October 19, 2010

Using Regression Discontinuity with Implicit Partitions: The Impacts of *Communidades Soliarias Rurales* on Schooling in El Salvador

PRELIMINARY DRAFT

Please do not cite without authors' permission.

Alan de Brauw
and
Daniel Gilligan*

Abstract

Regression discontinuity design is a useful tool for evaluating programs when a specific variable is used to determine program eligibility. In this paper, we show that under specific conditions, regression discontinuity can be used in instances when more than one variable are used to determine eligibility, without mapping those variables into a single measure, through the use of a distance metric and by creating an implicit partition between groups. We apply this model to the case of *Communidades Solidarias Rurales* in El Salvador, which used partitioned cluster analysis to determine the order communities would enter the program. Using data collected for the evaluation as well as data from the 6th National Census of El Salvador, we demonstrate that the program increased both *parvularia* and primary school enrollment among children aged 6 to 12 years old. Among children of primary school age, we further show that enrollment gains were largest among younger children and older girls.

*Alan de Brauw and Daniel Gilligan are Senior Research Fellows, International Food Policy Research Institute, 2033 K Street NW, Washington, DC 20006. We thank Amber Peterman, Margarita Beneke de Sanfeliu, and Mauricio Shi Artiga for contributions and suggestions that have strengthened this paper. Please direct correspondence to Alan de Brauw at alandebrauw@gmail.com or at the address listed above. All remaining errors are our responsibility.

Using Regression Discontinuity with Implicit Partitions: The Impacts of *Communidades Soliarias Rurales* on Schooling in El Salvador

Regression discontinuity methods have become increasingly popular in evaluating the impacts of social programs in the economics literature. In general, evaluations have been based on single thresholds that determine program eligibility. Since the threshold is arbitrary from the perspective of the unit of intervention, units that are just eligible for the program– or have values of the metric "close" to the threshold– can be compared with units that are just not eligible, to measure the local average treatment effect of the program.

It is not necessarily the case that one metric determines program eligibility. Rather, governments or agencies charged with determining program eligibility may instead choose to use two or more metrics. If two or more metrics are used, a common way to map these into a single measure is through a mathematical function, and if this procedure is used then regression discontinuity is still simple to use if specific assumptions are met (e.g. Imbens and Lemieux, 2008). A good example of such a procedure is when program eligibility is determined by a proxy means test, which effectively turns several measures into one measure, which can then be used to determine strict program eligibility.

However, one does not necessarily need to use a well-defined function exists to determine program eligibility. Other statistical procedures can be used to determine program eligibility that do not require mapping n characteristics into one variable through a function. For example, partitioned cluster analysis can be used to identify similar groups within data, which classifies individual observations into similar clusters of observations. If a subset of those clusters are then assigned a treatment, then the treatment status is completely determined by cluster membership, and therefore by the metrics used in assigning units to clusters. But an explicit threshold between treatment clusters and control clusters does not exist, so one cannot immediately perform regression discontinuity to determine program impacts.

In this paper, we develop a set of additional assumptions needed to use

standard regression discontinuity methods to evaluate programs that determine treatment status using partitioned cluster analysis or similar methods. The idea behind the estimator is that we use the distance metric that determines clustering in the data to implicitly define the threshold between treatment and control groups as a function of the distance between cluster centers. Under these quite reasonable assumptions, we show that the threshold can then be used in a sharp regression discontinuity estimator using the distance from each point to the threshold in estimation.

We then apply this methodology to evaluate a specific program, *Communidades Solidarias Rurales* (CSR) in El Salvador, that used partitioned cluster analysis to determine the order in which municipalities would enter CSR, as well as which municipalities would receive CSR. Using both data from the evaluation of CSR as well as census data from El Salvador, we compare schooling outcomes among households in municipios that entered CSR in 2006 with municipios entering in 2007. We find that close to the threshold, children in the 2006 entry group are more likely to have enrolled in *parvularia* at age 6, and that school enrollment rates among primary school age children increase by 4 percentage points. Using the census data, we further disaggregate these results by age and gender.

The paper proceeds as follows. First, we briefly review the one-dimensional regression discontinuity estimator, including assumptions necessary for the estimator to provide an unbiased estimate of the treatment effect. Second, we provide a brief description of partitioned cluster analysis. Third, we develop conditions for an N-dimensional regression discontinuity estimator to be valid. The fourth section presents basic information about CSR and describes the data sources used for analysis. The fifth section presents results and the sixth section concludes.

1 Regression Discontinuity Designs

Regression discontinuity designs are typically referred to as sharp and fuzzy designs. The estimator we will develop follows the sharp design, so we review it here. Following the notation of Imbens and Wooldridge (2008), we can consider two potential outcomes for unit i, namely $Y_i(0)$ and $Y_i(1)$, where the difference $Y_i(1) - Y_i(0)$ is the definition of the causal effect of the treatment. The observed outcome is equal to:

$$Y_i = (1 - W_i) \cdot Y_i(0) + W_i \cdot Y_i(1) \tag{1}$$

where $W_i \epsilon \{0, 1\}$ is the treatment indicator variable. The idea behind a regression discontinuity evaluation is that there is a variable X_i that completely determines whether or not a unit receives the treatment. Calling this threshold c, a unit will receive the treatment if $X_i \geq c$, which implies:

$$W_i = 1\{X_i \geq c\} \tag{2}$$

In a sharp regression discontinuity design, all units with a value of X_i that is at least c do receive the treatment, and those units with a value of X_i below c do not receive the treatment, effectively becoming the control group. The average treatment effect δ if the difference between the mean outcome for units with values of X_i just below the threshold (Y^-) and just above the threshold (Y^+). It can then be written as the difference in conditional expectations between units just above and below the threshold:

$$\delta = Y^+ - Y^- = \lim_{\varepsilon \to 0} E(Y_i(1)|X_i = c + \varepsilon) - E(Y_i(0)|X_i = c - \varepsilon) \tag{3}$$

for $\varepsilon > 0$.

To estimate δ, one needs to estimate both Y^+ and Y^-. Then Y^+ and Y^- must be estimated, and one can quite generally write the solution to the estimation problem in the form of non-parametric regressions:

$$\hat{Y}^{+} = \frac{\sum_{i} X_i > c \, Y_i K(\frac{X_i - c}{h})}{\sum_{i} X_i \geq c \, K(\frac{X_i - c}{h})} \tag{4}$$

$$\hat{Y}^{-} = \frac{\sum_{i} X_i < c \, Y_i K(\frac{c - X_i}{h})}{\sum_{i} X_i \geq c \, K \frac{c - X_i}{h})} \tag{5}$$

where $K(\cdot)$ represents a kernel estimator, and h represents the chosen bandwidth. The major problem here is to choose a kernel function that will identify the effect at the single point of interest, the threshold, as well as a proper bandwidth. Porter (2003) shows that as the bias in the estimate using the rectangular kernel is linear in the bandwidth h, whereas the bias in non-parametric estimators generally is of order h^2 in non-parametric estimators. In this paper, we vary the bandwidth to test the sensitivity of results to inclusion or exclusion of observations farther away from the threshold.[1]

Three assumptions are critical for the consistency of the sharp regression discontinuity estimator (Edmonds, Mammen, and Miller 2005). First, the probability of treatment must vary discontinuously at the threshold. Intuitively, the sharp cutoff point serves as an instrumental variable that affects program participation but does not independently affect outcomes. Second, observations just above and below the threshold must be similar in both their observed and unobserved characteristics. Third, one must assume that if the treatment did not occur, the outcome Y_i would be continuous at the threshold. In other words, there would be no sharp break in outcome measures in the population at large for those just below and just above the threshold without the program. Another way to think of the RDD estimator is as generating a locally randomized experiment (Lee and Lemieux, 2008), as if the continuity assumption is met, then the forcing variable takes on almost the same value at the threshold.

[1] See Ludwig and Miller (2007) and Imbens and Lemieux (2008) for details on methods of choosing the bandwidth in sharp regression discontinuity applications.

