ANOTHER LOOK AT THE REGRESSION DISCONTINUITY DESIGN

Erich Battistin [1], Enrico Rettore [2]

ABSTRACT

The attractiveness of the Regression Discontinuity Design (RDD) for the evaluation of social policies rests on its close similarity to a formal experimental design. On the other hand, it is of limited applicability since seldom units are assigned to the treatment group on the basis of pre-program characteristics observable to the analyst. Besides, it only allows identification of the mean impact of the program on a very specific sub-population of individuals. In this paper we show that the RDD straightforwardly generalizes to the instances in which individuals' eligibility is established with respect to an observable pre-program measure and eligible individuals self-select into the program. We derive the regularity conditions necessary for the identification of treatment effects and we show how this set-up turns out very convenient to build a specification test on conventional non-experimental estimators for the mean impact of the program. Data requirements are made explicit.

KEY WORDS: Program Evaluation; Second Control Group; Specification Tests.
JEL Classification: C4, C8.

Preliminary and incomplete. First draft 12th February 2002, this version 17th January 2003. This paper benefited from useful discussion with David Card, Hide Ichimura and Andrea Ichino and from comments by audiences at ESEM 2002, CEPR/IZA workshop "Improving Labour Market Performance" in Bonn, October 2002, Statistics Canada Symposium 2002 and LABORatorio conference "New perspectives in public policy evaluation" in Turin, November 2002.

1. INTRODUCTION

The central issue in the evaluation of public policies is to separate their causal effect from the confounding effect of other factors influencing the outcome of interest. Random assignment of units to the intervention produces treatment and control groups that are equivalent in all respects, except for their exposition status. Thus, in a completely randomized experiment any post-intervention difference between the two groups doesn't reflect pre-intervention differences by construction. As a result, differences between exposed and control units are entirely due to the intervention itself.

However, in most instances randomization is unfeasible either for ethical reasons or simply because assignment to the treatment can't be controlled by the analyst. Besides, even in those instances in which the analyst can randomize the assignment, units may not comply with the assigned status and either drop out of the intervention or seek an alternative program (see Heckman and Smith, 1995). A well-known and widely used example of randomized assignment is the JTPA program in the United States, which currently serves close to one million economically disadvantaged people every year (see Friedlander *et al.*, 1997). Random assignment occurs prior to the actual enrolment in the program, but a consistent fraction of those randomized into the treatment group don't participate. For certain components of the JTPA, such a non-complying behaviour seems to be non-negligible (see, for example, Heckman *et al.*, 1998b).

In this situation, the ideal experiment is not fully realized since participation turns out (at least partly) voluntary: training is provided only for those individuals who meet certain criteria of need and comply with the result of randomization. It follows that participation depends on observable and unobservable characteristics of individuals

[1]Institute for Fiscal Studies, 7 Ridgmount Street, London WC1E 7AE, UK, erich_b@ifs.org.uk
[2]Department of Statistics, University of Padova, Via Cesare Battisti 241, 35121 Padova, Italy, rettore@stat.unipd.it

that might be correlated with the outcome of interest. In this situation, differences between treated and control groups with respect to the outcome of interest might be the result of units' self-selection into the intervention.

The assessment of whether observed changes in the outcome of interest could be attributed to the intervention itself and not to other possible causes turns out to be even more complicated in a non-experimental setting. In this situation estimating cause-effect relationships that one might think to hold between the program and its outcomes typically depends on not testable assumptions about individuals' behaviour. Given that the ideal situation for policy evaluators is the complete knowledge of the mechanism leading individuals to participate into the treatment, and given that in most instances such a mechanism is unknown (either because of non-compliance of individuals in an experimental setting or because of the lack of knowledge arising from observational studies), the question then arises of how to make the most out of each design to obtain reasonable estimates of program effects.

There are instances in which the so called Regression Discontinuity Design (RDD) arises (see Campbell, 1964, Rubin, 1977, Trochim, 1984). According to this design, assignment is solely based on pre-intervention variables observable to the analyst and the probability of participation changes discontinuously as a function of these variables. To fix ideas, consider the case in which a pool of units willing to participate are split into two groups according to whether a pre-intervention measure is above or below a known threshold. Those who score above the threshold are exposed to the intervention while those who score below are denied it.

