Behaghel, Luc; Crépon, Bruno; Gurgand, Marc; Le Barbanchon, Thomas

Working Paper

Please call again: Correcting non-response bias in treatment effect models

IZA Discussion Papers, No. 6751

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Behaghel, Luc; Crépon, Bruno; Gurgand, Marc; Le Barbanchon, Thomas (2012) : Please call again: Correcting non-response bias in treatment effect models, IZA Discussion Papers, No. 6751, Institute for the Study of Labor (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/62568

DISCUSSION PAPER SERIES

IZA DP No. 6751

Please Call Again: Correcting Non-Response Bias in Treatment Effect Models

Luc Behaghel
Bruno Crépon
Marc Gurgand
Thomas Le Barbanchon

July 2012

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Please Call Again: Correcting Non-Response Bias in Treatment Effect Models

Luc Behaghel
Paris School of Economics (INRA), CREST, J-PAL and IZA

Bruno Crépon
CREST, J-PAL and IZA

Marc Gurgand
Paris School of Economics (CNRS), CREST, J-PAL and IZA

Thomas Le Barbanchon
CREST

Discussion Paper No. 6751
July 2012

IZA

P.O. Box 7240
53072 Bonn
Germany

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

IZA Discussion Paper No. 6751
July 2012

ABSTRACT

Please Call Again: Correcting Non-Response Bias in Treatment Effect Models*

We propose a novel selectivity correction procedure to deal with survey attrition, at the crossroads of the "Heckit" and of the bounding approach of Lee (2009). As a substitute for the instrument needed in sample selectivity correction models, we use information on the number of attempts that were made to obtain response to the survey from each individual who responded. We obtain set identification, but if the number of attempts to reach each individual is high enough, we can come closer to point identification. We apply our sample selection correction in the context of a job-search experiment with low and unbalanced response rates.

JEL Classification: C31, C93, J6

Keywords: survey non response, sample selectivity, treatment effect model, randomized controlled trial

Corresponding author:

Luc Behaghel
Paris School of Economics
48 Boulevard Jourdan
75014 Paris
France
E-mail: luc.behaghel@ens.fr

* We thank Laurent Davezies, Esther Duflo, and seminar participants at CREST for their comments on previous versions of this paper. We also thank ANPE, Unédic and DARES for their involvement in the experiment from which the data of the application are drawn, and for their financial support to the evaluation.

1 Introduction

Sample attrition is a pervasive issue for surveys in social sciences. The damage appears particularly clearly in randomized trials: while random assignment 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. This issue is serious in any type of treatment model, and practitioners have long been aware of the threat posed by data collection in that context.[1] Campbell (1969) lists "experimental mortality" (i.e., "the differential loss of respondents from comparison groups") as one of the nine threats to the internal validity of experimental and quasi-experimental designs. The concern is frequently raised in applied economics: examples include Hausman and Wise (1979) or the studies in the special issue of The Journal of Human Resources (spring 1998) dedicated to Attrition in Longitudinal Surveys.

The statistical and econometric literature has developed a variety of tools to deal with sample selectivity –of which attrition is one aspect– starting with seminal papers by Heckman (1976 and 1979) and turning less and less parametric up to the "worst-case", assumption-free approach developed by Horowitz and Manski (1995, 1998 and 2000). The toolbox also includes weighting procedures based on the assumption that data is "missing at random" conditional on some observables. Yet, applied economists may still feel perplexed in practice. While the virtues of conservative bounds à la Horowitz-Manski are clear, with commonly observed attrition rates above 15%, they yield quite wide identified sets. The other two approaches yield point identification. Yet missing-at-random assumptions are often hardly credible, given evidence in several empirical studies that attrition is correlated with the outcomes of interest[2]. Similarly, sample selectivity procedures face the practical difficulty of finding a credible instrument: "The practical limitation to relying on exclusion restrictions for the sample selection problem is that there may not exist credible 'instruments' that can be excluded from the outcome equation." (Lee, 2009, p. 1080). Middle-ground approaches such as Lee (2009) are a valuable compromise; in some instances however, they still yield quite large identified sets (e.g., Kremer et al., 2009).

[1]For instance, education scientist McCall writes in the early 1920s, before R.A. Fisher's reference book on *The Design of Experiments* (1935): "There are excellent books and courses of instruction dealing with the statistical manipulation of experimental data, but there is little help to be found on the methods of securing adequate and proper data to which to apply statistical procedures." (McCall, 1923)

[2]Evidence is obtained by confronting survey data to another more comprehensive data source (e.g. administrative data, or reports from parents or neighbors of non-respondents). See for instance Behrman, Parker and Todd (2009) or the application in section 3.

This paper proposes a novel and simple approach to correct sample selection bias resulting from non-response, at the crossroads of semi-parametric forms of the "Heckit" and of the bounding approach of Lee (2009). As a substitute for the instrument needed in sample selectivity correction models, we show that we can use basic information on the number of attempts that were made to obtain response to the survey from each individual who responded. The method can be applied whenever data collection entails sequential effort to obtain response, for instance trying to call several times in a phone survey, making several visits to the respondent's home, or even gradually offering higher incentives (gifts) to potential respondents. It does not require to randomize survey effort, as was proposed by DiNardo et al. (2006): as a result, there is no voluntary loss of information, as survey effort can be maximal for the whole sample. Further, correction can be made ex post, as long as the number of attempts have been recorded. We obtain non-parametric identification in a model of heterogenous treatment, provided that the selection mechanism can be represented by the latent variable threshold-crossing selection model, which is used in most of the literature.[3] Most of the time, only bounds are identified, but they are likely to be close when the number of attempts to obtain response is large enough.

The intuition of this result is the following. In the latent variable threshold-crossing selection model, individuals respond to the survey depending on an unobserved variable, call it V, that can be interpreted as reluctance to respond. Although this characteristic itself is unaffected by treatment status when treatment is exogenous (for instance when it is randomized), treatment may still affect one's response behavior. To fix ideas, assume that the response rate is lower in the control group. Respondent individuals in the treatment and control groups are thus different, with more "reluctant" (high V) individuals among respondents in the treatment group. This unobserved variable V can be correlated with counterfactual outcomes, implying potential bias.

A non-parametric way to neutralize this bias is thus to form a subset of treatment and control individuals that responded to the survey *and* have the same distribution of V. This is illustrated in figure 1. Because in the single index model people are ranked according to V, if 60% of the respondants in a group answered before the 18th call, they must have the 60% lowest values of V (those who responded after the 18th call or never responded must have the 40% highest values of V). If we have two groups (treatment and control) in which the answering behavior is different, it remains true that, if 60% of the respondants in the treatment group answered before the 18th call, and 60% of the respondants in the control group answered before the 20th call, each of these two subsamples contain the 60% lowest values of V within their group. When the groups are randomized (or can

[3]Vytlacil (2002) discusses the implications of that model.

be considered initially similar for any other reason), the 60% lowest V's in each group represent the same population. The insight of that paper is that when survey attempts are recorded, those two groups are identified in the data, and they can be compared with no risk of selection bias, because they have the same distribution of V. Bounds arise when we cannot find numbers of attempts that perfectly equalize response rates in the two groups. Implementation is very simple as it boils down to trimming the data. Naturally, the counterpart is that identification is local (restricted to the respondants) but this cannot be avoided when there is selection in the first place.

We apply our sample selection correction in the context of a controlled job-search experiment in which exhaustive administrative data is available, as well as a phone survey with richer information but low and unbalanced response rates. Using the administrative information, we reject that observations are missing-at-random in the phone survey, as attrition is correlated with the outcomes of interest. Moreover, point estimates suggest that using the phone survey sample we would over-estimate the program's impact by about 50%. Applying our sample selection correction procedure closes most of the gap between the estimates in the full and in the selected samples. Bounds *à la* Horowitz and Manski (2000) or Lee (2009) are, in this application, too wide to be very conclusive.

This paper contributes to a wide, 40-years old econometric literature on sample selection put in perspective by Manski (1989), and some 20 years later by Lee (2009). Lee (2009) illustrates how the two main approaches – the latent variable selection model and the bounding approaches – can converge when they use the same monotonicity assumptions on response behavior. An important distinction however, is whether the method implies using an instrument. The fact that an instrument is needed for the identification of the Heckit model not to depend on arbitrary parametric assumptions is often considered as a major drawback of the approach, as plausible instruments are rarely available. Our contribution to this literature is to show that the instrument may not be the issue: actual information on response conditions is enough to identify the same parameter as with an ideal instrument that would randomly vary the data collection effort across individuals.

However, our main contribution is probably to provide an additional, simple tool to practitioners. In that respect, it is related to proposals by Lee (2009) or DiNardo et al. (2006). The comparison with Lee (2009) is detailed in the paper. As noted above, Lee's procedure begins to be used in applied studies but may conduct to wide identified sets. DiNardo et al. (2006) propose to include randomization of survey effort in the design of surveys, as a way to generate instruments. They acknowledge that it may be difficult to persuade survey administrators to do so, and suggest to have only two levels of effort (for

instance, a maximum of 1 or 2 calls)[4], but do not recognize the fact that recording the actual number of calls or visits needed provides the same information.

The next section presents our sample selection correction approach. Section 3 gives an application, and section 4 concludes.

2 Sample selection correction using number of calls

In this section, we develop a new approach to sample selection correction using standard survey information on the data collection process. We first introduce the framework, recall and extend existing results on selection correction with instruments. We then present our approach, discuss its assumptions and compare it to the bounding approach.