2 Partitioned Cluster Analysis

Partition cluster analysis encompasses a set of iterative methods of breaking observations in data sets into distinct groups that are similar. The concept behind partition cluster analysis is to break up observations in large data sets into clusters of observations with relatively similar attributes. Observations are grouped along those attributes using an iterative procedure that proceeds as follows. First, the analyst chooses the number of groups, k, they are interested in forming. Second, the analyst chooses k initial points (centers), the distance from each observation to each initial point is calculated, and each observation is assigned the center nearest to it. The centers of each cluster are recalculated as either the mean or median of the points aligned with that center, and distances to the new k centers are computed. Clusters are then reformed if observations switch groups, and the procedure is repeated until all points no longer switch clusters.[2] We define the center of the k-th cluster as η_k, and the distance between any observation $\mathbf{X}_i$ and η_k can be defined as $d(\mathbf{X}_i, \eta_k)$. Cluster membership is then defined for all clusters k as the set of all X_{ik} such that $X_{ik} = \arg\min_{k \in K} (d(\mathbf{X}_i, \eta_k))$.

The analyst makes several choices that affect the results of a partitioned cluster analysis. First, the analyst chooses the number of clusters k; the initial centers of each cluster; the distance measure to be used; and the method of choosing new cluster centers (mean or median). After these choices are made, however, the results of the cluster analysis will always be the same; that is, they are always replicable with the same data. Furthermore, important from the perspective of performing a regression discontinuity analysis on partition clustered data, each point distinctly belongs to one cluster. Therefore the estimation strategy suggested is similar to that in sharp regression discontinuity.

[2]When the mean is used as the center, partitioned cluster analysis is usually known as cluster k-means analysis; when it is the median, it is known as cluster k-medians analysis.

3 Implicitly Defining the Threshold

When program assignment is performed using partitioned cluster analysis, then there is no true forcing variable. In this section, we describe additional assumptions under which an implicit forcing variable can be defined when partitioned cluster analysis is used with M traits.

To set up the estimator, assume a sample of N individuals indexed by i, who all have an M member vector of traits, $\mathbf{X_i}$. The traits are to be used to determine which individuals receive the "treatment" and which individuals do not receive the treatment. Assume that each element of $\mathbf{X}$ is positive (e.g. $X_j \geq 0 \; \forall \; j \in M$). A partition cluster analysis is performed on these individuals, to create K clusters, and each individual can then be indexed as $\mathbf{X}_{ik}$. Individuals are then partitioned into clusters, and the analyst chooses A of the clusters to receive the intervention (treatment clusters) and the remaining $B = K - A$ clusters remain as control clusters.

Given that each cluster has a well defined center, the centers implicitly define boundaries between all K of the clusters. These boundaries can be defined as follows. Consider any two cluster centers, η_c and η_d. There must be a set of points that are equidistant from the two cluster centers, $\mathbf{W}_j$, which define a boundary between those two cluster centers, $d(\mathbf{W}_j, \eta_c) = d(\mathbf{W}_j, \eta_d)$. This logic can extend to all $\frac{K(K-1)}{2}$ of the clusters; between any two clusters, there must be a boundary defined as the set of points that are equidistant from the two cluster centers.

Since boundaries exist between all clusters, it must be that a boundary exists between any adjacent treatment and control clusters. Furthermore, we can also define a further boundary that relates the closest center of a treatment cluster to the closest center of a control cluster. This boundary separates treatment clusters from control clusters. If we define this set of points as $\mathbf{Z}_j$, the set of points acts as the threshold between treatment and control observations. It can more formally be defined as:

$$\min_{a \in A} d(\mathbf{Z}_j, \eta_a) = \min_{b \in B} d(\mathbf{Z}_j, \eta_b) \tag{6}$$

So long as equation (6) conforms to three specific assumptions, the implicit function theorem states that equation (6) defines a function and so with the choice of a distance metric, that function defines a specific implicit forcing variable that can be used with which to perform RDD estimation.[3]

The first assumption is that the function defined by equation (6) must be continuous. If it is not continuous, then the control group for either certain treatment clusters or specific treatment observations might not be well-defined. Under certain conditions– for example, if an analyst was focusing on one specific treatment cluster among many– then local continuity would suffice for all points local to that specific treatment cluster and its center.

Second, it must be that the solution to equation (6) is unique; there can only be one vector Z_j should be the only set of points that in other words, we need one set of for each set of we must make two assumptions to ensure that a unique solution holds. If a unique solution does not exist, then two or more potential boundaries exist, and it is not possible to ascertain which of the boundaries would correctly define the forcing variable. From a mathematical perspective, if uniqueness did not exist, then the implicit function theorem would break down for equation (6).

Third, we make the assumption that all of the indicators used in partitioned cluster analysis increase the likelihood of being in the treatment. Formally, assume that $\mathbf{X}_t$ is a point in a treatment cluster. Then for any point $\mathbf{X}_j$, if $x_{js} \geq x_{ks}$ for $s = 1, \ldots, M$, then it must be $\mathbf{X}_j$ is also in the treatment. This assumption simply implies that the treatment clusters are all in generally the same neighborhood, as are the control clusters. It rules out choosing a cluster that is quite dissimilar from other clusters as a control cluster.

Under these three assumptions, the solution to equation (6) implicitly defines a boundary in one dimension– distance– between the treatment and control clusters, and as a result one can perform regression discontinuity using the distance from that set of points as the forcing variable. To illustrate this concept, consider the two dimensional example in Figure 1. The treatment cluster must

[3]The clear choice of a distance metric is the same distance metric used to perform partitioned cluster analysis.

be A by the third assumption, whereas there are two control clusters (B and C). The boundary between the treatment and control clusters is the set of points that are equidistant between the centers A and B when the center of B is closer, and then between A And C when the center of C becomes closer.

To finalize the estimator, given that the set of points $\mathbf{Z_j}$ defines the boundary between the treatment and control groups, the distance from an individual's traits $\mathbf{X_{ik}}$ to $\mathbf{Z_j}$ defines how close it is to the boundary. We can define this distance δ as:

$$\delta(\mathbf{X_{ik}}) =\mid \mathbf{X_{ik}} - \mathbf{Z_j} \mid \tag{7}$$

Given the process of partitioned cluster analysis described in the previous section, points that are arbitrarily close to the boundary would be the ones that, with a small change in one of the indicators making up $\mathbf{X_{ik}}$, individual i might switch from a treatment to a control cluster, or vice-versa. Therefore, the distance metric δ acts as the forcing variable in regression discontinuity. In parallel with a more explicit forcing variable, one can imagine that small shifts in the values of specific components of the cluster analysis would shift these observations closer to the threshold from one cluster to another, whereas those closer to the cluster centers would be less similar.

For regression discontinuity estimates using $\delta(\cdot)$ as the forcing variable, two of the three main assumptions that are critical for the consistency of the sharp regression discontinuity estimator still apply. The first assumption, that the probability of treatment varies discontinuously at the threshold, clearly applies. However, it is still necessary that observations just above and below the threshold are similar in both observed and unobserved characteristics, and one must assume that any outcome Y_i would be continuous at the threshold in the absence of a program. If the assumptions needed to define the implicit distance metric and these latter two assumptions are true, then this estimator can be used to identify program impacts using regression discontinuity.

4 Communidades Solidarias Rurales

El Salvador began the CSR program (previously called Red Solidaria) in 2005 to begin to alleviate poverty in rural areas of its poorest municipios. The program is targeted in two ways: geographically and categorically. We describe geographic targeting below; individuals targeted for the health transfer are meant to improve health and nutrition among children under 5 years old, the education transfer aims to improve school enrollment among children who have not completed primary school. For the education transfer, households receive the transfer if all children between the ages of 6 and 15 who have not completed primary school are enrolled in school are both enrolled and attend more than 80 percent of the time each month. Households who receive either the education transfer or the health transfer receive $15 per month, and households that are eligible for both transfers receive $20 per month.