This design features both advantages and disadvantages relative to its competitors. On the one hand, in a neighbourhood of the threshold for selection a RDD presents some features of a pure experiment. In this sense, it is certainly more attractive than a non-experimental design. Since subjects in the treatment and control group solely differ with respect to the variable on which the assignment to the intervention is established (and with respect to any other variable correlated to it), one can control for the confounding factors contrasting marginal participants to marginal non-participants. In this context, the term *marginal* refers to those units *not too far* from the threshold for selection. The comparison of mean outcomes for marginally treated and marginally control units identifies the mean impact of the intervention locally with respect to the threshold for selection. Intuitively, for identification at the cut-off point to hold it must be the case that any discontinuity in the relationship between outcomes of interest and the variable determining the treatment status is fully attributable to the treatment itself (this requires some regularity conditions at the threshold for selection discussed by Hahn *et al*., 2001).

On the other hand, the RDD features two main limitations. Firstly, its feasibility is by definition confined to those instances in which selection takes place on observable pre-intervention variables; as a matter of fact, this is not often the case. Secondly, even when such a design applies, it only permits to identify the mean impact of the intervention at the threshold for selection. In the realistic situation of heterogeneous impacts across units, this *local* effect might be very different from the effect for units away from the threshold for selection. To identify the mean impact on a broader population one can only resort to a non-experimental estimator whose consistency for the intended impact intrinsically depends on behavioural (and not testable) assumptions.

In this paper we show that the range of applicability of the RDD is wider than it has been thought before. It includes all those instances in which the relevant population is split into two subpopulations, eligibles and non-eligibles, provided that (i) the eligibility status is established with respect to a continuous variable and (ii) information on both non-eligible and eligible non-participant units is available. Then, the mean impact on participant units in a neighbourhood of the threshold for eligibility is identified, no matter how eligible units self-select into the program. We make an explicit connection to the literature on RDDs (only implicit references so far; Angrist, 1998; Van der Klaauw, 2002) and we derive the regularity conditions necessary for identification to hold. We also show that the discontinuity in the eligibility rule leads to regularity conditions for identification weaker than those arising in the standard RDD design (Hahn *et al*., 2001).

Secondly, as a straightforward corollary of the previous result, the selection bias at the threshold for eligibility turns out identifiable. Then, one can formally test whether any of the long array of existing non-experimental estimators is able to correct for such selection bias. On finding an estimator able to compensate for the selection bias even if only with reference to a particular subpopulation - namely, the units in a neighbourhood of the threshold for eligibility - one might feel more confident to use it on the broader population.

Links to related literature are established. In particular, we show that our first result is closely linked to Bloom (1984) and to Angrist and Imbens (1991). We also stress that our result is closely related to the idea stated by Rosenbaum (1987) of using two alternative comparison groups to better identify a program impact. Lastly, we point out the similarities between our specification test of a non-experimental estimator and the specification tests derived by Heckman and Hotz (1989) as well as the link to the characterization of the selection bias provided by Heckman *et al*. (1998a).

The remaining of this paper is organized as follows. Section 2 discusses similarities between a fully randomized experiment and a RDD. Section 3 generalizes the use of a RDD when participation into the treatment group is determined by self-selection. Section 4 shows how to validate the use of non-experimental estimators for the treatment effect using a RDD. Section 5 presents some concluding remarks.

2. REGRESSION DISCONTINUITY DESIGN AND RANDOMIZED EXPERIMENTS

This section highlights similarities between a fully randomized experiment and a RDD. Following the notation of the potential outcome approach to causal inference (see Rubin, 1974), let (Y^T,Y^{NT}) be the potential outcomes as a results of participating and not participating in the program, respectively. The causal effect of the treatment is then defined as the difference between these two outcomes, $?=Y^T-Y^{NT}$, which is not observable since being exposed to (denied) the program reveals Y^T (Y^{NT}) but conceals the other potential outcome.