2.1 Framework and notations

We use the potential outcome framework. $Z \in \{0, 1\}$ denotes the random assignment into treatment and $y(Z)$ is the potential outcome under assignment (or treatment) Z.[5] The observed outcome is $y = y(1)Z + y(0)(1 - Z)$. The parameter of interest is the average treatment effect:

$$E(y(1) - y(0)) \tag{1}$$

If Z is independent from $(y(0), y(1))$, as can be assumed under random assignment, then the average treatment effect can be estimated by comparing the empirical counterparts of $E(y|Z = 1)$ and $E(y|Z = 0)$. Alternatively, it is obtained by OLS estimation of:

$$y = \beta_0 + \beta_1 Z + \epsilon. \tag{2}$$

Attrition bias may arise from non observation of the value of y, resulting from non-response behavior (whether literally or as a result of missing observation in any sort of data). Define potential response under assignment Z as $R(Z) \in \{0, 1\}$. Just as for the outcome, $R(0)$ represents response behavior when a person is untreated and $R(1)$ when she is treated. Observed response behavior is $R = R(1)Z + R(0)(1 - Z)$.

[4]"until economists persuade data collection administrators of the value of sample selection correction, obtaining a binary [instrument for sample selectivity correction] will remain an ambitious goal practically, if not econometrically", DiNardo et al. (2006), p. 10.

[5]In this section, we assume perfect compliance: treatment is equal to assignment (equivalently, if there is imperfect compliance, we consider the intention-to-treat effect). Appendix A.3 provides an extension of our approach to the case with imperfect compliance.

When there is non-response, the observed mean value of y in treatment and control group measures $E(y|R=1,Z=1)$ and $E(y|R=1,Z=0)$ respectively. Therefore, the "naive" average treatment effect estimator (for instance the above OLS estimation on the respondents) measures:

$$\begin{aligned} E(y|R=1,Z=1) - E(y|R=1,Z=0) &= E(y(1)|R(1)=1) - E(y(0)|R(0)=1) \\ &= E(y(1)-y(0)) + \Delta_1 + \Delta_2, \end{aligned}$$

where the first equality obtains if Z is independent from $(y(0), y(1))$ and from $(R(0), R(1))$, and:

$$\begin{aligned} \Delta_1 &= E(y(1)-y(0)|R(1)=1) - E(y(1)-y(0)) \\ \Delta_2 &= E(y(0)|R(1)=1) - E(y(0)|R(0)=1). \end{aligned}$$

The first source of bias, Δ_1, results from treatment effect heterogeneity. It is present whenever the average treatment effect on those who respond to the survey ($R(1)=1$) is different from the effect in the whole population. The second source of bias, Δ_2, is a selection bias: it occurs whenever treated and control respondents are different in the sense that they have different counterfactuals $y(0)$. None of these terms can be directly estimated because they require $E(y(0)|R(1)=1)$ but $R(1)$ and $y(0)$ are not jointly observed.

Bias Δ_1 involves lack of external validity. Absent the selection bias, the "naive" estimator would be consistent for a population of respondents to a given survey, but may not extend to the general population. There is no way this bias can be avoided given only respondent outcomes are observed. In contrast, the problem raised by bias Δ_2 is one of internal validity. Even if our interest lies on $E(y(1)-y(0)|R(1)=1)$, this would not be estimated consistently if this second type of bias is present.

In the following, we restrict the parameter of interest to average treatment effect on respondents (or a subset of respondents), and we consider hypotheses under which the selection bias, if present, can be corrected.

Given the fundamental identification issue that characterizes the selection bias ($R(1)$ and $y(0)$ are not jointly observed), point identification of the causal treatment effect requires restrictions. Following Heckman (1976, 1979), a standard approach to sample selection correction relies on the latent selection model, whose identification requires instruments, i.e. determinants of selection (here response behavior) that do not determine the counterfactual outcomes. We present a semi-parametric version of that model but ar-

gue that proper instruments are difficult to find. We then show that, provided the survey records the number of calls after which individuals responded, identification is obtained based on that same model, even in the absence of instruments.

We assume the following latent variable threshold-crossing selection model:

Assumption 1

1. *(Latent variable threshold-crossing response model)*

$$R = \mathbf{1}(V < p(W, Z)), \tag{3}$$

2. *(Common support)* $p(W, 0)$ *and* $p(W, 1)$ *have non empty common support* $\boldsymbol{P}$ *as* W *varies. Denote* $\bar{p}$ *the upper bound of* $\boldsymbol{P}$*.*

Equation (3) adds some structure to the relation between response and treatment status, $R(Z)$. W is any variable related to response behavior, such as response incentives; as will prove useful in the following sections, W can also be thought as the maximum number of attempts of phone calls to survey one individual. p is an unknown function, and without any loss of generality, V follows a uniform distribution over $[0, 1]$, so that $p(W, Z)$ is the response rate as a function of W and Z. V is not observed and can be interpreted as the individual reluctance to respond to surveys. This latent variable threshold-crossing model is a fundamental element of the selectivity correction literature. Following Vytlacil (2002) equivalence result, the key embedded assumption is a form of monotonicity: individuals must not react in opposite ways to W and Z. [6]

2.2 Selectivity correction with instruments

We consider here that, in both treatment and control, W varies randomly accross observations. We have therefore the following independence assumption:

[6]Specifically, Vytlacil (2002) shows that the index model is equivalent to assuming the following condition: for all w, w', z, z', either $R_i(w, z) \geq R_i(w', z')) \ \forall \ i$, or $R_i(w, z) \leq R_i(w', z')) \ \forall \ i$. This monotonicity assumption is violated if assignment to treatment Z encourages some individuals to respond to the survey, but discourages some others. Also, the condition does not hold if some individuals are only sensitive to W, and some only to Z. Assume for instance that W takes only two values (each person is assigned to 1 or 2 attempts). There may be a person i_1 who responds to the first call anyway: $R_{i_1}(2, 0) < R_{i_1}(1, 1)$. By contrast, person i_2 is only available at the second call, but responds to the survey irrespective of treatment assignment: $R_{i_1}(2, 0) > R_{i_1}(1, 1)$. In that case, W and Z have monotonous impacts, but no single latent variable threshold-crossing response model exists.

Assumption 2

$$W, Z \quad \perp \quad y(0), y(1), V. \tag{4}$$

We then have the following identification result:

Proposition 1 ***: Identification with an instrument for response behavior.*** *Under assumptions 1 and 2, $E(y(1) - y(0)|V < \bar{p})$ is identified if there exists w_0 and w_1 in the data such that $p(w_0, 0) = p(w_1, 1) = \bar{p}$.*

Then:

$$E(y(1) - y(0)|V < \bar{p}) = E(y|R = 1, W = w_1, Z = 1) - E(y|R = 1, W = w_0, Z = 0). \tag{5}$$

This proposition builds on well-known properties of index selection models and adapts the semi-parametric identification results of Das, Newey and Vella (2003) to our setting with a binary regressor of interest and heterogeneous treatment effects. The proof is given in appendix A.1.1. To interpret equation (5), it is useful to think of V as an individual reluctance to respond to surveys. As W and Z are assigned randomly, V is equally distributed across treatment and control groups, and across different values of the instrument W. Given the response model in (3), the population of respondents is uniquely characterized by the value of $p(W, Z)$. If there are two couples of survey effort $(w_0, 0)$ and $(w_1, 1)$ such that $p(w_0, 0) = p(w_1, 1) = \bar{p}$, then these two couples are two different ways to isolate the same subpopulation of respondents. Therefore, comparing y across these two subpopulations (treated and controls) directly yields the impact of Z (on this specific subpopulation); i.e. the average treatment effect for those individuals with $V < \bar{p}$. Without further restrictions, we can only identify the parameter on the common support of respondents. The popular Heckman selectivity correction model is a version of this, with several parametric restrictions.

Of course, equation (5) is only useful to the extent that there exists such an instrument W, that varies sufficiently to cover the support of p. Unless this is planned in advance, it is usually extremely difficult to extract from the data some variables that have credible exogeneity properties and significant power to influence response. Therefore, as suggested above, randomizing survey effort could be a natural way to generate instrument W. One could for instance randomly assign the number of attempts to call each individual[7], or the

[7]Specifically, one could design the survey so that it is randomly decided that worker i will be called a maximum number of times W_i before considering him as a non respondent if he or she cannot be reached, and worker j will be called a maximum number of times W_j, etc.

value of the gift promised to those who respond. However, randomizing data collection effort amounts to wasting part of the sample (that on which data collection effort is not maximal). In most contexts, survey cost is high and the number of observations is limited; this may explain why, to the best of our knowledge, survey effort is not randomized in practice.

2.3 Realized number of calls as a substitute to instruments

We now consider the case when variable W is a maximum number of phone calls or home visits, or similar attempts to reach people, and it does not vary in the sample (everyone could potentially receive this maximum number of calls). We call N the number of attempts until a person is actually reached. The main insight of this paper is that W can be set to a value that is similar for all individuals in the sample: provided that the realization of N is recorded for each observation, the treatment impact is still identified. Therefore, ex ante randomization of W is not required and there is no consecutive loss of efficiency.

The empirical setup in this section is thus one where W is set to $W = w_{\max}$ for all individuals in the sample. Although assumption 2 still holds formally (but in a degenerate sense for W), it is clearer to state the following assumption:

Assumption 3

$$Z \quad \perp \quad y(0), y(1), V. \tag{6}$$

We also need to add some structure to the function $p(.,.)$:

Assumption 4

$$p(W, Z) \text{ is non-decreasing in } W, \quad \forall Z. \tag{7}$$

In the present context, where W is a maximum number of phone calls or visits, this is a very natural assumption, since contacts are iterative: the possibility to contact the person one more time should not decrease her chances to end up answering the survey. This is particularly reasonable is subjects are not aware of the planned maximum number of attempts.