To select municipios that would participate in CSR, geographic targeting took place in two steps. First, municipios were grouped into four Extreme Poverty Groups (EPGs) using partitioned cluster analysis. The two criteria that were used were the severe poverty rate, measured using representative data collected at the municipio level between 2001 and 2004, and the prevalence of severe stunting (HAZ scores below -3) among first graders, measured in a census of first graders in 2000. The two measures were deemed to be alternative measures of poverty that were quite uncorrelated with one another, as such measuring different dimensions of poverty. The first two EPGs, municipios rated in severe and high extreme poverty, were deemed eligible for CSR. Within each EPG, municipios were then ranked by a municipality marginality index (IIMM in Spanish), which is a declining index of welfare based on poverty, education levels and housing conditions. The IIMM was used to prioritize municipios for entry within each EPG.

The municipios in the severe EPG entered CSR in 2005 and 2006, and the municipios in the high extreme poverty group entered between 2007 and 2009. To identify program impacts after one year, one can compare individuals residing in a set of municipios that enter in one year with individuals residing

in the municipios that enter the next year. For this paper, we focus on the difference in outcomes among individuals living in the 2006 entry group with outcomes among individuals residing in municipios entering CSR in 2007. The comparison works for educational outcomes as follows. Households beginning to receive benefits in 2006 began to receive them relatively late in the year, long after school enrollment decisions were made. As such, the transfers were not conditional in the first year. However, transfers became conditional in 2007, when the new school year began in January. The 2007 entry group did not make decisions regarding school enrollment that were conditional until 2008.

4.1 Identification of Benefits using RDD

To identify benefits of CSR using RDD, we use the comparison between municipios entering in 2006 and those entering in 2007; therefore, the data measuring outcomes after the program began must be measured in 2007. We use this comparison primarily for two reasons. First, RDD identifies local average treatment effects, rather than average treatment effects. We are most interested in how a conditional cash transfer program affects the poor, so it is sensible to look for benefits among the poorest group possible. As the evaluation of CSR began too late to credibly measure school enrollment in the poorest group of municipios (those entering in 2005), we choose the 2006 entry group as the poorest possible group among which to measure impacts. Second, the data available on educational outcomes is perhaps the richest in 2007, as we can work with the Salvadoran census which took place in May, 2007.

Since the 2006 entry group was in the severe EPG and the 2007 entry group is in the high EPG, the municipios are separated by partitioned cluster analysis rather than by IIMM ranking As a result, we use the estimation strategy described in the previous sections to develop an implicit partition between the two groups. The municipios in the 2006 and 2007 entry groups are illustrated in Figure 2, and clearly demonstrate that a continuous partition can be drawn between the 2006 and 2007 entry groups equidistant from the two cluster centers conforming to the three assumptions in section 3. Therefore the set of points

defined by the cluster centers defines an implicit boundary between the treatment and control groups, and we measure the difference in distance between the cluster centers as the forcing variable.[4]

Because of the rolling entry of the program, we can only evaluate impacts after one year of receiving benefits from the program.[5] However, given that the primary educational outcome we can observe is enrollment and benefits of the program are conditional on school enrollment, it is reasonable to think that benefits defined as additional school enrollment is immediate. We plan to study school enrollment primarily among two groups. First, we examine children who are between the ages of 7 and 12 years old for the purposes of enrollment; by that, we mean that they were between 7 and 12 years old at the beginning of the calendar year. By law, children in El Salvador are supposed to be 7 years old before entering first grade, so if a child normally progresses the child completes primary school when the child is 12 years old at the beginning of the year. Second, in the census data we examine school enrollment among 6 year olds, who are also required to be in a pre-school called *parvularia* to receive CSR transfers.

5 Data Sources

We use two data sources to evaluate the impact of CSR on school enrollment among 6 to 12 year old children in El Salvador. The primary data set we use was collected by FUSADES in collaboration with researchers at the International Food Policy Research Institute, and households included in the sample were designed explicitly to evaluate the impact of CSR on several indicators of infant and maternal health, education, and nutritional status, including some of the indicators used in this paper. The baseline data set was collected in January

[4]Other papers that use community level treatments in a regression discontinuity framework to study program impacts include Ludwig and Miller (2007), Battistin and Rettore (2008), and Buddelmeyer and Skoufias (2004).

[5]Behrman and King (2009) show that the timing of evaluations can affect findings related to program impacts; however, due to the sequential nature of CSR entry it is not possible to even measure benefits after two years of implementation, as in many impact evaluations.

and February of 2008, and included retrospective measures of school enrollment in a specially designed education module so that measures of school enrollment prior to CSR entry could be constructed. The survey form also included sections on household demographics health, time allocation and off-farm labor, housing and other consumer durables, agriculture, migration, other income sources, consumer expenditures, and community participation in programs, including CSR.

The sample included 100 cantones in 50 municipios, distributed among the 2006 to 2008 entry groups. Prior to the baseline, 15 households with children under 3 years old or with a pregnant women resident, and 15 households with children between the ages of 6 and 12 were selected randomly within each canton from census lists, for a total of 30 targeted households per canton. In some cantones, fewer households were actually interviewed. For the purposes of this paper, the most important aspect of this sample is that it included 10 municipios that entered CSR in 2006 and 11 municipios that entered in 2007, with a total of 1280 households between those municipios; the whole data set included 2775 households after cleaning. We focus on children in those households in the impact estimation, though we include children in the 2008 entry groups in descriptive statistics for comparative purposes.[6]

The baseline survey specifically asked about each child's enrollment in school in 2005, 2006, and 2007; because the conditionality only began in the 2007 school year for children in households receiving benefits in the 2006 entry group, we can use the 2006 and 2007 measures to estimate impacts in a difference-in-difference framework. Although one might be concerned about recall bias in the school enrollment measures from 2006, it should be noted that there is no reason to believe that recall bias should differ on average between the 2006 and 2007 entry groups. We use the impact evaluation survey data primarily to study impacts among children aged 7 to 12 years.

Because the school year begins at the beginning of the calendar year in El Salvador and the conditionality also only begins at the beginning of the year,

[6]For purposes of the evaluation, the 2008 entry group are broken up into "early" 2008 and "late" 2008, as several of the municipios did not enter CSR until after the second round survey in late 2008.

we can also use the El Salvador census that took place in May of 2007 to study enrollment rates, and the enrollment rates we can study are then percentages among the population, rather than estimates based on the much smaller impact evaluation sample. Questions in the census asked about school enrollment for all children, and so we can create a data set containing all of the children living in rural parts of the 2006 and 2007 entry groups to provide alternative estimates of impact among the 2006 entry group. The main drawback is that the estimates are necessarily only single difference, but have the advantage of being based on the population, and so we can estimate impacts among much smaller groups; for example, among age and gender disaggregated groups. When using the census data, we estimate impacts among 7 to 12 year olds and among 6 year olds.

5.1 Descriptive Statistics, Enrollment

To begin to consider how school enrollment may have changed over time as a consequence of CSR entry, we initially measure the proportion of children of each age between 7 and 12 are enrolled in school, in 2006 and 2007 (Table 1). It is worth noting that with the lone exception of 7 year olds, across the entire impact evaluation sample, more than 90 percent of children report having been enrolled in school in 2006. In 2007, the percentages increase among all ages, by between 0.7 and 2.9 percentage points. However, given that the 2006 figures were collected through recall, it could be that there is a some recall bias in the 2006 figures.

That said, when we disaggregate 2006 and 2007 enrollment by entry group instead of by age, we find that almost all children in the 2006 entry group report being enrolled in school in the impact evaluation surveys (Table 2). 98.7 percent of children report being enrolled in the 2006 entry group, versus between 94.2 and 95.4 percent among other entry groups. Reported enrollment in 2006 was lower in all four entry groups, though again highest in the 2006 entry group. Therefore, these figures are not terribly suggestive of impacts. However, it is worth noting that there cannot be much heterogeneity in enrollment rates for

the 2006 entry group, since virtually all children are enrolled in school in 2007.