Let D be the binary variable for the treatment status, with D=1 for participants and D=0 for non-participants. If the assignment is determined by randomization, the treatment status doesn't depend on individuals' characteristics and the following condition holds true

$$(Y^T ?Y^{NT}) ? D? \tag{1}$$

Typically randomization is administered only to people who previously applied for a certain program, who in general are not representative of the overall population (as for the JTPA case discussed above). In this situation condition (1) holds with respect to the group of units actually randomized, not with respect to the overall population.

The attractiveness of randomization is that the difference between the mean outcome for treated units and the mean outcome for control units identifies the mean impact of the program

$$E(Y^T ? Y^{NT}) ? E(Y^T ?D ? 1) ? E(Y^{NT} ?D ? 0)? \tag{2}$$

since conditioning on D in the right-hand side of (2) is irrelevant by construction. In other words, randomization allows using information on non-participants to identify the mean counterfactual outcome for participants, i.e. what participants would have experienced had they not participated into the program.

Although a RDD lacks random assignment of units to the treatment group, it shares some interesting features with an experimental design. Throughout this section we will focus on the so called *sharp* RDD (Trochim, 1984), that is a RDD where the participation status is determined according to the following deterministic function of the observable characteristic S

$$D ? 1(S ? \bar{s}), \tag{3}$$

where $\bar{s}$ is a threshold for selection. This is actually the original formulation of such a design discussed by Campbell (1964). Units are assigned to the treatment if and only if they score at or above $\bar{s}$, implying that the probability of participation conditional on S is discontinuous at $\bar{s}$ stepping from zero to one as S crosses the threshold $\bar{s}$. Exploiting the relationship between S and D in (3), it follows that in a sharp RDD the following condition holds true

$$(Y^T ?Y^{NT}) ? D ?S ? \bar{s} \tag{4}$$

Because of the similarity with condition (1), a sharp RDD is often referred to as a quasi-experimental design (Cook and Campbell, 1979). In this context, conditioning on S allows to identify the average impact of the program on individuals scoring $\bar{s}$, thus a local version of the parameter in (2). In fact, in a neighbourhood of $\bar{s}$ this design presents the same features of a pure randomized experiment, since for any positive ? the following condition holds approximately

$$(Y^T ?Y^{NT}) ?S ? \bar{s} ? ? \quad ? \quad (Y^T ?Y^{NT}) ?S ? \bar{s} ? ?.$$

Note that to meaningfully define marginal units (with respect to $\bar{s}$), S needs to be continuous. In a finite sample for the condition to hold ? needs to go to zero at a proper rate as the sample size grows to infinity, implying a non-standard asymptotic theory for the resulting estimator of the mean impact (see Hahn *et al*., 2001, and Porter, 2002).

In some cases, units do not comply with the mandated status, dropping out of the program or seeking alternative treatments. Any of these violations of the original assignment might lead to biased conclusions on program effects, since conditions (1) or (4) are no longer valid. In fact, the presence of non-complying units when the assignment mechanism is given by (3) still makes the probability of participation discontinuous at $\bar{s}$, but the treatment status is no longer a deterministic function of the variable S (see Hahn *et al*., 2001, and Battistin and Rettore, 2002).

Two major drawbacks hamper the applicability of RDDs. First, in an observational study it is more often the case that units self-select into the treatment rather than being exogenously selected on a pre-program measure. Secondly, even in those instances in which the RDD applies, such a design is not informative about the impact on units away from $\bar{s}$. These are the two issues we look at in the next sections.

3. A GENERALIZATION OF THE SHARP REGRESSION DISCONTINUITY DESIGN

3.1 Identification results

This section discusses how to use the properties of a RDD within a set-up encountered in the evaluation of many social policies. Suppose that the program is targeted to a specific group of individuals whose eligibility depends on a known characteristic (such as age, income or unemployment duration) and that, conditional on eligibility, individuals choose to participate into the program according to a process unknown to the analyst. Without loss of generality, suppose that eligible individuals are those whose values of the variable S are above a certain threshold $\bar{s}$. Accordingly, the eligibility status is determined on the basis of a deterministic rule. If all eligible units participated into the program, a sharp RDD would arise and the mean impact on units in a neighbourhood of $\bar{s}$ would be identifiable.

