Please call again: correcting non-response bias in treatment effect models [*]

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

September 3, 2014

Abstract

We propose a novel selectivity correction procedure to deal with survey attrition in treatment effect models, at the crossroads of the "Heckit" model and 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 prior calls made to each individual before obtaining a response to the survey. We obtain sharp bounds to the average treatment effect on the common support of responding individuals. Because the number of prior calls brings information, we can obtain tighter bounds than in other non-parametric methods.

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

JEL : C31, C93, J6

[*] We thank Laurent Davezies, Xavier D'Haultfoeuille, Esther Duflo, Martin Huber, David McKenzie, as well as seminar participants at CREST, IFAU and the RES conference 2013 for their comments on previous versions of this paper.

[†] Paris School of Economics - Inra. Corresponding author. Address: 48 bd Jourdan, 75014 PARIS, FRANCE. Tel +33 1 43 13 63 66; fax +33 1 43 13 63 62. luc.behaghel@ens.fr

[‡] Crest

[§] Paris School of Economics - CNRS

[¶] Crest

1 Introduction

Sample attrition is a pervasive problem for surveys in the social sciences. The damage appears particularly clearly in randomized trials. Random assignment to treatment creates a treatment group and a control group that are at the same time comparable and representative of some 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 this 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 becoming 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

[1] For instance, the education scientist McCall, writing in the early 1920s before R.A. Fisher's reference book *The Design of Experiments* (1935) was published, observed: "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).

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).

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" model and 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 made to each individual before obtaining a response. The method can be applied whenever data collection entails sequential efforts to obtain a 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 the randomization of 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 heterogeneous 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 (sharp) bounds are identified, but this is in fact a common feature of selectivity models when there is no instrument or when instru-

[2] Evidence is obtained by comparing survey data with another, more comprehensive data source (e.g., administrative data, or reports from parents or neighbors of non-respondents). See for instance Behrman et al. (2009) or the application in Behaghel et al. (2012).

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

ments take a limited number of values. Because we bring information into the model, the identified set can be smaller than in Lee (2009).

The intuition of our 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. Treatment may affect the overall willingness of people to participate in a survey, but this model assumes that it does not change their relative rank in V within the treatment or control groups. As an example, assume that the response rate is lower in the control group, implying that the treatment has induced some more "reluctant" (high V) individuals to respond. As a result, respondent individuals in the treatment and control groups have different distributions of V, so that estimates of the treatment effect may be biased if V is correlated with counterfactual outcomes.

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 people are ranked according to V in the single-index model, if 60% of the respondents 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 must have the 40% highest values of V). If we have two groups (treatment and control) in which the answering behavior differs, it remains true that, if 60% of the respondents in the treatment group answered before the 18th call, and 60% of the respondents in the control group answered before the 20th call, then each of these subsamples contains the 60% lowest values of V in their group. When the groups are randomized (or can be considered initially similar for any other reason), the 60% lowest V's in each group represent the same population. The insight of this paper is that when survey attempts are recorded, these 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 the response rates in the two groups. Implementation then follows Lee (2009) and uses quantiles of the distribution of the outcome for marginal respondents to build sharp bounds.

This paper contributes to a wide econometric literature on sample selection reaching back forty years, overviewed 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 entails using an instrument. The fact that an instrument is needed in order 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 one would obtain 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 for practitioners. In this respect, it is related to proposals by Lee (2009) and DiNardo et al. (2006). The comparison with Lee (2009) is detailed in the paper. As noted above, Lee's procedure is beginning to be used in applied studies but may lead to wide identified sets. DiNardo et al. (2006) propose to include the 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 having 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. In fact, the approach proposed here, where the effort is maximal for every individual, is superior to the randomization of survey effort, because such randomization does not maximize the number of respondents.

In the next section, we first introduce the framework and notations. We then recall and extend existing results on selection correction with instruments. Section 4 presents our

[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.

approach in the case when survey effort can be thought as continuous. This case conveys the main intuition for our identification result. Section 5 considers the more realistic case when survey effort is discrete. Finally, we compare our approach to existing bounding approaches. An application of the method is provided in the working document version of this paper Behaghel et al. (2012).

2 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 $E(y|Z=0)$.

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

When there is non-response, the observed mean value of y in the treatment group and in the control group measures $E(y|R=1, Z=1)$ and $E(y|R=1, Z=0)$ respectively.

[5] We assume perfect compliance: treatment is equal to assignment (equivalently, if there is imperfect compliance, we consider the intention-to-treat effect).

Therefore, the "naive" average treatment effect estimator 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 $y(0)$. Neither 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 concerns the lack of external validity. In the absence of 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.[6] In contrast, the problem raised by bias Δ_2 is one of internal validity. Even if our interest lies in $E(y(1)-y(0)|R(1)=1)$, this would not be estimated consistently if the second type of bias is present.