Since the census includes the entire population, we can consider enrollment by age and gender for all of rural El Salvador as well as The census has enough data to support examining enrollment rates by both age and gender (Table 3). Among all children in rural El Salvador 6 to 12 years old, we find similar enrollment rates to those described using the impact evaluation data. Enrollment rates are fairly similar among boys and girls; girls are slightly more likely than boys to enroll in school at all ages. Enrollment rates also increase from 6 year olds, of whom 74 percent enroll, to 10 years old, when 95 percent enroll, and then drop off slightly. By and large, however, this table suggests that primary school enrollment is generally quite high, as suggested by the evaluation data.

From the perspective of the census, an open question is whether or not children in poorer areas are less likely to enroll in school after the poorest municipios have entered CSR. Therefore we next compute enrollment rates by gender for the 2005, 2006, and 2007 CSR entry groups (Table 4). Recall that enrollment is conditional among children in the 2005 and 2006 entry groups if their families are to receive the bono, whereas it is not among children in the 2007 entry groups, and that all children are eligible. Among six year olds, the difference is quite striking. According to the census, 6 year olds are 15 percentage points more likely to be enrolled in school in the 2005 and 2006 entry groups than in the 2007 entry group. This finding is particularly interesting, as parvularia attendance is required to receive the bono among 6 year olds, and enrollment rates are much lower among children who do not reside in the severe EPG. The difference is also large among 7 to 12 year olds, though not quite as large; the raw difference is closer to 7 percentage points between the groups receiving the bono and those that are not. We do not find qualitative differences between boys and girls. These figures are strongly indicative of measurable impacts on school enrollment; we will therefore add to our impact estimation section some information from the census.

In fact, the census figures are indicative of impacts in one other way. When we examine school enrollment rates among all 6 to 12 year olds by EPG, we

find that children in the severe EPG are actually more likely to enroll in school than children in the rest of rural El Salvador, including the moderate and low EPGs (Table 5). Given that we would expect lower enrollment rates among the poorer municipios, it seems likely that CSR has induced some additional enrollment.

In summary, we find that although school enrollment rates among children of age to be in primary school are relatively high, evidence from the census conducted in 2007 is suggestive of impacts of CSR on school enrollment. Furthermore, we do not have to rely on sample estimates of enrollment to make statements about enrollment, as the census figures are the best measures of enrollment rates available for rural El Salvador. We will use the impact evaluation data and census data to construct alternative estimates of impact in the next section.

6 Results

As we have data available on school enrollment in both 2006 and 2007 in the impact evaluation data, we construct an estimator of impacts using difference-in-differences. We primarily use the rectangular kernel and local linear regression to estimate impacts, given that Imbens and Wooldridge (2008) recommend that little is gained by using more complicated non-parametric weighting estimators. Therefore, for difference-in-difference impacts the most general equation that we estimate is:

$$Y_i = \alpha+\beta_1 T_i+\beta_2 G_i+\beta_3 T_i G_i+\beta_4 D_i+\beta_5 T_i D_i+\beta_6 T_i G_i+\beta_7 T_i G_i D_i+\varepsilon_i \; \forall \mid D_i \mid \leq h \tag{8}$$

where T_i references time; G_i references the CSR entry group; and D_i represents the difference in distance between the cluster centers, which is defined as negative when closer to the 2006 entry group cluster center and positive when closer to the 2007 entry group cluster center. The variable h represents the bandwidth, which we vary in estimation. The estimate of impact is β_3, which

estimates the difference in the slope after the program begins (in 2007). The coefficients $beta_4$ through β_7 represent the local slopes with respect to the difference in the distance to the two cluster centers, and is allowed to vary over time and by entry group. When the rectangular kernel is used, then we restrict β_4 through β_7 to be zero.

We initially estimate the impacts of CSR on the enrollment of 7 to 12 year olds in school based on the impact evaluation data. In doing so, we first difference municipio average enrollment rates, and graph them against the difference in distance between cluster centers (Figure 4), along with a linear fit on both sides of the threshold. The figure shows a classic pattern of impact for regression discontinuity. On both sides of the implicit threshold, changes in enrollment rates are rising towards the threshold among CSR transfer recipients, and then drop suddenly at the threshold, and begin to rise again farther from the threshold.

When we estimate the impacts close to the threshold, we find reasonably strong evidence that CSR increased primary school enrollment rates for 7-12 year olds between 2006 and 2007 in municipios entering the program in 2006 (Table 6). Using the rectangular kernel, coefficient estimates increase as we narrow the bandwidth; the point estimate at the most narrow bandwidth suggests an increase in enrollment of 5.2 percentage points at the threshold, and it is significant at the 5 percent level. On the other hand, the coefficients estimated using local linear regression fall somewhat as the bandwidth narrows, though most are statistically significant at the 5 or 10 percent level. At the most narrow bandwidth, the estimate is very consistent with that of the rectangular kernel, at 4.7 percentage points. Using other non-parametric kernels, we also find similar results, so it is safe to conclude that the impact evaluation data suggest an impact of around 5 percentage points at the implicit threshold.

These impact estimates are quite large given that enrollment rates were already above 90 percent prior to program entry. CSR has been very effective in getting the last few children into primary school who had not been enrolled. In fact, findings from the 2006 entry group described here are mirrored in

the 2007 and 2008 entry groups in follow-up surveys; by 2009, enrollment rates among 7 to 12 year olds are above 95 percent for all three entry groups (IFPRI-FUSADES, 2009).

6.1 Impact Results Using the Census

To use the El Salvador census data rather than the impact evaluation data to measure impacts, we must modify the strategy used as the data are cross-sectional and double difference estimation is not possible. That said, the census data have two distinct advantages. First, the census data by definition include all municipios, so there are additional degrees of freedom we can add to the analysis. Second, they are the census data, so estimates generated will be population estimates rather than sample estimates; they are immune to potential problems with outliers. Relatedly, we can break up the population into quite disaggregated gender/age groups to understand among whom the impact is coming.

Since we have only a single time period of data, we have to adjust our estimation strategy slightly. The most general equation we estimate with the census data is:

$$Y_i = \alpha + \beta_1 T_i + \beta_2 D_i + \beta_3 T_i D_i + \varepsilon_i \; \forall \mid D_i \mid \leq h \tag{9}$$

where Y_i again represents school enrollment, D_i is a dummy variable that represents the 2006 entry group, and D_i represents the the difference in distance between cluster centers. In some specifications, we also include a vector of child characteristics, X_i, which include age and gender. The coefficient β_1 now represents the impact of the bono associated with Communidades Solidarias Rurales, and we use both the rectangular kernel and local linear regression to estimate equation (9). The rectangular kernel implies that $\beta_2 = \beta_3 = 0$.

We initially graph average enrollment rates among 7 to 12 year olds at the municipio level for 2007 (Figure 5). The figure is clearly suggestive of impact. The enrollment rates in 2006 are almost all between 90 and 100 percent, and with the exception of one outlier they are tightly distributed. On the right hand

side of the threshold, enrollment rates are much more variable and are clearly lower on average. One outlier is clearly much lower than the other municipios as well. In general, however, this graph is quite consistent with the evaluation data, in that it is highly suggestive of impacts on the net enrollment rate.

Next, we estimate equation (9) using both the rectangular kernel and local linear regression (Table 7). The initial estimate implies that among the entire population of children in the 2006 entry group, school enrollment is 6.9 percentage points higher than in the 2007 entry group. However, this estimate includes a number of municipios that are not very close to the threshold; when we use local linear regression to account for any linear effects the distance from the threshold might have on this impact estimate, the estimate drops to 4.6 percent. Perhaps our best estimate of the impact at the threshold, however, comes at the most narrow bandwidth; both the rectangular kernel and the local linear regression estimates imply a 3.7 percentage point increase in net school enrollment.

The average impact of the bono associated with CSR at the threshold, then, is 3.7 percentage points. Yet there is enough data in the census, by definition, to potentially better isolate whether increased enrollment is taking place among older or younger children, or among boys or girls. Therefore we next investigate graphically and with regressions among which ages increased enrollment is occurring, and if increased enrollment is occurring among boys, girls, or both.