In fact, it is widespread evidence that not all eligible units participate into the program they are eligible for. Heterogeneity in the information available on the program, individual preferences and opportunity costs are factors likely to influence participation in several instances. In this situation, the probability of participation still varies discontinuously as a function of S but is no longer a deterministic function of this variable. In fact, the probability of participation is zero for those individuals who are not eligible for the program ($S<\bar{s}$), and is less than one for those who are instead eligible for it ($S=\bar{s}$). The resulting design is often called *fuzzy* RDD (Trochim, 1984).

As a result of both the eligibility rule and the process leading to participation, the population turns out split into three subgroups: *non-eligibles*, *eligible non-participants* and *participants*. To label these subgroups we introduce a further binary variable to distinguish, amongst individuals whoa are eligible for the treatment, those who actually receive it. Let Z be the program eligibility status, with Z=1 (Z=0) for individuals who are eligible (ineligible) for the program, and let D be the binary variable for the treatment status as defined above. Non-participants are a mixture of those individuals who don't meet eligibility criteria (Z=0) and those who choose not to enter the program, (Z=1, D=0). It is worth noting that, contrary to the case considered in the previous section, in this set-up participation of eligible individuals into the program does not take place by design but is due to self-selection.

Let

$$E(Y^{T} \mid Z=1, D=1, S=s) \tag{5}$$

be the mean outcome for eligible units scoring S=s and actually receiving the treatment, with $S\geq\bar{s}$. This quantity is identified exploiting information on participants for any given value of S. Let

$$E(Y^{NT} \mid Z=1, D=1, S=s) \tag{6}$$

be the counterfactual mean outcome for the same group of units, i.e. what their response would have been had they not participated into the program. The mean impact of the program on treated units scoring S=s is then defined as the difference between factual and counterfactual results in (5) and (6)

$$\tau(s) = E(Y^{T} \mid Z=1, D=1, S=s) - E(Y^{NT} \mid Z=1, D=1, S=s).$$

Accordingly, the mean impact on participants τ is obtained as a weighted average of these quantities, with weights given by the proportion of eligible units scoring S=s.

Neither $\tau(s)$ nor τ are directly identifiable, since the counterfactual mean outcome in (6) is not observed by construction. Nor we can replace it by the factual mean outcome observed for eligible non-participants. In fact, due to the self-selection process determining the group of participants (i.e. those for whom Z=1 and D=1) and the group of non-participants (i.e. those for whom E=1 and D=0), eligible non-participants are not a random sample from the pool of eligible units, implying that in general

$$E(Y^{NT} \mid Z=1, D=0, S=s) \tag{7}$$

is different from (6). Note that this result holds true for any given value of S, in particular when $S=\bar{s}$.

Suppose that the information on outcomes experienced by non-eligible units (Z=0) is available to the analyst. Since this group of units is by construction characterized by values of S below the threshold for selection $\bar{s}$, it cannot be used to approximate the counterfactual outcomes of participants. Nor we can use non-eligible units in a neighbourhood of $\bar{s}$ to approximate the counterfactual mean outcome of participant units in a neighbourhood of $\bar{s}$. The quantity

$$E(Y^{NT} \mid Z = 0, S = \bar{s}) \tag{8}$$

is in fact different from the counterfactual result (6) evaluated at $\bar{s}$ because of the non-random selection of units into the treatment group discussed above. Non-eligibles alone do not allow solving the problem.

It is the joint use of information on non-eligibles *and* eligible non-participants to allow solving the problem (at least for a particular subpopulation of participants). The key relationship to obtain this result is the following

$$E(Y^{NT} \mid Z = 1, S = \bar{s}) = E(Y^{NT} \mid Z = 0, S = \bar{s}), \tag{9}$$

which is a straightforward implication of the eligibility rule. In a neighbourhood of the cut-off point $\bar{s}$ eligible and non-eligible units are nearly alike with respect to S, so that in the counterfactual scenario the two marginal groups would have experienced the same mean outcome. This result rests on the sharp RDD as reviewed in the previous section. The left-hand side of equation (9) can be written as a weighted mean of outcomes experienced by eligible participants and eligible non-participants, respectively, in a neighbourhood of $\bar{s}$

$$E(Y^{NT} \mid Z = 1, D = 1, S = \bar{s})\pi(\bar{s}) + E(Y^{NT} \mid Z = 1, D = 0, S = \bar{s})(1 - \pi(\bar{s}))$$