In the following, we restrict the parameter of interest to the average treatment effect

[6] An externally valid estimator can be recovered, however, using observed covariates X that are available both for respondents and non-respondents (for instance from administrative data) and assuming some form of homogeneity of treatment impact conditional on X, as in Huber (2012, 2013) or Angrist and Fernandez-Val (2013).

on respondents, and we consider hypotheses under which the selection bias (Δ_2) can be corrected, if present.

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 uses the latent selection model, in which 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 this model but argue that proper instruments are difficult to find. We then show that, provided the survey records the number of calls made to each individual before they respond, identification can be obtained using 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)*

$$R = \mathbf{1}(V < p(W, Z)), \tag{2}$$

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

Equation (2) 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. It can also be considered as the maximum number of calls that can be made in the attempt to survey each individual, which will prove useful in the following sections. 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's (2002) equivalence result, the key embedded assumption is a form of monotonicity: different individuals must not react in opposite ways to a given change in the pair (W, Z).[7]

3 Selectivity correction with instruments

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

Assumption 2

$$(W, Z) \quad \perp \quad y(0), y(1), V. \tag{3}$$

We then obtain the following identification result:

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

$$E(y(1) - y(0)|V < \bar{p}) = E(y|R = 1, W = w_1, Z = 1) - E(y|R = 1, W = w_0, Z = 0). \tag{4}$$

[7] 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 $R_i(w, z) \leq R_i(w', z') \; \forall \; i$ (where $R(w, z)$ is the response function resulting from equation 2). This monotonicity assumption is violated if assignment to treatment Z encourages some individuals to respond to the survey, but discourages others. Also, the condition does not hold if some individuals are only sensitive to W, and others 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 only if treated: $R_{i_1}(2, 0) < R_{i_1}(1, 1)$. 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 monotonic impacts, but no single latent variable threshold-crossing response model exists.

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. Newey (2007) and Huber (2012, 2013) provide similar identification results for the non-separable case in the presence of a continuous instrument W. The proof is given in appendix A.1.1. To interpret equation (4), it is useful to think of V as an individual reluctance to respond to surveys. As W and Z are randomly assigned, V is equally distributed across treatment and control groups, and across different values of the instrument W.[8] Given the response model in (2), 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 simply different ways of isolating the same subpopulation of respondents. Therefore, comparing y across these two subpopulations (treatment and control) 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 restrictions, notably treatment effect homogeneity.

Of course, equation (4) is only useful to the extent that there exists such an instrument W that varies sufficiently to entail a non-empty common support $\mathbf{P}$. Unless this is planned in advance, it is usually extremely difficult to extract variables from the data that have credible exogeneity properties and sufficient power to influence response. Therefore, as suggested above, randomizing the survey effort could be a natural way to generate the instrument W. We could, for instance, randomly assign the number of attempts to call

[8] Notice that, as long as (W, Z) is independent from V and counterfactuals $y(0)$ and $y(1)$, W could be randomized conditional on Z, for instance if one wants to put more survey effort in the control group.

each individual,[9] or the value of the gift offered to those who respond. However, randomizing the data collection effort means wasting part of the sample (that on which data collection effort is not maximal). In most contexts, survey costs are 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.

4 Number of given calls as a substitute for instruments

We now consider the case where W is a maximum number of phone calls, 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 made before a person is actually reached. The main insight of this paper is that W can be set to the same value for all individuals in the sample: provided that 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.

When W is set to $W = w_{\max}$ for all individuals in the sample, assumption 2 still holds formally, but in a degenerate sense for W. It does imply $Z \perp y(0), y(1), V$, conditional on $W = w_{\max}$.

We also add some structure to the function $p(.,.)$:[10]

Assumption 3

$$p(W, Z) \textit{ is non-decreasing in } W, \quad \forall Z. \tag{5}$$

[9] Specifically, we 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, worker j will be called a maximum number of times W_j, and so on.

[10] Notice that assumption 3 is not implied by the latent variable model (equation 2), as the function p was so far unspecified.

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 of contacting a person one more time should not decrease the likelihood of that person eventually answering the survey. This is particularly reasonable if subjects are not aware of the planned maximum number of attempts.

Without loss of generality, let us consider the case where the share of respondents is higher among the treated, i.e., $p(w_{\max}, 0) \leq p(w_{\max}, 1)$. Assume there exists w_1 such that: $p(w_{\max}, 0) = p(w_1, 1)$. w_1 is the 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 3, $w_1 \leq w_{\max}$. 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), the existence of such a w_1 in practice is most unlikely. Here, however, we assume its existence for clarity: to present the identification result in a simple case and give the basic intuition. In practice, discreteness will entail 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 using information on the number of calls.*** *Under assumptions 1, 2 and 3, 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:*

$$E(y(1) - y(0)|V < p(w_{max}, 0)) = E(y|N \leq w_1, Z = 1) - E(y|N \leq w_{max}, Z = 0), \tag{6}$$

where w_1 is identified from $\Pr(N \leq w_1|Z = 1) = \Pr(N \leq w_{max}|Z = 0)$.[11]

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

$$E(y(1) - y(0)|V < p(w_{max}, 1)) = E(y|N \leq w_{max}, Z = 1) - E(y|N \leq w_0, Z = 0), \tag{7}$$

[11] Note that the set of individuals with $N \leq w_1$ is observable because $w_1 \leq w_{\max}$.

where w_0 is identified from $\Pr(N \leq w_{max}|Z = 1) = \Pr(N \leq w_0|Z = 0)$.