We initially plot average enrollment rates, by age and by the difference in distance between cluster centers (Figure 6). The figure suggests that impacts have occurred among younger age groups, but not as clearly among older age groups. For example, all of the enrollment rates for the 2006 entry group among children who are seven years old are again tightly distributed between 90 and 100 percent, whereas they are quite dispersed and much lower on average for children in the 2007 entry group. Yet for 10 year olds, for example, there is less dispersion in the 2007 entry group, although there are a few outliers. Almost all of the municipios, however, appear to have enrollment rates that are between

90 and 100 percent. Therefore we expect to observe larger impacts on younger children than on older children.

We next estimate the impacts of the bono associated with CSR on school enrollment rates by age and gender, only using the narrowest bandwidth (Table 8). The results are consistent with the illustrations above, but also reveal some interesting gender differences. First, we find that the impact is largest among seven year olds. According to the rectangular kernel results, among seven year olds enrollment is 8.9 percentage points higher in the 2006 group than in the 2007 group due to the bono payment. The coefficient estimates for all children remain significant but decline in magnitude until age 9. At age 10, the coefficient estimates are insignificant among both estimation methods, and among 11 and 12 year olds, they are both only significant for the rectangular kernel, and both are reasonably small in magnitude (around 2.5 percentage points for both). Therefore it is clear that the largest impacts are among younger children.

These results have two main implications. First, they suggest that one of the major impacts of the bono associated with CSR is to enroll children earlier in school than they might have otherwise; the impacts on 7 year olds are quite large, at least in relative terms. Second, they suggest that older girls become slightly more likely either to stay in school or to enter school as a consequence of receiving the bono. The former point is important as it foreshadows lower repeat rates, which have been correlated with receiving the bono and appear to be correlated with earlier school entry, which is a direct consequenece of the program. The latter point is important as it is clear that the experience of boys and girls is different when they are older with respect to CSR.

In summary, we find that impact estimates from two data sources, the data collected for the impact evaluation of CSR and rural households enumerated during the El Salvador census, both indicate similar impacts on school enrollment among 7 to 12 year olds. The impacts of CSR appear concentrated among younger children (aged 7 years) and older girls (aged 11 and 12 years). Next, we investigate whether CSR has also had an effect on the enrollment of children

in parvularia.

6.2 Impact Estimates on Parvularia Attendance

As discussed in the descriptive section, parvularia enrollment rates among 6 year olds, who are of age to attend parvularia, are much higher among children in the 2006 entry group than the 2007 entry group. Using the census data, we find that the difference in enrollment rates is also apparent when we graph municipio averages by the difference in distance between cluster centers (Figure 8). Clearly, school enrollment rates among six years olds are fairly consistently high among the 2006 entry group, whereas they are more variable and generally lower among the 2007 entry group. Graphically, there is fairly clear evidence of an impact on parvularia enrollment.

To more precisely estimate the impacts of CSR on school enrollment among six year olds, we follow the same estimation strategy as for school enrollment (Table 9). Not surprisingly, we find large impact estimates when we simply estimate equation (9) among 6 year olds without restricting the sample; the estimates are 16.9 percentage points at the mean and 15.3 percentage points using local linear regression (column 1). When we restrict the bandwidth to 5, the estimate using the rectangular kernel drops slightly to 14.9 percentage points, but the local linear regression result increases to 19.7 percentage points (column 3). The effect seems to be larger among girls (column 5) than among boys (column 4); whereas the impact appears to be around 16 percentage points among boys, the impact among girls is 23.9 percentage points, using the local linear regression results.

These results imply that even if CSR has not had large impacts on primary school enrollment– in part because large impacts were not possible given enrollment rates prior to CSR were quite high. As a result, a larger proportion of children in poorer municipios will have some experience with school prior to first grade entry, which may imply improved grade progression in the future.

7 Conclusion

In this paper, we first developed conditions necessary to locally estimate program impacts using regression discontinuity, when an explicit forcing variable is not used to determine program eligibility. In the case of CSR, partitioned cluster analysis was used to first assign municipios to specific clusters. We show that under a reasonable set of assumptions, an implicit threshold exists between clusters which can be used as an implicit threshold, with the distance to that threshold serving as a forcing variable. We recommend using the same distance measure that is used in partitioned cluster analysis as the measure of the forcing variable.

We then apply this estimation strategy to estimate the impacts on school enrollment among primary school aged children in rural El Salvador. We use two different data sets, one specifically collected for the impact evaluation as well as El Salvador's census data collected in 2007, to show that impacts on primary school enrollment were between 3.7 and 5.2 percentage points, depending upon the sample and estimation strategy. These impacts are relatively large given that enrollment was already above 90 percent for this age range. That said, because enrollment rates were already high, on one hand it might have been a better use of resources to target middle school enrollment rather than primary school enrollment. According to school censuses collected between 2005 and 2009, in municipios receiving benefits from CSR 9th grade enrollment is 36% lower than 6th grade enrollment had been three years previously, indicating that a large proportion of children drop out of school between 6th grade and 9th grade.

Although one can argue that middle school targeting might have had larger overall impacts, we also find large impacts on school enrollment among 6 year olds, and the majority of 6 year old children are attending parvularia. Given that El Salvador has long had very high first grade repetition, on the order of 22 percent nationwide in the data collected between 2001 and 2004 to generate a poverty map, to the extent that earlier enrollment is negatively correlated with first grade repetition (as argued in IFPRI-FUSADES, 2009), one might expect

that repetition rates will decline as children enter school earlier in El Salvador. The simplicity of targeting in CSR has made it reasonably easy to implement, and more complicated categorical targeting might have made the program more costly to implement well. For example, targeting both parvularia enrollment among 6 year olds and middle school might have confused implementing officials about eligibility of certain children, which is quite easy to determine in the program's current form.