where $\pi(\bar{s})=\Pr(D=1|Z=1,S=\bar{s})$ is the probability of self-selection into the program for units marginally eligible. Substituting the last expression in (9) we obtain

$$E(Y^{NT} \mid Z = 1, D = 1, S = \bar{s}) = \frac{E(Y^{NT} \mid Z = 0, S = \bar{s})}{\pi(\bar{s})} - \frac{E(Y^{NT} \mid Z = 1, D = 0, S = \bar{s})(1 - \pi(\bar{s}))}{\pi(\bar{s})}. \tag{10}$$

Namely, the counterfactual mean outcome for participants presenting $S=\bar{s}$ is a linear combination of the factual mean outcome for non-eligible units at $S=\bar{s}$ and of the factual mean outcome for eligible non-participants at $\bar{s}$. The coefficients of the linear combination add up to one and are function of the probability $\pi(\bar{s})$, which is identifiable. Hence, $\pi(\bar{s})$ - the mean impact on participants at $\bar{s}$ - is identifiable and it can be expressed as

$$\frac{E(Y \mid Z = 1, S = \bar{s}) - E(Y \mid Z = 0, S = \bar{s})}{\pi(\bar{s})},$$

by subtracting (10) from (5). The last expression can be interpreted as the ratio of the intention to treat effect, the mean impact we would observe if all eligible units actually took part in the program, to the mean impact of Z on D at $\bar{s}$. Results by Hahn *et al.* (2001) and Porter (2002) on non-parametric estimation in a RDD straightforwardly apply.

Note that condition (9) is the cornerstone on which we build the result. Otherwise stated, it is crucial that the eligibility rule is determined according to a *sharp* RDD. The major implication is that, although the RDD described in this section is *fuzzy*, the regularity conditions for identification of $\pi(\bar{s})$ are those required in a *sharp* RDD (and are therefore *weaker*). Moreover, to derive the result we don't need to specify how eligible units self-select into the treatment. Thus, identifiability of $\pi(\bar{s})$ doesn't require any behavioural assumption on the selection process. However, in order to obtain identification we need access to information for three different groups of units: participants, eligible non-participants and non-eligibles.

3.2 Related results

In a fully randomized experiment, Bloom (1984) deals with the case where some units assigned to the program do not actually participate (no-shows). Exploiting information on participants, eligible non-participants and non-eligibles the author proves that the mean impact on participants is identifiable. The result in the previous section can be seen as a special case of Bloom (1984) since it is as if randomization took place at the threshold for eligibility $\bar{s}$. In our case eligible non-participants at $\bar{s}$ play the role of Bloom's (1984) no-shows.

Our result (as well as Bloom's one) can also be derived as a special case of Angrist and Imbens (1991). The authors prove that, even if participation takes place as a result of self-selection, the mean impact on participants is identifiable provided that (i) there exists a random variable A affecting the participation into the program and orthogonal to the potential outcomes (Y^T,Y^{NT}) and (ii) the probability of participation conditional on A is zero for at least one value of A. Condition (i) qualifies A as an Instrumental Variable for the treatment status.

In Bloom's (1984) context, self-selection arises as a consequence of the non-complying behaviour of some units randomly assigned to the program. The natural choice for A in that case is the mandated status as it results from randomization. Condition (i) is satisfied since Pr(D=1|A=1)>Pr(D=1|A=0) and A is orthogonal to the potential outcomes while condition (ii) is satisfied since Pr(D=1|A=0)=0. In our case, since Z is orthogonal to the potential outcomes in a neighbourhood of $\bar{s}$ and Pr(D=1|Z=0)=0, Z meets the conditions stated by Angrist and Imbens (1991) in a neighbourhood of $\bar{s}$. Hence the identification of the mean impact on participants at $\bar{s}$ follows.

4. VALIDATING NON-EXPERIMENTAL ESTIMATORS OF THE MEAN IMPACT ON PARTICIPANTS

4.1 Specification tests

In the previous section we have shown that the existence of an eligibility rule allows to identify the mean impact of an intervention on marginally eligible participants even if participants are self-selected from the eligible pool. If the treatment effect is heterogeneous with respect to S, the impact on marginal participants is not informative on the impact of the intervention on units away from the threshold for eligibility. Nor non-eligible units and eligible non-participants can be used as valid comparison groups, since they differ systematically from participants (the former with respect to S and the latter with respect to the variables determining self-selection).