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

To understand this identification result, take the case $p(w_{\max}, 0) \leq p(w_{\max}, 1)$. Equation (6) 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 $E(y(0)|V < p(w_{\max}, 0))$. 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 us 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 by knowing who, among the treated, responded within w_1 calls.

To understand this latter aspect, note that, by definition of N and given the latent variable response model given by equation (2):

$$V < p(w, z) \Leftrightarrow (N \leq w, Z = z). \tag{8}$$

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 call. This is equivalent to saying that his 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. 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 (8) 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 both the control group 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 enables us 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 subpopulation of respondents (the respondents in the group with the lowest response rate or any subgroup who have responded after fewer calls, i.e. individuals with rather low levels of reluctance to respond). In Figure 1, 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 example, 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 appears to be widely applicable. Phone surveys can routinely keep track of the number of attempts that were made to call each person, so that the data requirement is limited. Assumption 1 is standard for a selectivity correction model and is discussed at length in Vytlacil (2002). Assumption 2 is standard for the identification of any causal parameter (as it boils down to $Z \perp y(0), y(1), V$ in our setup). Assumption 3 is the one that is specific to our approach. But it is extremely reasonable when one considers a specific scalar 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 implement, as it amounts to removing part of the sample.

Finally, when response rates are identical in the control and treatment groups, then there is no selectivity bias under the latent variable response model, 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 is identical. In that case, the sample analog to $E(y|N \leq w_1, Z = 1) - E(y|N \leq w_{\max}, Z = 0)$ in equation (6) is the simple comparison of means among respondents.

5 Discreteness of the number of calls

Our main result, stated in proposition 2, holds when an adequate w_1 can be found. In the case of a number of calls, W and N are discrete: $w = 1, 2, ..., w_{\max}$. As a consequence, it is not always possible to identify the exact cut-off w_1 and remove the corresponding marginal respondents. However, we can bound the average treatment effect: proposition 3 below will show that we can define sharp bounds using quantiles of Y, as proposed by Lee (2009).

Take the case $p(w_{\max}, 0) \leq p(w_{\max}, 1)$ and find w_1 such that $\Pr(N \leq w_1|Z = 1) \geq \Pr(N \leq w_{\max}|Z = 0)$ (it could perfectly well be $w_1 = w_{\max}$, therefore such a w_1 exists). If we take w_1 such that $\Pr(N \leq w_1|Z = 1)$ is closest to $\Pr(N \leq w_{\max}|Z = 0)$, then $\Pr(N \leq w_1 - 1|Z = 1) < \Pr(N \leq w_{\max}|Z = 0)$: i.e. $w_1 - 1$ calls generate a response rate in the treatment group that is below the response rate in the control group. Overall, we have $p(w_1 - 1, 1) < p(w_{\max}, 0) \leq p(w_1, 1)$.

We use 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 $\Pr(N \leq w_{\max}|Z=0)$.[12] In this problem, $E(y|p(w_{\max},0) \leq V < p(w_1,1), Z=1)$ is unknown, but if it can be bounded; then $E(y|V < p(w_{\max},0), Z=1)$ is also bounded and so is the average treatment effect.

When y is continuous, sharp bounds can be derived following Lee (2009) and using quantiles of the distribution of Y. More precisely, 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 $p(w_1-1,1) \leq V < p(w_1,1)$, also in the treatment group. Among the population B, the population A is present in the proportion $\alpha = [p(w_1,1)-p(w_{\max},0)]/[p(w_1,1)-p(w_1-1,1)]$. 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 $E(y|y \leq y_\alpha, Z=1, N=w_1)$. Symmetrically, it cannot be larger than $E(y|y \geq y_{1-\alpha}, Z=1, N=w_1)$. As a result, $E(y|p(w_{\max},0) \leq V < p(w_1,1), Z=1)$ is bounded by $E(y|y \leq y_\alpha, Z=1, N=w_1)$ and $E(y|y \geq y_{1-\alpha}, Z=1, N=w_1)$, which can be computed from the data. This is formally stated in the following proposition:

Proposition 3 ***: Partial identification using information on the number of calls.*** *Assume assumptions 1, 2 and 3 hold, y is continuous, and* $p(w_{max},0) \leq p(w_{max},1)$. *There exists* w_1 *such that:*

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

[12] In proposition 2, the latter term, $E(y|p(w_{\max},0) \leq V < p(w_1,1), Z=1)$, had a weight of zero.