Without loss of generality, consider the case where $p(w_{\max}, 0) \leq p(w_{\max}, 1)$, i.e. the share of respondents is higher among treated. Assume there exists w_1 such that:

$$p(w_{\max}, 0) = p(w_1, 1) \tag{8}$$

w_1 is a maximum number of calls that would be sufficient to obtain exactly the same response rate among the treated ($Z = 1$) as among the non-treated ($Z = 0$). Because of assumption 4, $w_1 \leq w_{\max}$. Notice that, if W was continous, w_1 would always exist. As W is in essence discrete in the present setup (number of calls or attempts), existence of such w_1 in practice is most unlikely. However, we assume here its existence for clarity in order to present the identification result in a simple case and give the basic intuition. In practice, discreteness will require identification of bounds rather than point estimates, and we will turn to this complication in the next section. For now, we have the following identification result:

Proposition 2 ***: Identification with information on the number of calls.*** *Under assumptions 1, 3 and 4, and if* $p(w_{max}, 0) \leq p(w_{max}, 1)$*,* $E(y(1) - y(0)|V < p(w_{max}, 0))$ *is identified if there exists* w_1 *in the data such that* $p(w_{max}, 0) = p(w_1, 1)$*.*

Then:

$$E(y(1) - y(0)|V < p(w_{max}, 0)) = E(y|N \leq w_1, Z = 1) - E(y|N \leq w_{max}, Z = 0), \tag{9}$$

where w_1 *is identified from*

$$\Pr(N \leq w_1|Z = 1) = \Pr(N \leq w_{max}|Z = 0), \tag{10}$$

and the set of individuals with $N \leq w_1$ *is observable because* $w_1 \leq w_{max}$*.*

If $p(w_{max}, 0) \geq p(w_{max}, 1)$*, then define similarly* w_0 *such that* $p(w_0, 0) = p(w_{max}, 1)$ *and*

$$E(y(1) - y(0)|V < p(w_{max}, 1)) = E(y|N \leq w_{max}, Z = 1) - E(y|N \leq w_0, Z = 0), \tag{11}$$

where w_0 *is identified from*

$$\Pr(N \leq w_{max}|Z = 1) = \Pr(N \leq w_0|Z = 0). \tag{12}$$

The proof is in appendix A.1.2 and results on convergence and inference are in appendix A.2.1. The result is valid for any maximum number of calls $w_{\max}$. If $w_{\max}$ is set high enough, we could have $\max p(w_{\max}, 0), p(w_{\max}, 1) = \bar{p}$ and we would identify $E(y(1) - y(0)|V \leq \bar{p})$ just as before.

To understand this identification result, take the case $p(w_{\max}, 0) \leq p(w_{\max}, 1)$. Equation (9) means that $E(y(1) - y(0)|V \leq p(w_{\max}, 0))$ is point identified by simply truncating the treatment sample. The average outcome of those control individuals that respond before the $w_{\max}$th call (all respondents in that case) is $E(y|N \leq w_{\max}, Z = 0)$ and this identifies $E(y(0)|V < p(w_{\max}, 0))$. And the average outcome of those treatment individuals that respond before the w_1th call (a subset of respondents in that case) is $E(y|N \leq w_1, Z = 1)$ and this identifies $E(y(1)|V < p(w_{\max}, 0))$. The sample selection problem is due to the fact that those who respond to the survey in the treatment group are no longer comparable to respondents from the control group, as the treatment affects response behavior. In our example the treatment has induced some additional individuals to respond: let's call them the "marginal respondents". The latent variable response model implies that individuals can be ranked according to their reluctance to respond (V), and this ranking is not modified by assignment to treatment. It is then possible to truncate the treatment sample by removing those marginal respondents. In proposition 1, this is done by manipulating W. Proposition 2 states that this is not necessary and we can truncate the sample simply knowing who, among the treated, responded within w_1 tries.

To understand this latter aspect, we need to note that, by definition of N and given the latent variable response model given by equation (3):

$$V < p(w, z) \Leftrightarrow (N \leq w, Z = z). \tag{13}$$

This is proved formally in appendix A.1.2, but it is natural that, when variable W is a maximum number of contacts, the response model means that a person in treatment group z who has V such that $V < p(w, z)$ will be reached at most at the wth contact. This is equivalent to saying that her realized N will be at most w. For this reason, N is sufficient to characterize individuals in terms of their latent V and we do not need exogenous variation in W for that. Respondents in the control group are such that $V < p(w_{\max}, 0)$. We need to identify respondents in the treatment group such that $V < p(w_{\max}, 0) = p(w_1, 1)$. Equivalence (13) states that they are fully characterized by $N \leq w_1$.

Figure 1 illustrates this process. Individuals are ranked according to their unobserved reluctance to respond V, and treatment does not affect the ranking, so that below any level of the latent reluctance to respond, people with the same characteristics are present in the control and the treatment group. Without actual instruments, the information provided by the number of calls before the person responded acts as a proxy for V and makes it possible to identify the marginal respondents. They can therefore be removed from the treatment–control comparison, thus restoring identification. Identification is

only "local", however, in the sense that it is only valid for a sub population of respondents (the respondents in the group with the lowest response rate or any subgroup who have responded after fewer phone calls). In the figure, individuals have been called up to 20 times and the response rate is lower in the control group. In the treatment group, the same response rate as in the control group is obtained for individuals that have responded within 18 calls (thus in this exemple, w_1=18). People who responded at the 19th or 20th call have no counterpart in the control group, so they are removed: comparing outcomes in the remaining control and treatment samples maintains balanced groups in terms of response behavior, thus neutralizing the selectivity bias.

Proposition 2 seems to be widely applicable. Phone surveys can routinely keep track of the number of attempts that were made to call the person, so that the data requirement is limited. Assumption 1 is standard for a selectivity correction model and it will be discussed at length below. Assumption 3 has to be assumed for identification of any causal parameter. Assumption 4 is the one specific to our approach. But it is extremely reasonable when one considers a specific variable W, such as the maximum number of calls allowed, as long as this maximum number of calls is not known to the individuals in advance. Last, non-parametric estimation is very easy to carry, as it boils down to removing part of the sample.

Finally, notice that, when response rates are identical in the control and treatment groups, then under the latent variable response model, there is no selectivity bias, in the sense that $E(y(1) - y(0)|R = 1)$ is identified. The reason is that identical response rates imply $p(w_{\max}, 0) = p(w_{\max}, 1)$, which, under the response model assumed here, means that the distribution of V of respondents in the two groups are identical. It is easy to check that the estimator from proposition 2 is then identical to simple comparison of means among respondents.

2.4 Discreteness of the number of calls

Our main result, stated in proposition 2, holds when w_1 can be found. In practice, W and N are discrete. As a consequence, it is not always possible to identify the exact cut-off w_1 and drop corresponding marginal respondents. However, one can bound the average treatment effect. In the propositon below, we consider the case when the outcome variable Y is bounded. When it is not, we can use quantiles of Y to bound the parameter, as done in Lee (2009); this is explained at the end of this section.

Take the case $p(w_{\max}, 0) < p(w_{\max}, 1)$. When N is discrete, we can find w_1 such that

$\Pr(N \leq w_1|Z=1) > \Pr(N \leq w_{\max}|Z=0)$ (it could perfectly be $w_1 = w_{\max}$, therefore such w_1 exists). Then proposition 2 can be restated as follows :

Proposition 3 ***: Partial identification with information on the number of calls.*** *Assume Y is bounded with lower and upper bounds $(\underline{y}, \overline{y})$. Under assumptions 1, 3 and 4, and if $p(w_{max}, 0) < p(w_{max}, 1)$, $E(y(1) - y(0)|V < p(w_{max}, 0))$ is set-identified with lower and upper bounds $(\underline{\Delta}, \overline{\Delta})$ such that:*

$$\begin{aligned}
\underline{\Delta} &= \frac{\Pr(N \leq w_1|Z=1)}{\Pr(N \leq w_{max}|Z=0)}[E(y(1)|N \leq w_1, Z=1) - \overline{y}] + \overline{y} \\
&- E(y|N \leq w_{max}, Z=0) \\
\overline{\Delta} &= \frac{\Pr(N \leq w_1|Z=1)}{\Pr(N \leq w_{max}|Z=0)}[E(y(1)|N \leq w_1, Z=1) - \underline{y}] + \underline{y} \\
&- E(y|N \leq w_{max}, Z=0)
\end{aligned}$$

where w_1 is such that

$$\Pr(N \leq w_1|Z=1) > \Pr(N \leq w_{max}|Z=0).$$

If $p(w_{max}, 0) > p(w_{max}, 1)$, we can find accordingly w_0 such that $\Pr(N \leq w_0|Z=0) > \Pr(N \leq w_{max}|Z=1)$ and define bounds symmetrically.

This is proved formally in appendix A.1.3 and results useful for estimation and inference are in appendix A.2.2. This proposition uses the fact that the unknown parameter in the treatment group, $E(y|V < p(w_{\max}, 0), Z=1)$ is a weighted difference of $E(y|V < p(w_1, 1), Z=1)$ which is observed, and $E(y|p(w_{\max}, 0) \leq V < p(w_1, 1), Z=1)$ which is unknown, with weights that depend on the observed proportions $\Pr(N \leq w_1|Z=1)$ and $\Pr(N \leq w_{\max}|Z=0)$. The lower and upper bounds correspond to the maximum and minimum possible values of $E(y|p(w_{\max}, 0) \leq V < p(w_1, 1), Z=1)$: $\overline{y}$ and $\underline{y}$ respectively.