In order to identify the mean impact on the overall population of participants, one has to resort to one of the long array of non-experimental estimators available in the literature which adjust for the selection bias under different assumptions (see Heckman *et al*., 1999, and Blundell and Costa Dias, 2000, for a review). The main problem with this approach is that alternative estimators for the parameter of interest are consistent under assumptions that most of the times are not testable.

Over the years the literature took two main routes to deal with this problem. The first route amounts to seek whether any over-identifying restriction on the data generating process arises from a behavioural theory of the phenomenon under investigation, possibly exploiting such a kind of restriction to test the assumptions on which the non-experimental estimator rests (see Rosenbaum, 1984 and Heckman and Hotz, 1989).

The second route is feasible only when an experimental design has been run, so that an experimental estimate of the impact comes available. Then, besides estimating the mean impact, one can exploit the experimental set up to study the selection bias and to assess whether the non-experimental estimators are able to reproduce the experimental estimate (see LaLonde, 1986 and Heckman *et al*., 1998a). When information from a randomized experiment is available, one can meaningfully check how closely non-experimental comparison groups methods approximate experimental impact estimates. At the same time, this allows us to assess the performance of alternative non-experimental estimators for the treatment effect, thus suggesting the best strategy to follow when experimental data are not available.

In this section we show that if information is available on the three groups of units resulting from the set-up of Section 3.1, then one can test the validity of any non-experimental estimator on a specific subpopulation. To fix the ideas, we will focus on the well-known matching estimator, but the same line of reasoning applies to other non-experimental estimators. The key assumption on which the matching estimator rests is that all the variables driving the self-selection process *and* correlated to the outcome are observable to the analyst. Formally, the assignment to the treatment is told *strongly ignorable* given a set of characteristics x if, conditional on x, the treatment can be thought as randomly assigned to units provided that at each value x there is a positive probability of being treated

$$(Y^T ? Y^{NT}) ? D ? x ? 0 ? Pr(D ? 1 ? x) ? 1 ? \quad (11)$$

If this condition holds, then it is as if units were randomly assigned to the treatment with a probability depending on x; the counterfactual outcome for participants presenting characteristics x can be approximated by the actual outcome of non-participants presenting the same characteristics. Since units presenting x have a common probability to enter the program, then an operational rule to obtain an *ex post* experimental-like data set is to match participants to non-participants on such probability (the so called *propensity score*), whose dimension is invariant with respect to the dimension of x (see Rosenbaum and Rubin, 1983).

The critical assumption of this procedure is that the set x is rich enough to guarantee the orthogonality condition in (11). In principle, this imposes strong requirements on data collection. Moreover, the violation of the second

condition in (11) would raise the so called common support problem (see for example Heckman *et al.*, 1998a, and Lechner, 2001).

Let

$$sb(s) = E(Y^{NT} \mid E = 1, D = 1, S = s) - E(Y^{NT} \mid E = 1, D = 0, S = s) \quad (12)$$

be the *selection bias* affecting the raw comparison of eligible participants to eligible non-participants. The first term on the right-hand side is a counterfactual mean outcome while the second is a factual one. This quantity captures pre-intervention differences between eligible units self-selected in and out of the intervention, respectively, at each level of S, with $S = \bar{s}$.

Using the results of the previous section, the mean counterfactual outcome for participants is identifiable in a neighbourhood of $\bar{s}$ by means of (10). This also implies that the selection bias for units marginally eligible, $sb(\bar{s})$, is identifiable as the difference between (10) and (7) evaluated at $\bar{s}$. Note that the identification of the counterfactual term on the right-hand side of (12) at $\bar{s}$ exploits information on the subgroup of non-eligible units closest to the group of eligible units. Apparently, identification is precluded as S moves away from $\bar{s}$.