Define:

$$\begin{aligned}
\overline{y} &= E(y|y \geq y_{1-\alpha}, Z=1, N=w_1) \\
\underline{y} &= E(y|y \leq y_{\alpha}, Z=1, N=w_1) \\
\textit{with } y_q &= G^{-1}(q) \textit{ where } G \textit{ is the c.d.f of } Y \textit{ conditional on } N=w_1 \textit{ and } Z=1 \\
\textit{and } \alpha &= \frac{\Pr(N \leq w_1|Z=1) - \Pr(N \leq w_{max}|Z=0)}{\Pr(N \leq w_1|Z=1) - \Pr(N \leq w_1 - 1|Z=1)} /
\end{aligned}$$

Then $E(y(1) - y(0)|V < p(w_{max}, 0))$ *is set-identified with sharp lower and upper bounds* $(\underline{\Delta}, \overline{\Delta})$ *such that:*

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

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

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

This is proved formally in appendix A.1.3, and results useful for estimation and inference are given in appendix A.2. Estimation includes several steps. First, identify the number of calls w_1 in the treated population ($Z=1$) that makes the response rate in that population closest (but higher) to the response rate among all controls ($Z=0$). Compute all respondent shares $\Pr(N \leq w_1|Z=1)$, $\Pr(N \leq w_1 - 1|Z=1)$, and $\Pr(N \leq w_{\max}|Z=0)$. Then isolate the treated population that answered the w_1th call but didn't answer the (w_1-1)th call and compute the relevant α-based quantiles of Y in this population to form $\overline{y}$ and $\underline{y}$. Finally, use the average outcomes in the proper populations of respondents ($N \leq w_1$ if $Z=1$ and $N \leq w_{\max}$ if $Z=0$) to compute the bounds on treatment effect.

Note that when y is discrete, the quantiles y_α and $y_{1-\alpha}$ are not well-defined, so that Lee bounds are not immediately applicable. However, bounds for $E(y|p(w_{\max}, 0) \leq V <$

$p(w_1, 1), Z = 1)$ can be easily derived when y is bounded. In particular, consider the case where y is binary ($y \in \{0; 1\}$). Define $\gamma = \Pr(y = 1|Z = 1, N = w_1)$. Then an upper bound for $E(y|p(w_{\max}, 0) \leq V < p(w_1, 1), Z = 1)$ is $\overline{y} = \min(1, \frac{\gamma}{\alpha})$, and a lower bound is $\underline{y} = \max(0, \frac{\alpha+\gamma-1}{\alpha})$. The rest of proposition 3 holds. Estimation and inference results for the discrete case are also given in appendix A.2.

Identification is illustrated in Figure 2.a. As compared with 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 therefore set the number of calls, w_1=19, for which the response rate is higher but as close as possible 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 need to know it for the population below the dotted line. If we knew the average outcome for the population above the dotted line (with $N = 19$), we could infer the result. As we do not know this average outcome, we must bound it. In order to do so, we can consider the population with $N = 19$ (population B): α is the share of that population that is above the dotted line (population A). As explained above, the α and $(1 - \alpha)$ quantiles of the outcome in population B provide bounds to the average outcome in population A.

The width of the identified set is:

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

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

An alternative approach to cope with the discreteness of the survey effort is to make extra assumptions on functional forms of the model.[13] With no loss of generality, there are functions g_0 and g_1 such that:

$$E(y(z)|V < p(w, z)) = g_Z(p(w, z)), \qquad \forall z, \forall w = 1, 2, ..., w_{\max}.$$

[13] This alternative approach was suggested by a referee.

According to Proposition 2, values of $p(w,z)$ are directly given for any observation using that $p(w,z) = \Pr(N \leq w, Z = z)$, and no instrument is required. However, as the number of calls or visits is discrete, the response probability $p(w,z)$ also takes discrete values. Therefore, functions g cannot be estimated and the model is partially identified, unless functions g have less parameters than there are values of $p(w,z)$. Naturally, if one is ready to make such functional form hypothesis, one can estimate a general version of the usual selection model:

$$E(y|R = 1, N \leq w, z) = z[g_1(p(w,z)) - g_0(p(w,z))] + g_0(p(w,z)).$$

Then, with g_0 and g_1 in hand, one can estimate the average treatment effect on the common support of respondents.[14] Recalling that $\bar{p} = \min(p(w_{\max}, 0), p(w_{\max}, 1))$, we obtain:

$$E(y(1) - y(0)|V \leq \bar{p}) = g_1(\bar{p}) - g_0(\bar{p}).$$

6 Comparison with other bounding approaches

It is useful to compare our approach with 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 relaxing 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

[14] The familiar Das, Newey and Vella (2003) model applies when $g_0 - g_1 = \beta$, a constant impact. Under the parametric model, inclusion of control variables is straightforward. In contrast, inclusion of control variables in our bound estimation is difficult because, as those variables would generally be correlated with V, a different parameter would be bounded for different values of the control variables.

less information (the number of attempts made before reaching individuals does not need to be observed).[15] It does, however, require the outcome of interest to be bounded; moreover, as illustrated by Lee (2009), it may generate very large bounds if response rates are not very high.