The width of the identified set is

$$\overline{\Delta} - \underline{\Delta} = \left(\frac{\Pr(N \leq w_1|Z=1)}{\Pr(N \leq w_{\max}|Z=0)} - 1\right)(\overline{y} - \underline{y}).$$

In order to minimize the width of the identification set, it is natural to choose w_1 such that $\Pr(N \leq w_1|Z=1)$ is closest to $\Pr(N \leq w_{\max}|Z=0)$. The more phone calls there are, the more likely it is to find a group of respondents among the treated which is close to that in the control group, and the smallest the identification set.

This process is illustrated in Figure 2.a. As compared to Figure 1, there is no longer a number of calls in the treatment group for which the response rate is exactly the same as in the control group. We would therefore set the number of calls, w_1=19, for which the response rate is higher but closest to that in the control group. We observe the outcome for the corresponding group of individuals ($N \leq 19$) in the treatment group, but we would need to know it for the population below the dotted line. If we knew the average outcome for the population between the dotted line and $N = 19$, we would infer the result. As we do not observe this average outcome, we must bound it. Obviously, the bounds will be smaller the smaller this area. This will more likely happen when there is a large number of attempts N. Intuitively, when the number of attempts is very large, we converge towards the case with point identification.

In the case when Y is not bounded, we can follow Lee (2009) and use quantiles of the distribution of Y. Specifically, define $\underline{w_1}$ such that $\Pr(N \leq \underline{w_1}|Z=1) = p(\underline{w_1},1) < \Pr(N \leq w_{\max}|Z=0) = p(w_{\max},0)$. This is a number of calls that generates a response rate among the treated that is below the response rate in the control group.[8] As explained, we need to bound $E(y|p(w_{\max},0) \leq V < p(w_1,1), Z=1)$. Call A the population such that $p(w_{\max},0) \leq V < p(w_1,1)$ in the treatment group. We can consider a larger set of individuals (call it B) such that $p(\underline{w_1},1) \leq V < p(w_1,1)$ also in the treatment group. Among the population B, the population A is in proportion $\alpha = [p(w_1,1)-p(w_{\max},0)]/[p(w_1,1)-p(\underline{w_1},1)]$. Call y_α the αth quantile of the distribution of Y in population B. The mean of Y in population A cannot be lower than $E(y|y \leq y_\alpha, Z=1)$. Symmetrically, it cannot be larger than $E(y|y \geq y_{1-\alpha}, Z=1)$. As a result, if Y is not bounded, bounds in proposition 3 still hold with $\overline{y} = E(y|y \geq y_{1-\alpha}, Z=1)$ and $\underline{y} = E(y|y \leq y_\alpha, Z=1)$, which can be computed from the data. Notice that $\underline{w_1}$ should be chosen so as to make $p(\underline{w_1},1)$ as high as possible.

Back to Figure 2.a, we would set $\underline{w_1} = 18$ and use the quantiles defined within the population between $N = 18$ and $N = 19$ in the treatment group (i.e. population B). Area A is between the dotted line and $N = 19$.

[8] As always, we consider the case where $p(w_{\max},0) < p(w_{\max},1)$.

2.5 Comparison with bounding approaches

It is useful to compare our approach to the alternative, increasingly influential approach to the sample selection problem: the construction of worst-case scenario bounds of the treatment effect. This comparison will shed light on the trade-off between releasing identifying assumptions and improving what can be identified.

The assumption-free approach proposed by Horowitz and Manski (1995, 1998 and 2000) requires both weaker hypotheses (response behavior does not need to be monotonic) and less information (the number of attempts before reaching individuals does not need to be observed).[9] It does however require the outcome of interest to be bounded; moreover, as illustrated by Lee (2009) or in the application below, it may generate very large bounds if response rates are not very high.

The approach suggested by Lee (2009) is much closer to our approach. It provides tight bounds on treatment effects under the assumption that selection into the sample is monotonic (in a less restrictive sense than above), i.e., considering response $R(Z)$ as a function of assignment to treatment, $R(1) \geq R(0)$ for all individuals (or the reverse). The bounds are given by proposition 1a in Lee (2009, p. 1083). The width of the identified set can be substantially smaller than in Horowitz and Manski (2000), as it depends on the difference in response rates between the control and the treatment group, rather than on their sum. As in this paper, point identification is achieved when response rates are balanced.

Let us compare Lee (2009) with our framework: (i) our approach requires observing the actual survey effort leading to response[10]; (ii) both approaches impose monotonicity conditions on potential response behavior, but in our approach, the monotonicity condition is stronger as it bears jointly on the impact of assignment to treatment and on the impact of survey effort. The counterpart is that in many cases (when the number of attempts to reach people is large enough) we should have closer bounds, because we can "cut" the sample into smaller pieces.

Concerning identification results, the two approaches lead to point identification when response rates are balanced. Actually, when response rates are balanced between treated and controls, the monotonicity assumption implies that respondents in the two groups

[9] See in particular Horowitz and Manski (2000). Assume that y is bounded: $-\infty < y_{min} \leq y \leq y_{max} < \infty$. In its simplest form, the approach is to consider two extreme cases. In the best case, the outcome of all non-respondents from the control group is y_{min} and the outcome of all treated non-respondents is y_{max}; vice-versa in the worst case. If non respondents are in proportion nr_0 (resp. nr_1) in the control (resp. treatment) group, then the width of the identified interval is $(nr_0 + nr_1)(y_{max} - y_{min})$.

[10] Generally, Lee's approach applies to any selection model, whereas this paper is only relevant to selection created by survey non-response.

represent the exact same population: there is no sample selection issue to start with. Notice that the usual "balanced response" argument only holds under the Lee (2009) response monotonicity hypothesis. In that sense, the numerous papers that check that response rates are balanced and then follow up with estimation implicitly make this hypothesis.

When response rates are not balanced, both approaches yield set identification (our approach can provide point identification only in cases that are quite unlikely in practice). Figure Figure 2.a (commented above) and Figure 2.b illustrate the difference. In the two cases, individuals are ranked according to their unobserved propensity to respond, and treatment does not affect the ranking, so that at a given level V corresponding to a given response rate, individuals in the control and the treatment groups are comparable. In Lee's approach, the mean outcome for all treated respondents is observed and, in order to bound the mean outcome for the population below the dotted line, bounds are derived for the share above the line: bounds for the parameter of interest are small when this share is relatively small. In our approach, knowledge of the number of attempts to reach a person, N allows us to tighten the bounds if it allows us to observe groups within which the marginal respondents are a smaller share. In the figure, individuals between $N = 19$ and the dotted line are less numerous than all individuals above the dotted line. Therefore, this approach should be more informative typically when there is a large number of attempts.

3 Application

In this section, we analyze non response in the context of a job search experiment (which actually initiated the research in this paper). We briefly present the program that is evaluated, the data, and evidence on sample selection bias. We then implement the sample selection correction proposed in this paper, and compare it to alternative (bounding) approaches. Our correction appears to reduce the sample selection bias, whereas the identified intervals from the bounding approaches are too wide to be conclusive here.

3.1 The program and the data

The phone survey used in this application took place in the context of a job search experiment (presented in more details in Behaghel, Crépon and Gurgand, 2012). In 2007-08, the French unemployment benefit provider (Unédic) mandated private companies to provide intensive counseling to job seekers. To be eligible, job seekers had to be entitled with unemployment benefits for at least 365 days. The program was implemented as

a randomized control trial: eligible job seekers were randomly assigned to the standard track (control group, with less intensive counseling) or to the treatment.[11]

There are 2 sources of information to measure the program impact : administrative data and a phone survey. From the administrative data (i.e. the unemployment registers), we use one key variable, *exit*, which indicates whether the job seekers exited registered unemployment before the phone survey (before March 2008). Because individuals are benefit recipients we can be quite confident that an exit from the registers, which imply the suspension of benefits, is meaningful and related to a transition to employment (this view can be challenged when unemployed are not eligible to benefits; see Card, Chetty and Weber, 2008). However, the administrative data does not allow to measure other relevant dimensions of impact, such as job quality on which the program put strong emphasis. To measure these dimensions, a phone survey was run in March and April 2008. The initial sample included around 800 job seekers out of the 4,300 individuals who had entered the experiment between April and June 2007 (see table 1). 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.[12] The interviews were conducted by a survey company. The questionnaire was rather 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 response rate to the phone survey is low: out of 798 individuals in initial the sample, only 57% responded (see Table 1). This motivates investigating the risk of sample selection bias. To do so, we use the exhaustive administrative data as a benchmark: are the results on *exit* the same if one considers the full sample as if one restricts the analysis to the sample of respondents to the phone survey? Table 2 shows OLS estimates of intention-to-treat effects in the two samples. Estimated effects are about 50% larger when the sample is restricted to the phone survey respondents (a 13.6 percentage point impact, compared to 9.6 percentage points in the full sample). This is suggestive of a quantitatively significantly bias. Note however that the difference could come either from treatment effect heterogeneity (component Δ_1 in equation 3) or from sample selection bias (component Δ_2). A caveat is that standard errors are large, so that the difference is not statistically significant and could simply be due to sampling variations.

[11]Participation to the intensive scheme was not compulsory, so that compliance was imperfect. For the sake of simplicity, we focus on the intention-to-treat effect. The generalization to the identification of local average treatment effects is available on demand.

[12]The analysis uses survey weights accordingly.