Then, let

$$sb(s,x) = E(Y^{NT} \mid E = 1, D = 1, x, S = s) - E(Y^{NT} \mid E = 1, D = 0, x, S = s)$$

be the selection bias on the specific subpopulation indexed by x, where x are the variables exploited to account for the selection bias in a matching estimation of the intervention impact. If the orthogonality condition in (11) holds, then sb(s,x)=0 uniformly with respect to x and S. In particular, a necessary condition for the matching estimator to work is that $sb(\bar{s},x)=0$, which is directly testable.

Operationally, in a neighbourhood of $\bar{s}$ any test of the equality of the mean outcomes of the non-eligible units and of the eligible non-participants, respectively, conditional on x is a test of the strong ignorability of the assignment to intervention, thus a test of the validity of the matching estimator. Clearly, the rejection of the null hypothesis is sufficient to conclude that condition (11) does not hold.

On the other hand, on accepting the null hypothesis one might feel more confident in using the matching estimator but by no means it can be said that the validity of the estimator has been proved. In fact, the specification test tells nothing on whether the strong ignorability condition holds away from $\bar{s}$.

4.2 Related results

Since the RDD can been seen as a formal experiment at $\bar{s}$, the specification test developed above displays a similarity to what Heckman *et al.* (1998a) develop in an experimental set-up. In both cases there is a benchmark estimate of the intervention mean impact - the RDD estimate in the former, the experimental one in the latter - to which the analyst is ready to attach credibility. Then, the analyst tests non-experimental estimators against the benchmark to discover whether the assumptions they rest upon are met.

The similarity between the two approaches stops here. On the one hand, the availability of an experimental set-up as in Heckman *et al.* (1998a) allows to fully characterize the selection bias and to test non-experimental estimators with reference to the population of participants. If a RDD is available, this is feasible only with reference to the population of participants at $\bar{s}$.

On the other hand, it is very often the case that an intervention is targeted to a population of eligible units among which it is actually delivered only to those showing up to participate while it is much less frequent to have available an experimental set-up. Then, the three groups of units needed to implement the results in this paper in principle should be available. Whether they are actually available it depends on the design of the data collection. This opens the door to a routinely application of the specification test based on the RDD as a tool to validate non-experimental estimators of the mean impact on participants.

Rosenbaum (1987) in his discussion of the role of a second control group in an observational study gives an example (example 2 on p. 294), which resembles very closely the set-up we refer to. The Advanced Placement (AP) Program provides high school students with the opportunity to earn college credits for work done in high school. Not all high schools offer the AP program, and in those that do, only a small minority of students participate. Two comparison groups naturally arise in this context, (i) students enrolled in high school not offering the program and (ii) students enrolled in high schools offering the program who did not participate.

Then, Rosenbaum (1987) goes on discussing how the availability of two comparison groups can be exploited to test the strong ignorability condition needed to believe the results of a matching estimator.

Apparently, the first comparison group resembles our pool of non-eligible units while the second comparison group resembles our pool of eligible non-participant units. The main difference between Rosenbaum's example and our set-up is that in the former case the rule according to which high schools decide whether to offer the AP program or not is unknown to the analyst while in our set-up the eligibility rule is known. It is exactly this feature to allow identifying the mean impact on participants as well as the selection bias at $\bar{s}$.

5. CONCLUSIONS

The main message from this paper is that every time an intervention is targeted to a population of eligible units but is actually administered to a sub-set of self-selected eligible units, it is worth collecting information separately on three groups of units: non-eligibles, eligible non-participants and eligible participants. Also, the variables with respect to which eligibility is established have to be recorded.

The relevance of distinguishing between non-eligibles and eligible non-participants to improve the comparability of treatment and control groups has already been stressed in the literature (see, amongst others, Heckman *et al*., 1998a). We have shown that, if the eligibility rule is based on a continuous variable and information is observed for both non-eligibles and eligible non-participants, the mean impact on participants who are marginally eligible for the program is identified, no matter how self-selection of participants takes place. We have also shown that the resulting design fits a *fuzzy* RDD but that discontinuities in the eligibility rule lead to regularity conditions for identification characterizing a sharp RDD. Finally, we have shown that as a straightforward consequence of the previous result also the selection bias for units on the margin between eligibility and non-eligibility is identifiable. Such a results suggests the use of a specification test in a neighborhood of the threshold for eligibility so that the properties of non-experimental estimators can be assessed. By design, such a test is informative on the performance of non-experimental estimators only for a particular subgroup of units, thus results cannot be generalized to the whole population (unless we are willing to impose further identifying restrictions). The value of the specification test is that if it rejects the estimator locally then this is sufficient to reject it altogether.