8 References

Imbens and Lemieux 2008

Imbens and Wooldridge 2008

Edmonds, Mammen, Miller 2005

Ludwig and Miller 2007

Porter 2003

Lee and Lemieux 2008

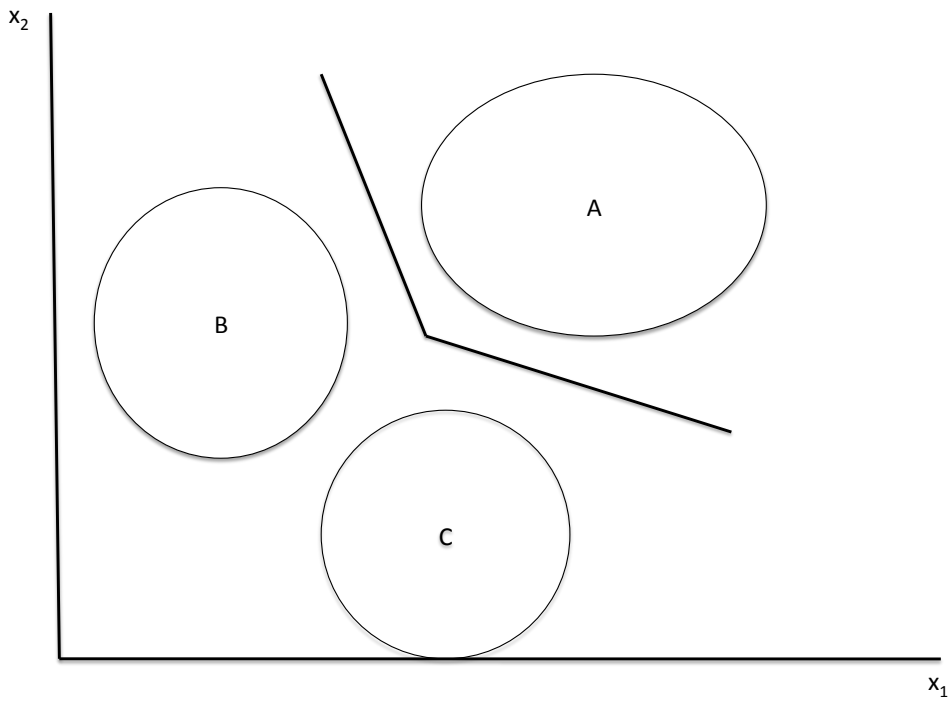

Figure 1: Illustration of Implicit Threshold in Two Dimensional Case

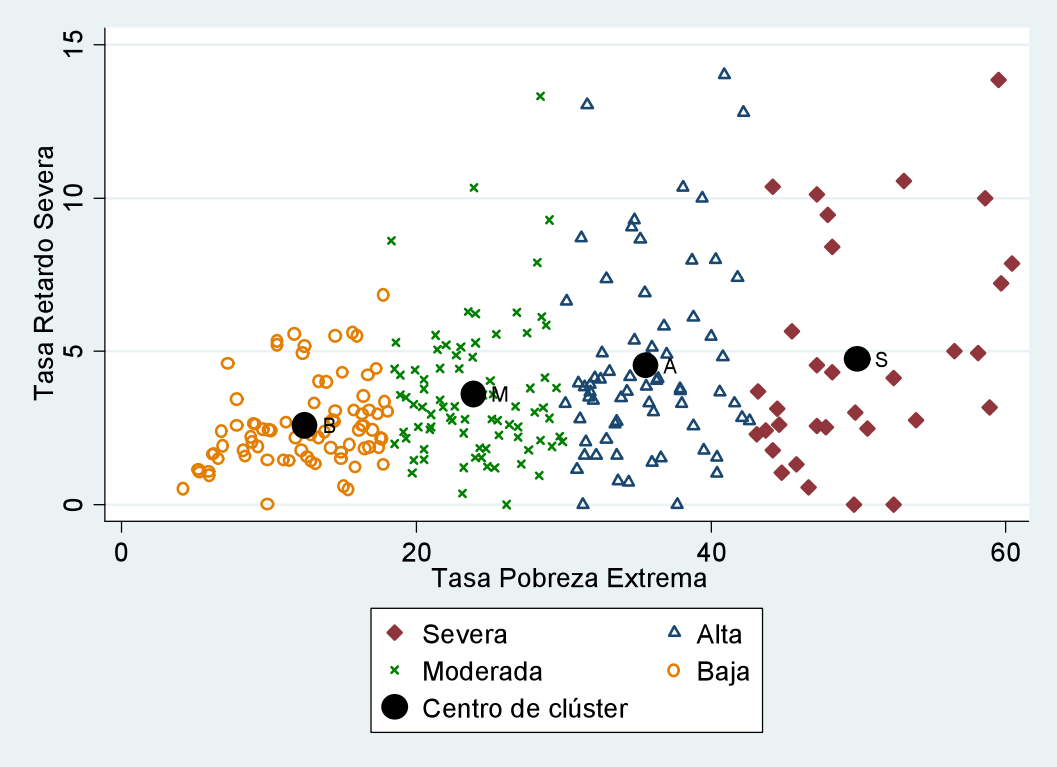

Figure 2. Municipalities by Extreme Poverty Group Clusters with Cluster Centers

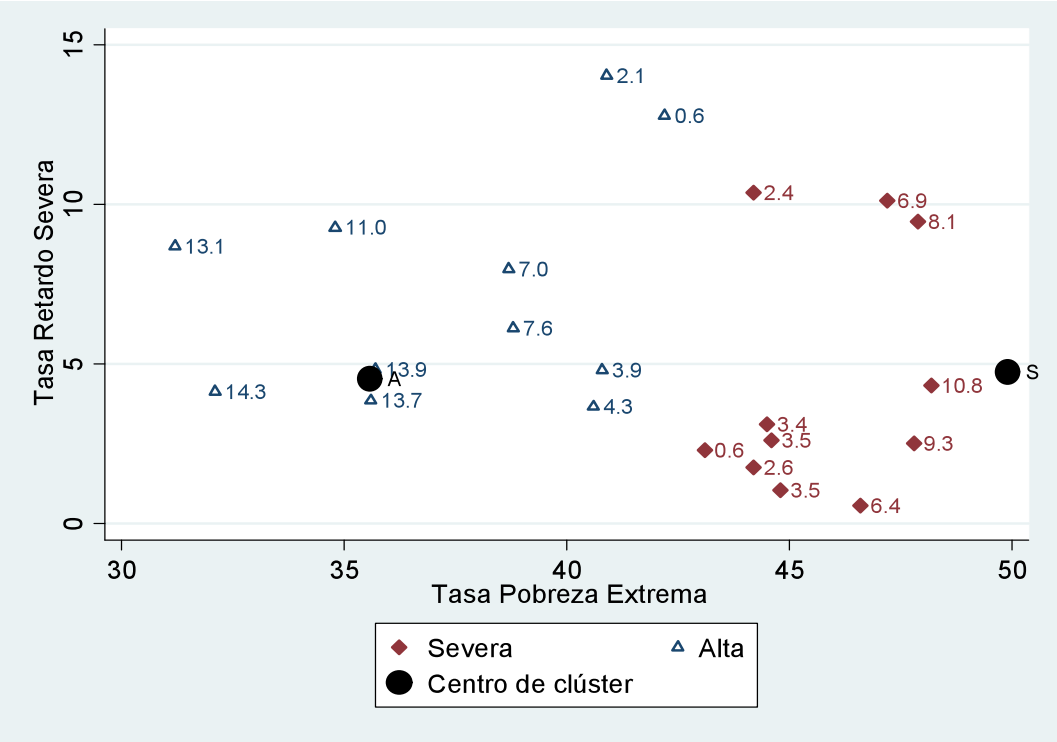

Figure 3. Municipality Distance from Implicit Cluster Threshold for Severe and High Extreme Poverty Group Municipalities, 2006 and 2007 *Red Solidaria* Entrants

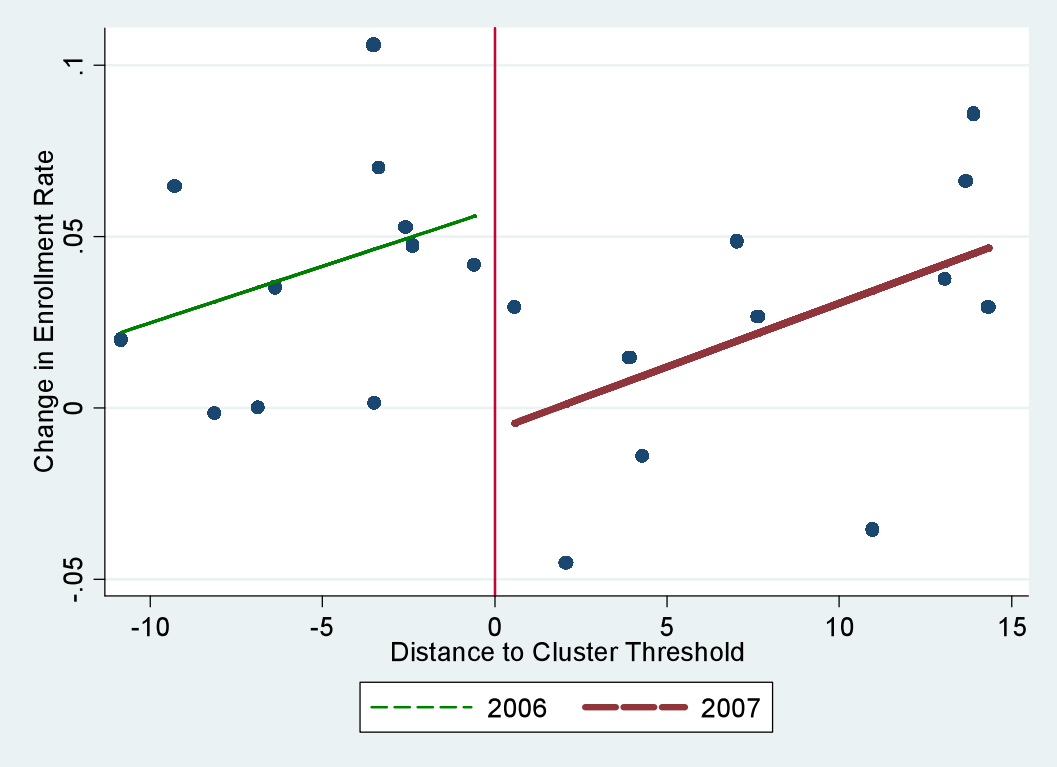

Figure 4. Change in enrollment rate of 7-12 year olds from 2006-2007 by distance from implied cluster threshold, 2006 and 2007 entry groups
Source: Impact Evaluation Survey Data, 2008