The approach proposed 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 treatment groups, rather than on their sum.

Let us compare Lee (2009) with our framework. (i) Our approach requires observation of the actual survey effort leading to the response.[16] (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, we should have closer bounds, because we can further trim the sample making the difference in response rates look smaller (see Figure 2).

Concerning identification results, the two approaches lead to point identification when response rates are balanced. Actually, when response rates are balanced between treat-

[15] 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}, and vice versa in the worst case. If non-respondents are present in the proportion nr_0 (resp. nr_1) in the control (resp. treatment) group, then the width of the identified interval is $(nr_0 + nr_1)(y_{max} - y_{min})$.

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

ment and control groups, the monotonicity assumption entails that respondents in the two groups represent the exact same population: there is no sample selection issue to start with.

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 2.a (commented above) and Figure 2.b illustrate the difference. In both cases, individuals are ranked according to their unobserved propensity to respond, and treatment does not affect the ranking, so that at a given level of V corresponding to a given response rate, individuals in the control and 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 made 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 Figure 2.a, individuals above the dotted line with $N = 19$ 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.

7 Conclusion

In this paper, we argue against the view that finding plausible instruments is the key impediment to sample selection correction along the lines of the Heckman (1976, 1979) model. Under the hypothesis of that model, and in the context of survey attrition, basic information on the number of prior calls made to each individual before they responded is enough to obtain narrow bounds of treatment effect, even in a non-parametric model with heterogeneous treatment effects and a general 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 between individuals.

The fact that most of the time we can only expect to identify bounds is not specific to this approach. Any semi-parametric version of the familiar latent threshold-crossing selectivity model with an actual instrument would behave similarly when the instrument is not continuous: this applies in particular when survey effort is randomized in order to correct the selectivity bias. Point estimation only obtains under parametric assumptions. Furthermore, randomizing the survey effort does not maximize the number of respondents. Therefore, it is always optimal to apply the maximal effort to every individual, record the number of prior calls and use the correction method in this paper.

If the instrument is not the issue, this does not mean that there is no problem with that class of sample selection correction models. The true cost lies in the restrictions that the model implies on response behavior. Clearly, if bounding approaches yield sufficiently narrow identified sets, they should be preferred because they involve less stringent restrictions. However, Horowitz and Manski's (2000) bounds are quite wide when response rates are below 80%, which is by no means the exception in the social sciences. And the assumptions made by Lee (2009) are not so different from ours: extending the monotonicity assumptions is not such a large cost compared with the substantial gains in terms of identification in cases where response rates are unbalanced.

References

Angrist, J. and Fernandez-Val, I. (2013), Extrapolate-ing: External validity and overidentification in the late framework, *in* 'Advances in Economics and Econometrics: Theory and Applications', Vol. 3 of *Econometric Society Monographs*, Tenth World Congress.

Behaghel, L., Crépon, B., Gurgand, M. and Le Barbanchon, T. (2012), 'Please call again: Correcting non-response bias in treatment effect models', *IZA Discussion Paper* **6751**.

Behrman, J., Parker, 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.

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

DiNardo, J., McCrary, J. and Sanbonmatsu, L. (2006), 'Constructive proposals for dealing with attrition', *Mimeo* .

Hausman, J. and Wise, D. (1979), 'Attrition bias in experimental and panel data: The gary income maintenance experiment', *Econometrica* **47**(2), 455–73.

Heckman, 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.

Huber, M. (2012), 'Identification of average treatment effects in social experiments under alternative forms of attrition', *Journal of Educational and Behavioral Statistics* **37**(3), 443–474.

Huber, M. (2013), 'Treatment evaluation in the presence of sample selection', *Econometric Reviews* **forthcoming**.

Imbens, G. and Manski, C. (2004), 'Confidence intervals for partially identified parameters', *Econometrica* **72**(6), 1845–1858.

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

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. (2007), 'Nonparametric continous/discrete choice models', *International Economic Review* **48**(4), 1429–1439.