REFERENCES

Angrist, J.D (1998), "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants", Econometrica, 66.

Angrist, J.D, and G.W. Imbens, (1991), "Sources of Identifying Information in Evaluation Models", *NBER Technical Working Paper* 117.

Battistin, E., and E. Rettore (2002), "Testing for programme effects in a regression discontinuity design with imperfect compliance", *Journal of the Royal Statistical Society A*, Vol. 165, No. 1, pp. 1-19.

Bloom, H.S. (1984), "Accounting for No-Shows in Experimental Evaluation Designs", *Evaluation Review*, 8, pp. 225-246.

Blundell, R., and M. Costa Dias (2000), "Evaluation methods for non-experimental data", *Fiscal Studies*, 21, 4, pp. 427-468.

Campbell, D.T. (1964), *Reforms as experiment*.

Cook, T.D., and D.T. Campbell (1979), *Quasi-Experimentation. Design and Analisys Issues for Field Settigs*, Boston: Houghton Mifflin Company.

Friedlander, D., Greenberg, D.H., and P.K. Robins (1997), "Evaluating Government Training Programs for the Economically Disadvantaged", *Journal of Economic Literature*, 35, 4, pp. 1809-1855.

Hahn, J., Todd, P., and W. Van der Klaauw (2001), "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design", *Econometrica*, 69, 3, pp. 201-209.

Heckman, J.J., and V.J. Hotz (1989), "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training", *Journal of the American Statistical Association*, 84, pp. 862-874.

Heckman, J.J., and J. Smith (1995), "Assessing the case for social experiments", *Journal of Economic Perspectives*, 9, 2, pp. 85-110.

Heckman, J.J., Ichimura, H., Smith, J., and P. Todd (1998a), "Characterizing Selection Bias Using Experimental Data", *Econometrica*, 66, pp. 1017-1098.

Heckman, J.J., Smith, J., and C. Taber (1998b), "Accounting for Dropouts in Evaluations of Social Experiments", *The Review of Economics and Statistics*, 80, 1, pp. 1-14.

Heckman, J.J., Lalonde, R., and J. Smith (1999), "The Economics and Econometrics of Active Labor Market Programs", in Ashenfelter, A. and D. Card (eds.) *Handbook of Labor Economics, Volume 3*, Amsterdam: Elsevier Science.

LaLonde, R. (1986), "Evaluating the econometric evaluations of training programs with experimental data", *American Economic Review*, 76, pp. 604-20.

Lechner, M. (2001), "A note on the common support problem in applied evaluation studies", *Discussion Paper 2001-01*, Department of Economics, University of St. Gallen.

Porter, J. (2002), "Asymptotic bias and optimal convergence rates for semiparametric kernel estimators in the regression discontinuity model", unpublished manuscript, Harvard University.

Rosenbaum, P.R. (1984), "From Association to Causation in Observational Studies: The Role of Tests of Strongly Ignorable Treatment Assignment", *Journal of the American Statistical Association*, 79, 385, pp. 41-48.

Rosenbaum, P.R. (1987), "The Role of a Second Control Group in an Observational Study", *Statistical Science*, 2, 3, pp. 292-306.

Rosenbaum, P.R., and D.B. Rubin (1983), "The central role of the propensity score in observational studies for causal effects", *Biometrika*, 70, pp. 41-55.

Rubin, D.B. (1974), "Estimating causal effects of treatments in randomized and nonrandomized studies", *Journal of Educational Psycology*, 66, pp. 688-701.

Rubin, D.B. (1977), "Assignment to Treatment Group on the Basis of a Covariate", *Journal of Educational Statistics*, 2, pp. 4-58.

Trochim, W. (1984), *Research Design for Program Evaluation: the Regression-Discontinuity Approach*, Beverly Hills: Sage Publications.

Van der Klaauw, W. (2002), "Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Approach", *International Economic Review*.