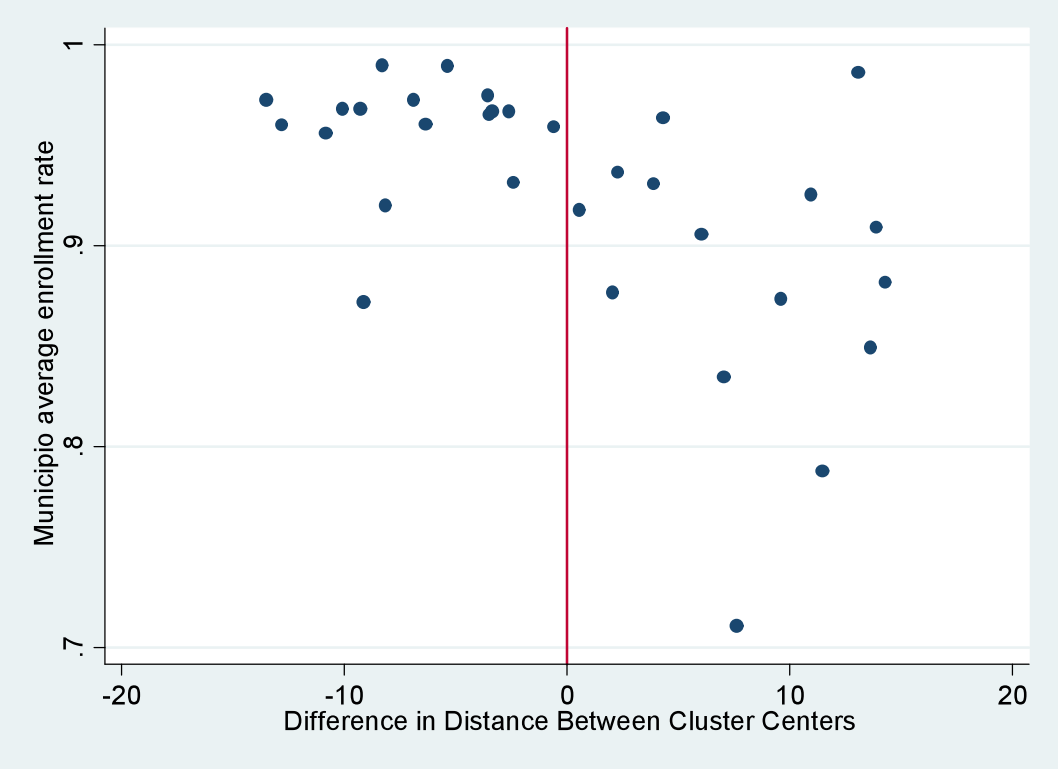

Figure 5. Average Net Enrollment Rates, 7-12 year olds, Municipio Level, Comparing 2006 Entry Group to 2007 Entry Group, El Salvador Census, 2007

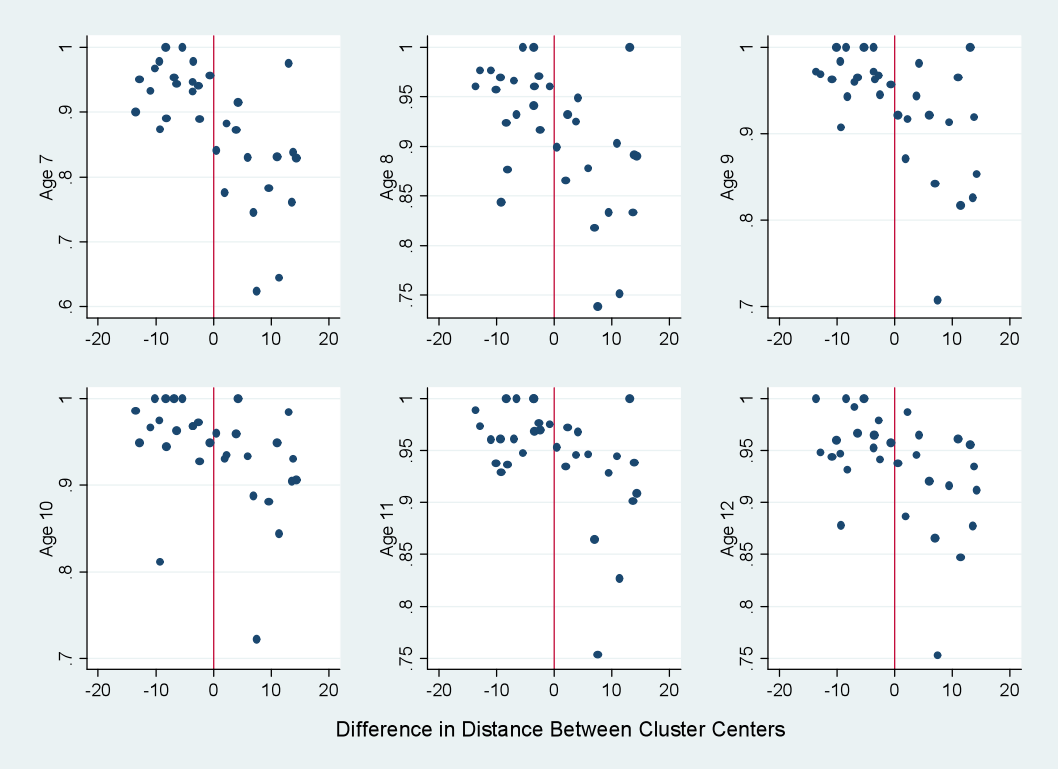

Figure 6. Average Net Enrollment Rates, 7-12 year olds by age, Municipio Level, Comparing 2006 Entry Group to 2007 Entry Group, El Salvador Census, 2007

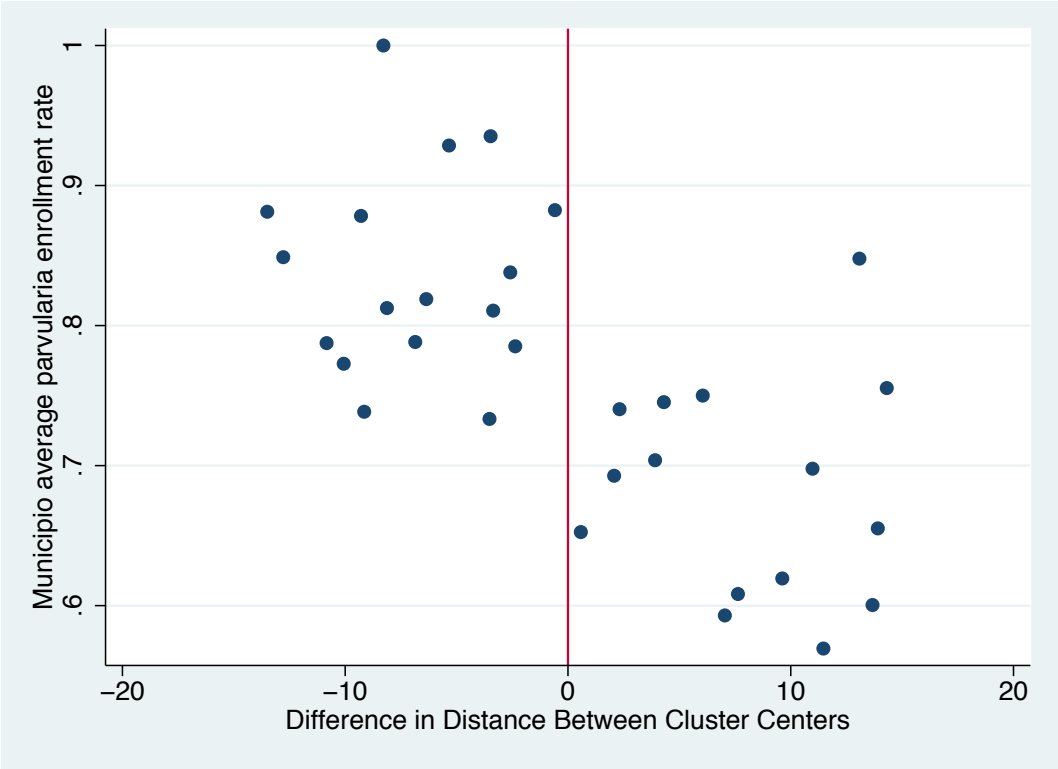

Figure 7. Average Net Enrollment Rates, 6 year olds, Municipio Level, Comparing 2006 Entry Group to 2007 Entry Group, El Salvador Census, 2007