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

N = number of calls until reaching individuals

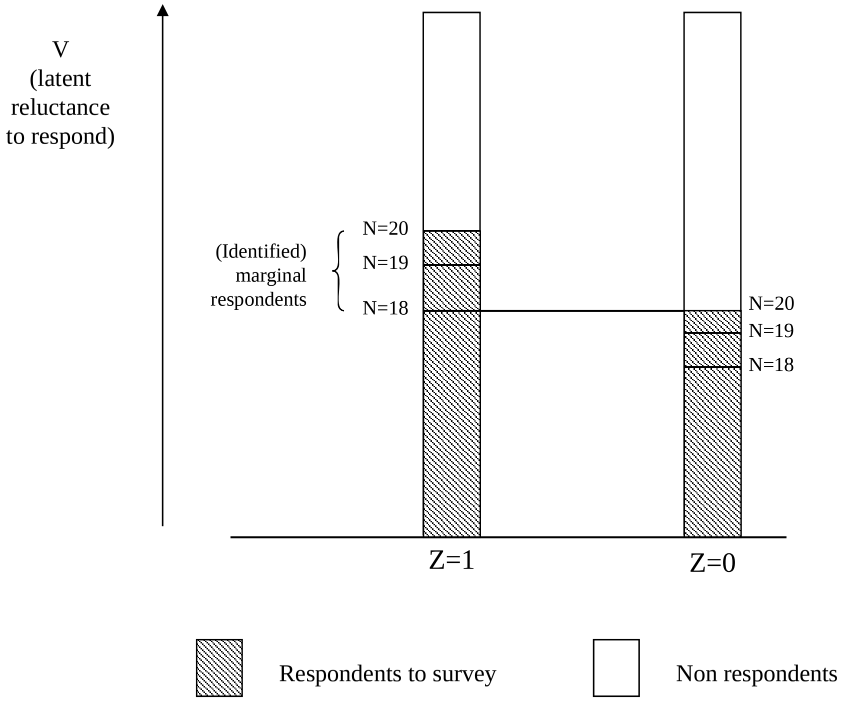
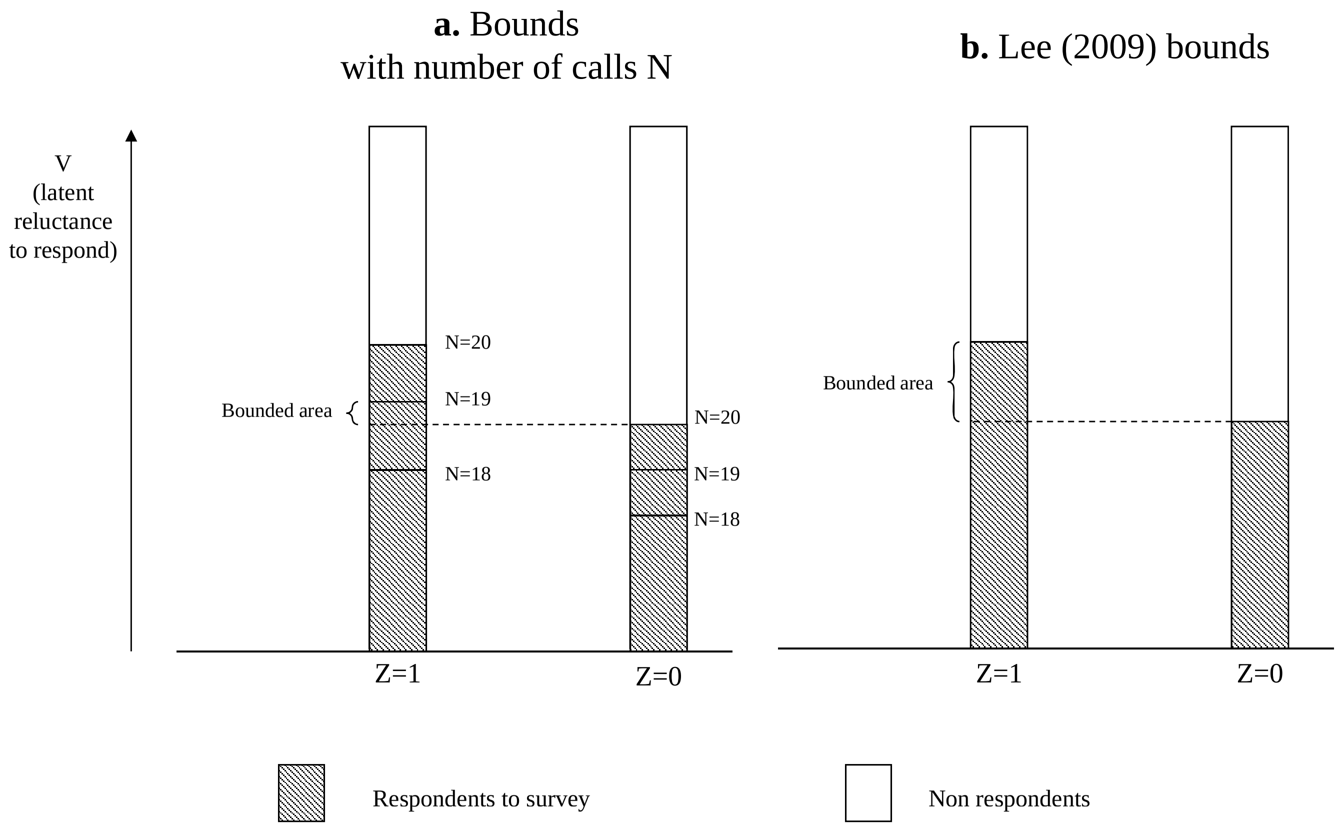


Figure 2: Identification under proposition 3 and under Lee's (2009) bounds



a. Bounds with number of calls N

b. Lee (2009) bounds

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 exist* 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 4 in the text.