Table 3 provides evidence that sampling variations are not the only reason why the full sample and the sample of respondents yield different results: it shows that response behavior is statistically correlated with exit from unemployment registers. For instance, in the control group (line 1), there are sizable differences in exit rates between those who respond (column 2) and those who do not respond to the long telephone survey (column 1). The exit rate is 14.5 percentage points lower among respondents than among non respondents (column 4). As a consequence, considering the respondents only to estimate the exit on the whole population leads to a 7.2 percentage point downward bias (a 15% difference, see column 5). The fact that the phone survey over-represents job seekers with lower employment prospects can be interpreted in various ways: for instance, job seekers who have found a job are harder to reach, or they do not feel they have to respond to surveys related to the public employment service anymore. There is similar evidence of a downward bias for job seekers assigned to treatment.

Another important piece of evidence that non response is not "as good as random" in the phone survey is given by table 4, which shows that non response is correlated with treatment assignment: job seekers respond more when they are assigned to the private scheme than when assigned to the control group (the response rate increases by 13.7 percentage points, with a standard error of 4.0).

To sum up, the low and unbalanced response rates cast doubts on the validity of the phone survey in order to measure the program impact. These doubts are reinforced by the comparison with the exhaustive administrative data. In what follows, we implement different approaches to control for attrition bias in this data, and check whether these corrections close the gap between results with the full sample and results with respondents to the phone survey only. Recall that due to treatment effect heterogeneity it could be the case that the true effect on respondents (estimated with sample selection correction) actually widens this gap. Comparing estimates on the whole population and our corrected estimates cannot be a formal test for validity of our correction.

3.2 Selection correction

Table 5 displays estimates based on three correction approaches. In the first two columns, we recall the estimate on the whole population and the estimate on respondents. In columns 3 and 4, we report "bounding" estimates. The Horowitz and Manski (2010) bounds and the Lee (2009) bounds are large, so that the telephone survey brings limited information.

In the last column, we report estimates derived from our proposed correction method.[13] In order to implement the correction, we need to find the number of calls at which to truncate the treatment group and restore the balance with the control group. Figure 3 displays response rates according to the number of phone calls and the assignment status. To restore the initial balance between experimental groups, the sample needs to be truncated between 6 and 7 phone calls in the group assigned to treatment. Following proposition 3, we assume two polar situations for marginal workers who respond after 7 phone calls when treated and would not have responded had they be in the control group. If we assume that they are employed (unemployed), we obtain a lower (upper) bound of the treatment effect. Even though the resulting identified set is quite large (2.6 points), it turns out to be strictly between the effect on the whole population and the naive effect on the respondents. The correction thus tends to close the gap between average treatment effect on respondents and on the whole population. Taken at face value, this implies that treatment effect heterogeneity is less an issue than internal validity of the effect on respondents. More interestingly, the identified set is far smaller than usual bounding estimates (around 20 points).

4 Conclusion

In this paper, we argue against the view that finding plausible instruments is the key impediment to sample selection correction models in the line of the Heckman (1976, 1979) model. If that model is correct, basic information on the number of calls (or number of visits) that were performed before the individual responded is enough to obtain narrow bounds of treatment effect, even in a semi-parametric model with heterogeneous treatment effect and a flexible specification of the latent threshold-crossing selection equation. The somewhat counter-intuitive result is that, despite the fact that reluctance to respond may well be correlated with potential outcomes, the actual effort made to get a response contains the same information as if survey effort was randomly allocated to individuals.

If the instrument is not the issue, it does not mean that there is no issue with such sample selection correction models. The true cost, however, lies in the restrictions that the model implies on response behavior. Clearly, if bounding approaches yield sufficiently narrow identified sets, they should be preferred as they imply less stringent restrictions. However, Horowitz and Manski (2000) bounds are quite large when response rates are below 80%, which is by no way the exception in social sciences. And the assumptions made by Lee (2009) are not so different from ours: extending the monotonicity assumptions may

[13] Estimates and standard errors are derived in appendix.

not be such a large cost compared to the substantial gains in terms of identification in cases where response rates are unbalanced, as in our application or in Kremer, Miguel and Thornton (2009).

References

Angrist, J., Imbens, G. and Rubin, D. (1996), 'Identification of causal effects using instrumental variables', *Journal of the American Statistical Association* **91**, 444–455.

Behaghel, Crepon and Gurgand (2012), 'Private and public provision of counseling to jobseekers : Evidence from a large controlled experiment', *IZA Discussion Paper* (6518).

Behrman, J., P. S. and Todd, P. (2009), Medium term impacts of the oportunidades conditional cash transfer program on rural youth in mexico, *in* S. Klasen and F. Nowak-Lehmann, eds, 'Poverty, Inequality and Policy in Latin America', MIT Press.

Campbell, D. (1969), 'Reforms as experiments', *American Psychologist* **24**, 409–29.

Card, Chetty and Weber (2008), 'The spike at benefit exhaustion : Leaving the unemployment system or starting a new job ?', *American Economic Review* **97**(2), 113–118.

Das, M., Newey, W. K. and Vella, F. (2003), 'Nonparametric estimation of sample selection models', *Review of Economic Studies* **70**(1), 33–58.

Fitzgerald, J., Gottschalk, P. and Moffitt, R. (1998), 'An analysis of sample attrition in panel data: The michigan panel study of income dynamics', *Journal of Human Resources* **33**(2), 251–299.

Heckman, J. J. (1976), 'The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models', *The Annals of Economic and Social Measurement* **5**, 475–492.

Heckman, J. J. (1979), 'Sample selection bias as a specification error', *Econometrica* **47**(1), 153–61.

Horowitz, J. L. and Manski, C. F. (1995), 'Identification and robustness with contaminated and corrupted data', *Econometrica* **63**(2), 281–302.

Horowitz, J. L. and Manski, C. F. (1998), 'Censoring of outcomes and regressors due to survey nonresponse: Identification and estimation using weights and imputations', *Journal of Econometrics* **84**(1), 37–58.

Horowitz, J. and Manski, C. (2000), 'Nonparametric analysis of randomized experiments with missing covariate and outcome data', *Journal of the American Statistical Association* **95**(449), 77–84.

Kremer, M., Miguel, E. and Thornton, R. (2009), 'Incentives to learn', *The Review of Economics and Statistics* **91**(3), 437–456.

Krueger, A. B. (1999), 'Experimental estimates of education production functions', *The Quarterly Journal of Economics* **114**(2), 497–532.

LaLonde, R. J. (1986), 'Evaluating the econometric evaluations of training programs with experimental data', *American Economic Review* **76**(4), 604–20.

Lee, D. (2009), 'Training, wages, and sample selection: Estimating sharp bounds on treatment effects', *Review of Economic Studies* **76**, 1071–1102.

Manski, C. (1989), 'Anatomy of the selection problem', *Journal of Human Resources* **24**(3), 343–60.

Newey, W. K. and McFadden, D. (1986), Large sample estimation and hypothesis testing, *in* R. F. Engle and D. McFadden, eds, 'Handbook of Econometrics', Vol. 4 of *Handbook of Econometrics*, Elsevier, chapter 36, pp. 2111–2245.

Vytlacil, E. (2002), 'Independence, monotonicity, and latent index models: An equivalence result', *Econometrica* **70**(1), 331–341.