Table 1. School Enrollment Rates by Age, 2006 and 2007, Baseline Impact Evaluation Survey, El Salvador

Age	2006		2007	
	Percent Enrolled	Number of Observations	Percent Enrolled	Number of Observations
7	89.9	614	92.7	628
8	93.9	657	95.8	614
9	96.3	614	97.0	657
10	95.4	542	98.1	614
11	93.6	469	96.3	542
12	91.0	436	92.1	469

Source: IFPRI-FUSADES Impact Evaluation Survey, 2008. Includes all CSR entry groups.

Table 2. School Enrollment, Children aged 7-12 in year previous to survey for baseline and at present for follow-up, by year of entry into Red Solidaria

	Year of Entry into Red Solidaria			
	2006	2007	early 2008	late 2008
Enrolled in 2007	0.987	0.949	0.942	0.954
	(0.004)	(0.011)	(0.009)	(0.010)
Enrolled in 2006	0.963	0.931	0.923	0.934
	(0.008)	(0.014)	(0.012)	(0.013)

Notes: Standard errors in parentheses clustered at the canton level. The averages for 2006 are reweighted to reflect the demographic composition of the children enrolled in 2007.
Source: Impact Evaluation Baseline Survey, 2008

Table 3. Percent of Children Enrolled in School, 2007, Rural El Salvador, by Age and Gender

Age of Child	All Children	Males	Females
6	74.2%	73.2%	75.2%
7	88.0%	87.3%	88.6%
8	92.3%	91.8%	92.8%
9	93.8%	93.3%	94.3%
10	95.0%	94.6%	95.4%
11	94.9%	94.7%	95.1%
12	94.6%	94.5%	94.7%

Source: Censo de El Salvador, 2007.

Table 4. School Enrollment Rates, 2007, Rural El Salvador, by Year of Entry to Red Solidaria and Gender

	All Children		Males		Females	
Year of Entry	6	7 to 12	6	7 to 12	6	7 to 12
2005	80.5%	94.7%	81.6%	94.5%	81.3%	95.0%
2006	82.8%	95.4%	83.6%	95.3%	83.6%	95.5%
2007	65.9%	88.5%	67.2%	88.0%	67.2%	88.8%

Source: Censo de El Salvador, 2007.

Table 5. School Enrollment Rates, 2007, Rural El Salvador, by Poverty Group

Grupo de Pobreza	Enrollment Rate, 6-12 Year Olds	Number of Observations
Severe	92.3%	33183
High	86.7%	111767
All Other	89.4%	550747

Source: Censo de El Salvador, 2007.

Table 6. Regression Discontinuity Results for Impact of Red Solidaria on Change in Enrollment Rates for 7-12 year olds from 2006-2007, Comparing 2006 Entrants to 2007 Entrants

Outcome Variable	Full Sample (1)	Euclidean Distance Bandwidth=10 (2)	Euclidean Distance Bandwidth=8 (3)	Euclidean Distance Bandwidth=5 (4)
Impact on Change in Enrollment, 2006-2007				
OLS Estimation (Rectangular Kernel)	0.015 (0.019)	0.031 (0.018)*	0.030 (0.020)	0.052 (0.023)**
LLR Estimation	0.066 (0.028)**	0.071 (0.027)**	0.082 (0.026)***	0.047 (0.037)
Nonparametric Kernel				
Gaussian	0.020 (0.016)	0.034 (0.020)*	0.035 (0.019)*	0.052 (0.021)**
Epanechnikov	0.030 (0.017)*	0.038 (0.020)*	0.045 (0.020)**	0.054 (0.021)***
Number of Obs.	3239	2306	2105	1534

Notes: Standard errors in parentheses clustered at the municipio level. *-indicates significance at the 10 percent level; **- indicates significance at the 5 percent level; ***- indicates significance at the 1 percent level. For kernel estimates, standard errors are bootstrapped using 100 replications of the data. Bandwidth refers to the distance on either side of the threshold; where a bandwidth is specified, any observations outside the bandwidth are excluded.

Table 7. Impacts of Transfer Associated with Communidades Solidarias Rurales on Net School Enrollment, 7-12 Year Olds, El Salvador Census

	Population	Bandwidth=10	Bandwidth=5
Rectangular Kernel	0.069	0.057	0.037
	(0.016)**	(0.018)**	(0.013)**
Local Linear Regression	0.046	0.029	0.037
	(0.017)**	(0.018)	(0.020)*
Number of Obs.	37065	21662	14441
Municipios	32	22	11

Notes: Standard errors clustered at the municipio level are in parentheses. *- indicates significance at the 10 percent level, and **- indicates significance at the 5 percent level. All regressions include a full set of age and gender dummies.

Table 8. Impact of Comunidades Solidarias Rurales on School Enrollment, by Age and Gender, Bandwidth=5, El Salvador Census Data, 2007

Age of Child	All		Boys		Girls	
	OLS	LLR	OLS	LLR	OLS	LLR
7	0.089	0.098	0.096	0.117	0.081	0.074
	(0.023)**	(0.046)**	(0.026)**	(0.035)**	(0.025)**	(0.067)
8	0.042	0.054	0.037	0.051	0.047	0.058
	(0.015)**	(0.024)**	(0.016)**	(0.026)*	(0.016)**	(0.024)**
9	0.039	0.04	0.033	0.022	0.047	0.06
	(0.013)**	(0.019)**	(0.010)**	(0.014)	(0.017)**	(0.026)**
10	-0.001	-0.019	0.001	0.019	-0.004	-0.048
	(0.012)	(0.021)	(0.021)	(0.028)	(0.017)	(0.023)*
11	0.026	0.015	0.018	-0.03	0.033	0.067
	(0.006)**	(0.011)	(0.009)*	(0.007)	(0.013)**	(0.019)**
12	0.024	0.023	0.014	0	0.034	0.047
	(0.014)*	(0.017)	(0.011)	(0.017)	(0.019)	(0.021)**

Notes: Standard errors clustered at municipio in parentheses. Each cell represents a separate regression. *- indicates significance at the 10 percent level; **- indicates significance at the 5 percent level. Regressions compare individuals in 2006 entry municipios with 2007 entry municipios.

Table 9. Impact of CSR on School Enrollment among children age 6, using Regression Discontinuity

	No Bandwidth	Bandwidth=10	Bandwidth=5		
	All children	All children	All children	Boys	Girls
	(1)	(2)	(3)	(4)	(5)
Rectangular	0.169	0.160	0.148	0.135	0.162
Kernel	(0.019)**	(0.021)**	(0.029)**	(0.038)**	(0.033)**
Local Linear	0.153	0.149	0.197	0.162	0.239
Regression	(0.031)**	(0.041)**	(0.060)**	(0.079)*	(0.048)**
Number of Obs.	6209	3665	2509	1294	1209

Notes: Standard errors clustered at the municipio level in parentheses. *- indicates significance at the 10 percent level; **- indicates significance at the 5 percent level.