A.1.2 Proof of proposition 2

Proof 2 *For any given* (w,z)*, let us denote (A):* $V < p(w,z)$ *and (B):* $(N \leq w, Z=z)$*. We start by proving that under assumptions 1 and 3, (A)* $\Leftrightarrow$ *(B) (equation 8).*

(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 3, 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 that:

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

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 2:

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

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). In the case were* $w_1 = 1$, *results hold with* $\Pr(N \leq w_1 - 1|Z = 1) = 0$.

Using $V < p(w, z) \Leftrightarrow (N \leq w, Z = z)$ *(proved in A.1.2) and independence 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:

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

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 given in the text for* $E(y(1)|V < p(w_{max}, 0)) - E(y(0)|V < p(w_{max}, 0))$*. Sharpness derives directly from the proof in Lee (2009).*

A.2 Estimation and inference of the truncation model

In this appendix, we define the estimators of the bounds of the identified set. We prove their consistency and derive their asymptotic properties. We focus on the lower bound $\underline{\Delta}$. The proof is similar for the upper bound. $\underline{\Delta}$ is defined in proposition 3 as:

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

Its estimator, $\widehat{\underline{\Delta}}$, is the sample analog of $\underline{\Delta}$. Let us review its different components.

The estimator of the conditional mean of Y for the respondents in the control group and its properties are standard. Let us define $\hat{\nu}$ this estimator. It writes: $\hat{\nu} = \frac{\sum YR(1-Z)}{\sum R(1-Z)}$. It converges to the true value $\nu = E(Y|Z=0, R=1) = E(y|N \leq w_{\max}, Z=0)$. Let us call m the number of observations in the sample. We have: $\sqrt{m}\,(\widehat{\nu} - \nu) \to N(0, V^C)$ where $V^C = \frac{Var(Y|Z=0, N \leq w_{max})}{E(1-Z)\bar{p}}$.

$\underline{\Delta}$ comprises two proportions: $\bar{p} = \Pr(N \leq w_{\max}|Z=0)$ and $\tilde{p} = \Pr(N \leq w_1|Z=1)$. Their sample analogs are respectively: $\widehat{\bar{p}} = \frac{\sum R(1-Z)}{\sum 1-Z}$ and $\widehat{\tilde{p}} = \frac{\sum \mathbf{1}(N \leq \widehat{w_1})RZ}{\sum Z}$. Note that $\widehat{\tilde{p}}$ depends on $\widehat{w_1}$, which is also estimated: $\widehat{w_1} = \min n : \frac{\sum Z\mathbf{1}(N \leq n)}{\sum Z} \geq \widehat{\bar{p}}$. Because of N discreteness, the estimator of w_1 converges faster than $\sqrt{m}$. As a consequence, it can be omitted in the computation of the asymptotic variance of $\tilde{p}$. Finally, we have the usual results: $\sqrt{m}\left(\widehat{\bar{p}} - \bar{p}\right) \rightarrow N\left(0, V^{\bar{p}}\right)$ and $\sqrt{m}\left(\widehat{\tilde{p}} - \tilde{p}\right) \rightarrow N\left(0, V^{\tilde{p}}\right)$ where $V^{\bar{p}} = \frac{\bar{p}(1-\bar{p})}{E(1-Z)}$ and $V^{\tilde{p}} = \frac{\tilde{p}(1-\tilde{p})}{E(Z)}$

Let us define $\widehat{\mu}$ the estimator of $\mu = E(y|N \leq w_1, Z = 1)$. It writes: $\widehat{\mu} = \frac{\sum YZ\mathbf{1}(N \leq \widehat{w_1})}{\sum Z\mathbf{1}(N \leq \widehat{w_1})}$. As above, we can abstract from the uncertainty associated to the estimation of w_1 ($\hat{w}_1$ converges faster than $\sqrt{m}$). We have the usual result: $\sqrt{m}\,(\widehat{\mu} - \mu) \rightarrow N(0, V^Y)$ with $V^Y = \frac{Var(Y|Z=1, N \leq w_1)}{\bar{p}E(Z)}$

The last term that needs to be estimated is $\overline{y} = E(y|y \geq y_{1-\alpha}, Z = 1, N = w_1)$. Let us define $\widehat{\overline{y}}$ as its estimator. When Y is continuous, it can be characterized as in Lee (2009).