Figure 1: Identification in proposition 2

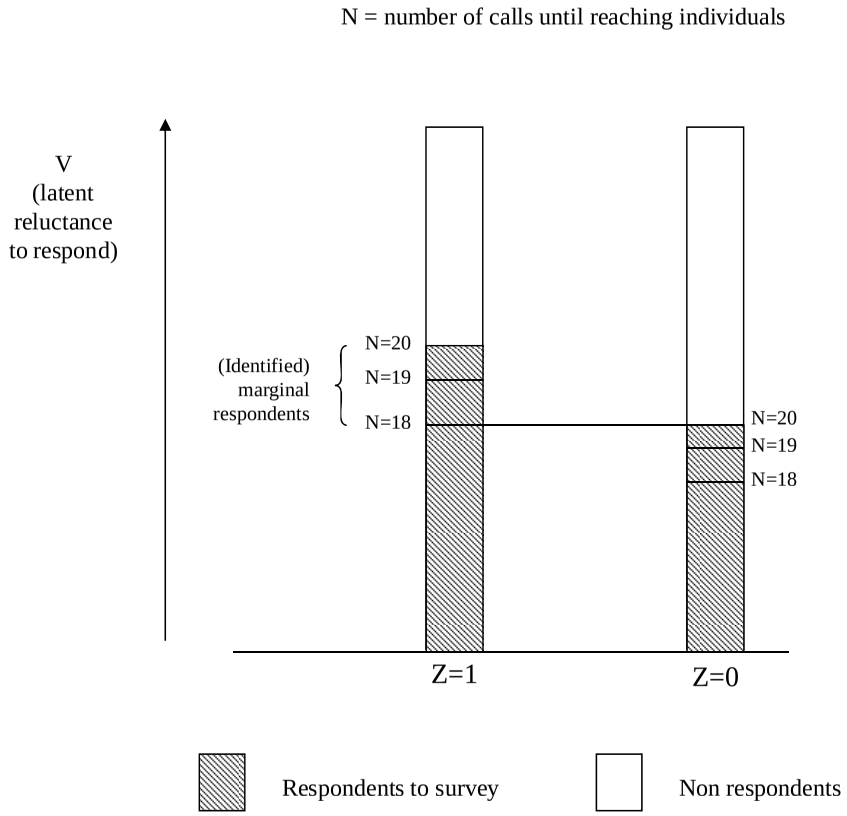

Figure 2: Identification under Lee (2009) bounds and proposition 3

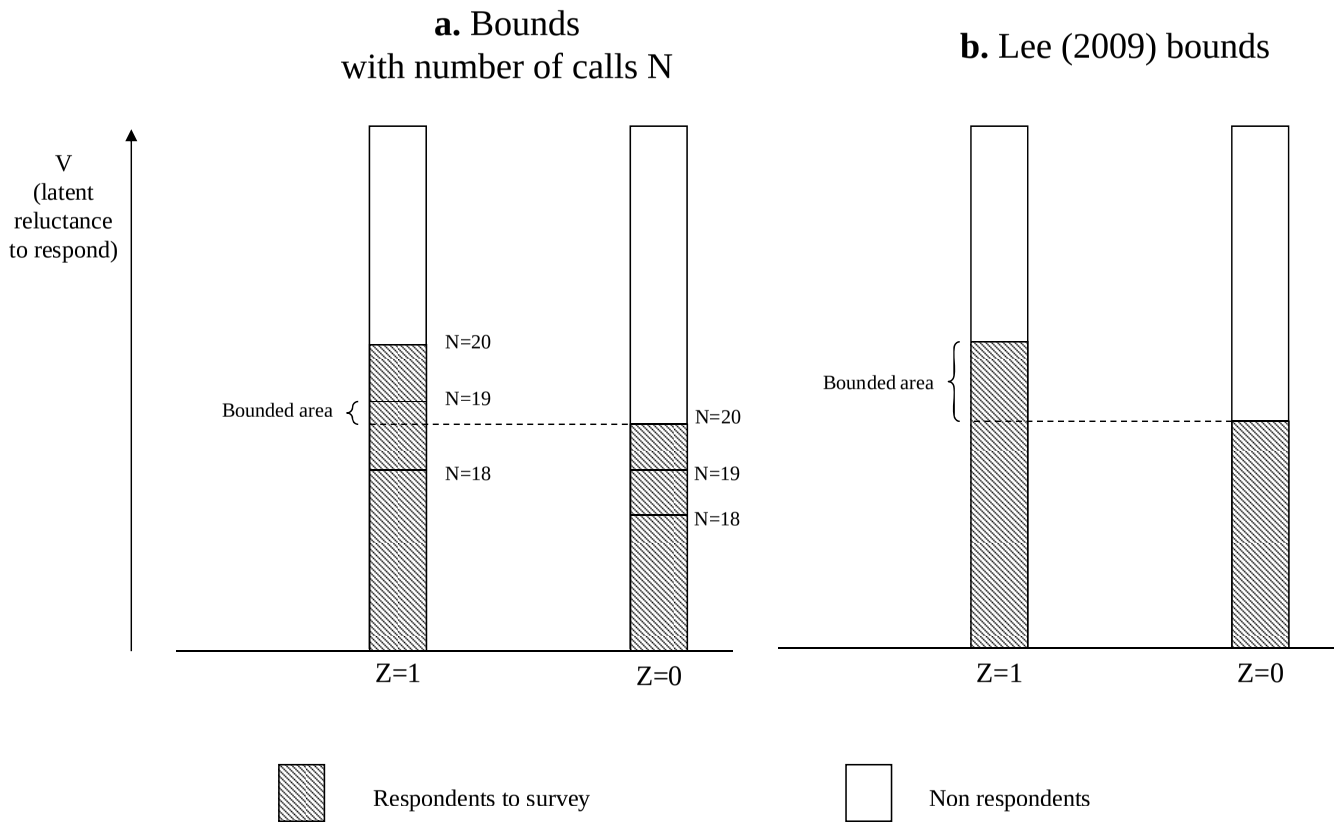

Figure 3: Response rates by assignment according to the number of phone calls

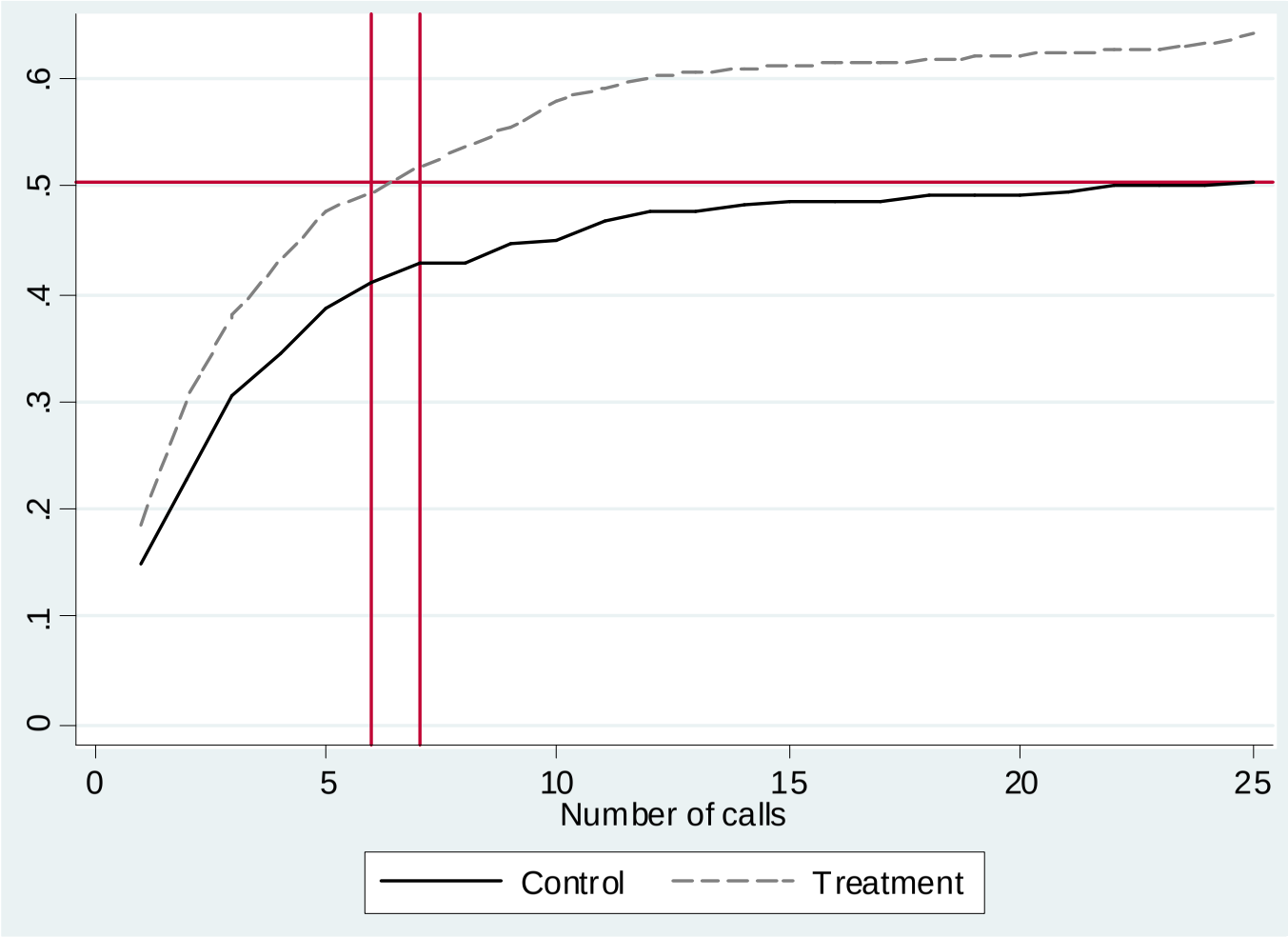

Table 1: Population sizes

	Full sample	Respondents to phone survey
Sample size	798	493
Initial population size	4324.82	2478.09
Weighted response rate	1.00	0.57

Note : Response rate is computed using the sampling weights of the telephone survey.

Table 2: Program impact on exit from the unemployment registers without correcting for sample selection

	Full sample	Respondents to phone survey
Treatment	0.096	0.136
	(0.040)	(0.054)
N	798	493

Note : Linear probability model. Observations weighted according to the sampling design of the telephone survey. Robust standard errors in parenthesis.

Table 3: Exit from the unemployment registers depending on the response status in the phone survey

	Non respondents (a)	Respondents (b)	All (c)	Difference (b)-(a)	Difference (c)-(a)	p value (c)-(a)
Control group	0.523	0.378	0.450	-0.145	-0.072	0.031
	(0.048)	(0.046)	(0.034)	(0.067)	(0.033)	.
Treatment group	0.602	0.515	0.546	-0.088	-0.031	0.063
	(0.038)	(0.028)	(0.023)	(0.047)	(0.017)	.

Note : Observations weighted according to the sampling design of the telephone survey. Standard errors are below the effects in parenthesis.

Table 4: Impact of assignment on response to phone survey

	Response
Treatment	0.137
	(0.040)
N	798

Note : Linear probability model. Observations weighted according to the sampling design of the telephone survey. Robust standard errors in parenthesis.

Table 5: Program impact on exit from unemployment records with corrections for sample selection

	Without correction		Horowitz Manski	Lee	Truncation
Sample	all	respondents			
	(1)	(2)	(3)	(4)	(5)
Treatment	0.096	0.136	[-0.356;0.498]	[0.003;0.217]	[0.101;.127]
	(0.040)	(0.054)	(0.037);(0.033)	(0.056);(0.049)	(0.061);(0.061)
Size	798	493	493	493	423

Note : Linear probability model, with telephone survey weights. Standard errors in parenthesis. Columns (1) and (2) recall table 2. Estimates with standard bounding techniques *à la* Horowitz and Manski and *à la* Lee are in columns (3) and (4). Our truncation procedure is in the last column.

