not respond to the long telephone survey. Standard errors are in parenthesis. The computation uses the sampling weights of the long telephone survey multiplied by the sampling weights of the short telephone survey and the design weights due to changing assignment probabilities.

Table 4: Impact of assignment on response

| | Enriched adm data | Long telephone survey |
|---|---|---|
| Private scheme | 0.0 | 5.6 |
| | (1.9) | (2.1) |
| Public scheme | −0.8 | 2.4 |
| | (1.7) | (1.5) |
| Size | 8137 | 8137 |

**Note** : Linear probability models for whether the employment status is available, in the enriched administrative data and in the long telephone survey, respectively. Fixed effects by constant probability areas and by periods have been introduced. Robust standard errors are in parenthesis. Estimations use sampling-weights of the long telephone survey multiplied by sampling-weights of the short telephone survey in the enriched administrative data.

Table 5: Bounds on treatment effects

| | Enriched adm data | Long telephone survey sample | | |
|---|---|---|---|---|
| | Exit to Employment | Employment | Permanent Employment | Exit to Employment |
| | | A. Intention-To-Treat | | |
| Private scheme | 0.8 ; 0.8<br>(2.5) (2.8) | 0.2 ; 11.1<br>(3.7) (3.4) | −0.9 ; 10.0<br>(4.1) (2.2) | 1.1 ; 12<br>(4.0) (2.9) |
| Public scheme | 5.2 ; 6.1<br>(2.1) (2.4) | 1.1 ; 5.8<br>(2.8) (2.5) | −1.1 ; 3.6<br>(3.0) (1.5) | 3.5 ; 8.3<br>(2.8) (2.2) |
| | | B. LATE | | |
| Private scheme | 3.4 ; 3.7<br>(6.4) (7.2) | 0.5 ; 25.0<br>(8.0) (7.1) | −2.8 ; 21.7<br>(8.7) (4.6) | 2.1 ; 26.7<br>(8.5) (6.1) |
| Public scheme | 13.6 ; 16.2<br>(5.3) (6.1) | 2.9 ; 17.5<br>(6.6) (6.0) | −5.9 ; 8.7<br>(7.4) (3.6) | 7.7 ; 22.4<br>(7.0) (5.5) |

**Note** : For each outcome (employment, permanent employment and exit to employment), the first number is the lower bound and the second number the upper bound of the treatment effect, taking into account non response under the assumption of non-response monotonicity. The resulting intervals are to be compared with the naive estimates of table 2, columns 2-5. Robust standard errors are in parenthesis. The LATE estimator is obtained by instrumenting treatment status by treatment assignment. Fixed effects are introduced to control for variations of assignment probabilities by areas and by periods. The estimation uses the sampling weights of the long telephone survey. In the case of the enriched administrative data, the weights are multiplied by the sampling weights of the short telephone survey.