$$\begin{pmatrix} \widehat{\overline{y}} = \frac{\sum Y\mathbf{1}(Y \geq \widehat{y}_{1-\widehat{\alpha}})Z\mathbf{1}(N = \widehat{w_1})}{\sum \mathbf{1}(Y \geq \widehat{y}_{1-\widehat{\alpha}})Z\mathbf{1}(N = \widehat{w_1})} \\ \widehat{y}_{1-\widehat{\alpha}} = \min y : \frac{\sum \mathbf{1}(Y \geq \widehat{y}_{1-\widehat{\alpha}})Z\mathbf{1}(N = \widehat{w_1})}{\sum Z\mathbf{1}(N = \widehat{w_1})} \geq 1 - \widehat{\alpha} \\ \widehat{\alpha} = \frac{\widehat{\tilde{p}} - \widehat{\bar{p}}}{\widehat{\tilde{p}} - \widehat{\tilde{q}}} \\ \widehat{\tilde{q}} = \frac{\sum Z\mathbf{1}(N \leq \widehat{w_1} - 1)}{\sum Z} \end{pmatrix}$$

The proof of its consistency and the derivation of its asymptotic properties also follow Lee (2009), which uses theorems 2.6 and 7.2 of Newey and MacFadden (1986). Applying directly Lee's results, we have: $\sqrt{m}\left(\widehat{\overline{y}} - \overline{y}\right) \rightarrow N(0, V^{\overline{y}})$ where

$$\begin{aligned} V^{\overline{y}} &= \frac{Var(Y|Z = 1, N = w_1, Y \geq y_{1-\alpha})}{E(Z1(N = w_1))\alpha} + \frac{(y_{1-\alpha} - \overline{y})^2(1 - \alpha)}{E(Z1(N = w_1))\alpha} + \left(\frac{y_{1-\alpha} - \overline{y}}{\alpha}\right)^2 V^{\alpha} \\ V^{\alpha} &= \frac{1}{(\tilde{p} - \tilde{q})^2}\left(V^{\tilde{p}} + V^{\bar{p}}\right) + \left(\frac{\tilde{p} - \bar{p}}{(\tilde{p} - \tilde{q})^2}\right)^2 \left(V^{\tilde{p}} + V^{\tilde{q}}\right) \end{aligned}$$

To sum up, the estimator of $\underline{\widehat{\Delta}}$ writes: $\underline{\widehat{\Delta}} = \frac{\widehat{\tilde{p}}}{\widehat{\bar{p}}}[\hat{\mu} - \widehat{\overline{y}}] + \widehat{\overline{y}} - \hat{\nu}$. We can then apply the delta method. Thus the asymptotic variance of $\underline{\widehat{\Delta}}$ is:

$$\left(\frac{\tilde{p}}{\bar{p}}\right)^2 (V^Y + V^{\overline{y}}) + \left(\frac{\tilde{p}(\mu - \overline{y})}{\bar{p}^2}\right)^2 \frac{\bar{p}(1 - \bar{p})}{E(1 - Z)} + \left(\frac{\mu - \overline{y}}{\bar{p}}\right)^2 \frac{\tilde{p}(1 - \tilde{p})}{E(Z)} + V^{\overline{y}} + V^C$$

Following Imbens and Manski (2004), a 95% confidence interval for $E(y(1) - y(0)|V <$

$p(w_{\max}, 0))$ given those bounds is: $[\widehat{\underline{\Delta}} - C_m.\widehat{\sigma_{\underline{\Delta}}}/\sqrt{m}, \widehat{\overline{\Delta}} + C_m.\widehat{\sigma_{\overline{\Delta}}}/\sqrt{m}]$ where C_m satisfies: $\Phi(C_m + \sqrt{m}\frac{\widehat{\overline{\Delta}} - \widehat{\underline{\Delta}}}{\max(\widehat{\sigma_{\overline{\Delta}}}, \widehat{\sigma_{\underline{\Delta}}})}) - \Phi(-C_m) = 0.95$

When Y is binary, we can apply the previous formula. The only term that needs to be updated is the estimator of $\bar{y}$. It writes: $\widehat{\overline{y}} = \min(1, \frac{\widehat{\gamma}}{\widehat{\alpha}})$ where $\widehat{\alpha}$ is defined above and $\widehat{\gamma} = \frac{\sum YZ1(N=\widehat{w_1})}{\sum Z1(N=\widehat{w_1})}$. Let us assume that the econometrician knows how the minimum is attained. When $\overline{y} = 1$, the asymptotic variance $V^{\bar{y}}$ can be dropped. When $\overline{y} = \frac{\gamma}{\alpha}$, we have: $V^{\bar{y}} = \frac{Var(Y|Z=1,N=w_1)}{E(Z1(N=w_1))\alpha^2} + \left(\frac{\gamma}{\alpha^2}\right)^2 V^{\alpha}$.