A Appendix

A.1 Proofs of propositions in the text

A.1.1 Proof of proposition 1

Proof 1 *Under assumption 1,*

$$\begin{aligned} E(y|R=1,Z=0,W=w) &= E(y(0)|R=1,Z=0,W=w) \\ &= E(y(0)|\mathbf{1}(V\leq p(w,0))=1,Z=0,W=w) \\ &= \frac{E(y(0)\mathbf{1}(V\leq p(w,0))|Z=0,W=w)}{\Pr(\mathbf{1}(V\leq p(w,0))=1|Z=0,W=w)} \\ &= \frac{E(y(0)\mathbf{1}(V\leq p(w,0)))}{\Pr(\mathbf{1}(V\leq p(w,0))=1)} \\ &= E(y(0)|V\leq p(w,0)). \end{aligned}$$

Similarly,

$$E(y|R=1,Z=1,W=w) = E(y(1)|V\leq p(w,1)).$$

This holds for any couples $(w_0,0)$ *and* $(w_1,1)$*. Consequently, if there exists* w_0 *and* w_1 *such that* $p(w_0,0)=p(w_1,1)=\bar{p}$*, then we have:*

$$E(y(1)-y(0)|V\leq\bar{p})=E(y|R=1,W=w_1,Z=1)-E(y|R=1,W=w_0,Z=0),$$

which is equation 5 in the text.

A.1.2 Proof of proposition 2

Proof 2 *For any given* (w,z)*, denote (A):* $V<p(w,z)$ *and (B):* $(N\leq w,Z=z)$*. We start by proving that under assumptions 1 and 4, (A)* $\Leftrightarrow$ *(B) (equation 13).*

(A) $\Rightarrow$ *(B):* $V<p(w,z)$ *implies that the person responds when a maximum of* w *attempts are made. The number of attempts until the person is reached is therefore less or equal* w*.*

not (A) $\Rightarrow$ *not (B):* $V\geq p(w,z)$ *implies that the person does not respond when a maximum of* w *attempts are made. Given assumption 4, it also implies that the person does not respond when a maximum of 1, or 2,..., or* $w-1$ *attempts are made. This in*

turn implies that N cannot be equal to 1,2,...,w. Therefore, $N > w$, noting $N = \infty$ for individuals who never respond (whatever the number of attempts).

This equivalence result implies

$$\begin{aligned} p(w,z) &\equiv \Pr(R=1|W=w,Z=z) \\ &= \Pr(V<p(w,z)) \\ &= \Pr(N\leq w|Z=z). \end{aligned}$$

Take the case $p(w_{max},0) < p(w_{max},1)$. If there exists w_1 such that $p(w_{max},0) = p(w_1,1)$, then w_1 is such that

$$\Pr(N\leq w_1|Z=1) = \Pr(N\leq w_{max}|Z=0).$$

Moreover, using the equivalence result and assumption 3,

$$\begin{aligned} E(y|Z=z,N\leq w) &= E(y(z)|Z=z,N\leq w) \\ &= E(y(z)|Z=z,V<p(w,z)) \\ &= E(y(z)|Z=z,\mathbf{1}(V<p(w,z))=1) \\ &= \frac{E(y(z)\mathbf{1}(V\leq p(w,z))|Z=z)}{\Pr(\mathbf{1}(V\leq p(w,z))=1|Z=z)} \\ &= \frac{E(y(z)\mathbf{1}(V\leq p(w,z)))}{\Pr(\mathbf{1}(V\leq p(w,z))=1)} \\ &= E(y(z)|V\leq p(w,z)). \end{aligned}$$

This holds in particular for $(w,z) = (w_{max},0)$ and $(w_1,1)$. Therefore,

$$E(y(1)-y(0)|V<p(w_{max},0)) = E(y|N\leq w_1,Z=1) - E(y|N\leq w_{max},Z=0).$$

Because p is non-decreasing, $w_1 \leq w_{max}$ and the sample with $(N\leq w_1, Z=1)$ is observable.

A.1.3 Proof of proposition 3

Proof 3 *Take the case $p(w_{max},0) < p(w_{max},1)$. Because p is non-decreasing, $\exists w_1 \leq w_{max}$ such that $p(w_{max},0) < p(w_1,1)$. (If inequality is not strict, we are back to Proposition 2).*

Using $V < p(w,z) \Leftrightarrow (N\leq w, Z=z)$ (proved in A.1.2) and independance assumptions, we have:

$$
\begin{aligned}
& E(y(1)|N \leq w_1, Z = 1) \\
= \; & E(y(1)|V < p(w_1, 1)) \\
= \; & E(y(1)|V < p(w_{max}, 0)) \Pr(V < p(w_{max}, 0)|V < p(w_1, 1)) \\
+ \; & E(y(1)|p(w_{max}, 0) \leq V < p(w_1, 1)) \Pr(p(w_{max}, 0) \leq V < p(w_1, 1)|V < p(w_1, 1))
\end{aligned}
$$

where we have decomposed $(V < p(w_1, 1))$ *depending on whether* V *is lower or higher than* $p(w_{max}, 0)$.

Also:

$$
\begin{aligned}
\Pr(V < p(w_{max}, 0)|V < p(w_1, 1)) &= \frac{\Pr(N \leq w_{max}|Z = 0)}{\Pr(N \leq w_1|Z = 1)} \\
\Pr(p(w_{max}, 0) \leq V < p(w_1, 1)|V < p(w_1, 1)) &= \frac{\Pr(N \leq w_1|Z = 1) - \Pr(N \leq w_{max}|Z = 0)}{\Pr(N \leq w_1|Z = 1)}
\end{aligned}
$$

By manipulating the previous equations we get:

$$
\begin{aligned}
& E(y(1)|V < p(w_{max}, 0)) \\
= \; & \frac{\Pr(N \leq w_1|Z = 1)}{\Pr(N \leq w_{max}|Z = 0)} E(y(1)|N \leq w_1, Z = 1) \\
- \; & \frac{\Pr(N \leq w_1|Z = 1) - \Pr(N \leq w_{max}|Z = 0)}{\Pr(N \leq w_{max}|Z = 0)} E(y(1)|p(w_{max}, 0) \leq V < p(w_1, 1))
\end{aligned}
$$

In this expression, $E(y(1)|p(w_{max}, 0) \leq V < p(w_1, 1))$ *is not observed. If we set it to* $\underline{y}$ *or* $\overline{y}$ *(its bounds), we obtain the bounds in text for* $E(y(1)|V < p(w_{max}, 0)) - E(y(0)|V < p(w_{max}, 0))$.

A.2 Estimation and inference of the truncation model

A.2.1 Results for proposition 2

For notation convenience, denote Δ the treatment effect identified in proposition 2 ($\Delta = E(y(1) - y(0)|V \leq \bar{p})$). In addition to proposition 2 assumption, we assume $\bar{p}$ is actually attained in the control group. Estimation and inference results are inspired by Lee (2009).

First we define the estimates as sample analogs to the parameters defined in proposition 2.

$$\begin{pmatrix} \widehat{\Delta^R} = \frac{\sum YZ\mathbf{1}(N\leq \widehat{n_{\widehat{p}}})}{\sum Z\mathbf{1}(N\leq \widehat{n_{\widehat{p}}})} - \frac{\sum YR(1-Z)}{\sum R(1-Z)} \\ \widehat{n_{\widehat{p}}} = \min n : \frac{\sum Z\mathbf{1}(N\leq \widehat{n_{\widehat{p}}})}{\sum Z} \geq \widehat{p} \\ \widehat{p} = \frac{\sum R(1-Z)}{\sum 1-Z} \end{pmatrix}$$

Second we verify consistency by showing that the estimator solves a well defined GMM problem and applying theorem 2.6 of Newey and MacFadden (1986). To do so, we need one additional assumption, i.e. N has bounded support. It is sufficient to prove consistency of $\mu_0 = E(Y|Z=1, N \leq n_{\bar{p}})$ Denote $\theta_0' = (\mu_0, n_{\bar{p}_0}, \bar{p}_0)'$ the true value of the parameters vector and $d' = (Y, Z, N)'$ the data. Define the moment function $g(d, \theta)$:

$$g(d,\theta) = \begin{pmatrix} (Y-\mu)Z\mathbf{1}(N \leq n_{\bar{p}}) \\ (\mathbf{1}(N \leq n_{\bar{p}}) - \bar{p})Z \\ (\mathbf{1}(N \leq w) - \bar{p})(1-Z) \end{pmatrix}$$

Recall that w is the maximum number of attempts, such that $R = \mathbf{1}(N \leq w)$. The estimator μ_0 is the solution to $min_\theta(\sum g(d,\theta))'(\sum g(d,\theta))$.

Third we verify asymptotic normality by applying theorem 7.2 of Newey and MacFadden (1986). We define $g_0(\theta) = E(g(d,\theta))$ and $\widehat{g_n}(\theta) = n^{-1}\sum g(d,\theta)$. We also define G the derivative of $g_0(\theta)$ at $\theta = \theta_0$. We can verify the assumptions of theorem 7.2 [14] and obtain that the asymptotic variance is $V = G^{-1}\Sigma(G^{-1})'$ where Σ is the asymptotic variance of $\widehat{g_n}(\theta)$. Σ is equal to:

$$\Sigma = \begin{pmatrix} E((Y-\mu_0)^2\mathbf{1}(N \leq n_{\bar{p}_0})|Z=1)E(Z) & 0 & 0 \\ 0 & \bar{p_0}(1-\bar{p_0})E(Z) & 0 \\ 0 & 0 & \bar{p_0}(1-\bar{p_0})E(1-Z) \end{pmatrix}$$

Define $f(.)$ as the density of N conditional on $Z=1$. Then G is equal to:

$$G = \begin{pmatrix} -\bar{p_0}E(Z) & Mf(n_{\bar{p}_0})E(Z) & 0 \\ 0 & f(n_{\bar{p}_0})E(Z) & -E(Z) \\ 0 & 0 & -E(1-Z) \end{pmatrix}$$

[14] the only difficulty is stochastic equicontinuity

where $M = E(Y - \mu_0 | Z = 1, N = n_{\bar{p_0}})$. Its inverse G^{-1} is :

$$
\begin{aligned}
G^{-1} &= \frac{1}{\bar{p_0} f(n_{\bar{p_0}})(E(Z))^2 E(1-Z)} \times \\
&\begin{pmatrix} -f(n_{\bar{p_0}})E(Z)E(1-Z) & Mf(n_{\bar{p_0}})E(Z)E(1-Z) & -Mf(n_{\bar{p_0}})(E(Z))^2 \\ 0 & \bar{p_0}E(Z)E(1-Z) & -\bar{p_0}(E(Z))^2 \\ 0 & 0 & -\bar{p_0} f(n_{\bar{p_0}})(E(Z))^2 \end{pmatrix}
\end{aligned}
$$

Hence the upper left term of the variance matrix is the sum of three terms :

$$
\begin{aligned}
V(1,1) &= \frac{Var(Y | Z = 1, N \leq n_{\bar{p_0}})}{\bar{p_0}E(Z)} + \\
&\frac{(1-\bar{p_0})(E(Y - \mu_0 | Z = 1, N = n_{\bar{p_0}}))^2}{\bar{p_0}E(Z)} + \\
&\frac{(E(Y - \mu_0 | Z = 1, N = n_{\bar{p_0}}))^2}{E(1-Z)}
\end{aligned}
$$

The first term (V^Y) is the usual variance of the mean estimator when there is no uncertainty concerning the trimming procedure. The second term (V^N) reflects the fact that once the fraction to be trimmed is known there is still uncertainty about the right number of calls under which the sample should be trimmed. The third term (V^P) is the part of the variance of the estimator due to uncertainty about the true fraction to be trimmed.

To sum up, we have shown that $\sqrt{(}m) \left(\widehat{\Delta} - \Delta \right) \rightarrow N(0, V^Y + V^N + V^P + V^C)$ in distribution, where m is the total sample size, V^Y, V^N, V^P are defined just above and V^C is the variance of the conditional mean in the control group: $V^C = \frac{Var(Y|Z=0, N \leq w)}{E(1-Z)p_0}$.

A.2.2 Results for proposition 3

When N is discrete, we prove the convergence and derive the asymptotic properties of the bounds of the identified set. We focus on the lower bound $\underline{\Delta}$. The proof is similar for the upper bound. The estimator of the lower bound, $\widehat{\underline{\Delta}}$, is the sample analogs of $\underline{\Delta}$. The convergence and precision of the second term in $\underline{\Delta}$, the conditional mean of Y for the respondents in the control group is the same as in the previous section. Let's focus on the first term. The estimator can be written :

$$\frac{\widehat{\tilde{p}}}{\widehat{\bar{p}}}[\hat{\mu} - \overline{y}] + \overline{y}$$

where $\tilde{p} = \Pr(N \leq n_{\bar{p}} | Z = 1)$, so that $\widehat{\tilde{p}} = \frac{\sum \mathbf{1}(N \leq \widehat{n_{\bar{p}}}) RZ}{\sum Z}$. $\widehat{\bar{p}}$ and $\hat{\mu}$ have already been defined above.

First, the convergence and precision of $\hat{\mu}$, can be proven using the same theorem of Newey and McFadden (1986), as their assumptions are still verified (the parameter set is compact and the moment function is continuous almost everywhere on this compact). Note that the cutoff of the number of calls in the treatment group, denoted w_1 in proposition 3, corresponds in fact to $n_{\bar{p}}$ in the previous subsection; we use both notations indifferently below. Because of N discreteness, the estimator of $n_{\bar{p}}$ converges faster than $\sqrt{m}$. As a consequence, it can be omitted in the computation of the asymptotic variance:

$$\sqrt{m}\,(\widehat{\mu} - \mu) \to N(0, V^Y)$$

where to simplify notation, we abstract from subindex 0 which indicates true parameter in the previous subsection ($\mu = \mu_0$).

Second, we apply the delta method to the estimator of $\underline{\Delta}$. In addition to the asymptotic variance of $\widehat{\mu}$, we need to use the convergence properties of $\widehat{\bar{p}}$ and $\widehat{\tilde{p}}$:

$$\sqrt{m}\left(\widehat{\bar{p}} - \bar{p}\right) \quad \to \quad N(0, \frac{\bar{p}(1-\bar{p})}{E(1-Z)}) \tag{14}$$

$$\sqrt{m}\left(\widehat{\tilde{p}} - \tilde{p}\right) \quad \to \quad N(0, \frac{\tilde{p}(1-\tilde{p})}{E(Z)}) \tag{15}$$

Then the asymptotic variance of $\widehat{\underline{\Delta}}$ is:

$$\left(\frac{\tilde{p}}{\bar{p}}\right)^2 V_Y + \left(\frac{\tilde{p}(\mu - \bar{y})}{\bar{p}^2}\right)^2 \frac{\bar{p}(1-\bar{p})}{E(1-Z)} + \left(\frac{\mu - \bar{y}}{\bar{p}}\right)^2 \frac{\tilde{p}(1-\tilde{p})}{E(Z)} + V_C$$

A.3 Extension to non compliance

In this appendix, we extend the results of proposition 2 to the case where compliance is imperfect. 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 5

1. *SUTVA*
2. *(Monotonicity):*
$$T(1) \geq T(0)$$
3. *(Exclusion)*
$$y(T) = y(T(Z)) \equiv \tilde{y}(Z)$$
4. *(Independence)*
$$Z \perp y(1), y(0), T(1), T(0)$$

(Note that we changed notation for the sake of readability: $y(0)$ and $y(1)$ now denote potential outcome under the different treatment statuses; $\tilde{y}(0)$ and $\tilde{y}(1)$ correspond to potential outcomes under the different assignment statuses that were noted $y(0)$ and $y(1)$ above.) It is well known that under this set of assumptions, the usual Wald estimator identifies the local average treatment effect on compliers (LATE):

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

We now consider non response. We extend assumption **??** to account for imperfect compliance.

Assumption 6

1. *(Latent variable threshold-crossing response model):*
$$R = \mathbf{1}(V < \tilde{p}(W,Z)),$$
2. *(Independence):*
$$\begin{aligned} W, Z &\perp \tilde{y}(0), \tilde{y}(1), \tilde{N}(0), \tilde{N}(1), T(0), T(1), V \\ Z &\perp W \end{aligned}$$

Proposition 4 ***Identification with the actual number of calls leading to response under imperfect compliance.*** *Under assumption 5 and 6 and a binary outcome,* $E(y(1)-y(0)|V\leq\bar{p},T(1)-T(0)=1)$ *is identified from the observation of* y, T, Z *and* N*:*

$$E(y(1)-y(0)|V\leq\bar{p},T(1)-T(0)=1)=\frac{E(y|N\leq w_1,Z=1)-E(y|N\leq w_0,Z=0)}{E(T|N\leq w_1,Z=1)-E(T|N\leq w_0,Z=0)},$$

with w_0,w_1 *such that*

$$\Pr(N\leq w_1|Z=1)=\Pr(N\leq w_0|Z=0)=\bar{p}.$$

Proof 4 *Under assumption 6, Proposition 2 applies:*

$$\begin{aligned}E(\tilde{y}(1)-\tilde{y}(0)|V\leq\bar{p}) &= E(y|N\leq w_1,Z=1)-E(y|n\leq w_0,Z=0) &(16)\\ E(T(1)-T(0)|V\leq\bar{p}) &= E(T|N\leq w_1,Z=1)-E(T|n\leq w_0,Z=0) &(17)\end{aligned}$$

(note that $y=Z\tilde{y}(1)+(1-Z)\tilde{y}(0)$ *and* $N=Z\tilde{N}(1)+(1-Z)\tilde{N}(0)$*.)*

By the law of iterated expectations,

$$\begin{aligned}E(\tilde{y}(1)-\tilde{y}(0)|V\leq\bar{p}) = {} & E(y(1)-y(0)|V\leq\bar{p},T(1)-T(0)=1)\times\Pr(T(1)-T(0)=1|V\leq\bar{p})\\ & +0\times\Pr(T(1)-T(0)=0|V\leq\bar{p})\\ & +E(y(1)-y(0)|V\leq\bar{p},T(1)-T(0)=-1)\times\Pr(T(1)-T(0)=-1|V\leq\bar{p})\end{aligned}$$

In the absence of defiers ($T(1)\geq T(0)$*), the last term is 0. Therefore*

$$E(y(1)-y(0)|V\leq\bar{p},T(1)-T(0)=1) = \frac{E(\tilde{y}(1)-\tilde{y}(0)|V\leq\bar{p})}{E(T(1)-T(0)|V\leq\bar{p})} \qquad (18)$$

Combining equations 16, 17, and 18 yields the result.

Proposition 4 implies that $E(y(1)-y(0)|V\leq\bar{p},T(1)-T(0)=1)$ can be estimated using the standard Wald estimator, after truncating the sample following the order of the number of phone calls needed, up to the point where the same share of the initial population is represented in the treatment and in the control group. Once this truncation is done, control and treatment respondents are statistically identical, and the standard IV argument applies.