NBER WORKING PAPER SERIES

EXTERNAL AND INTERNAL VALIDITY OF A GEOGRAPHIC QUASI-EXPERIMENT EMBEDDED IN CLUSTER-RANDOMIZED EXPERIMENT

Sebastian Galiani
Patrick J. McEwan
Brian Quistorff

Working Paper 22468
http://www.nber.org/papers/w22468

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2016

We are grateful to Matias Cattaneo, Juan Carlos Escanciano, Luke Keele, Rocío Titiunik, the anonymous referees, and participants of the Advances in Econometrics conference at the University of Michigan for their helpful comments, without implicating them for errors or interpretations. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

External and Internal Validity of a Geographic Quasi-Experiment Embedded in Cluster-Randomized Experiment
Sebastian Galiani, Patrick J. McEwan, and Brian Quistorff
NBER Working Paper No. 22468
July 2016
JEL No. O22

ABSTRACT

This paper analyzes a geographic quasi-experiment embedded in a cluster-randomized experiment in Honduras. In the experiment, average treatment effects on school enrollment and child labor were large—especially in the poorest blocks—and could be generalized to a policy-relevant population given the original sample selection criteria. In contrast, the geographic quasi-experiment yielded point estimates that, for two of three dependent variables, were attenuated. A judicious policy analyst without access to the experimental results might have provided misleading advice based on the magnitude of point estimates. We assessed two main explanations for the difference in point estimates, related to external and internal validity.

Sebastian Galiani
Department of Economics
University of Maryland
3105 Tydings Hall
College Park, MD 20742
and NBER
galiani@econ.umd.edu

Patrick J. McEwan
Department of Economics
Wellesley College
106 Central St.
Wellesley, MA 02481
pmcewan@wellesley.edu

Brian Quistorff
Department of Economics
University of Maryland
College Park, MD 20742
bquistorff@gmail.com

1. Introduction

In a typical regression-discontinuity design, treatments are assigned on the basis of a single, continuous covariate and a cutoff. The identification of treatment effects relies on the assumption that the relation between potential outcomes and the assignment variable is continuous at the cutoff (Hahn, Todd, and van der Klaauw, 2001; Lee and Lemieux, 2010). The assumption is particularly credible if a stochastic component in the assignment variable (e.g., a noisy test score) ensures that the agents cannot precisely manipulate their values of the assignment variable. That is, agents are subject to "local" random assignment (Lee, 2008; Lee and Lemieux, 2010).[1]

Researchers have extended the continuity results to regression-discontinuity designs in which assignment is based on a vector of variables (Imbens and Zajonc, 2009; Keele and Titiunik, 2015). A special case is the geographic discontinuity design (GDD), in which exposure to a treatment depends on the latitude and longitude of agents with respect to an administrative or territorial boundary. Researchers compare treated and untreated agents residing near boundaries, using parametric and/or non-parametric methods (e.g., Black, 1999; Dell, 2011; Keele and Titiunik, 2015). In the words of Lee and Lemieux (2010), these are often "nonrandomized" discontinuity designs since agents are usually aware of boundaries (and associated treatments) and can precisely choose their locations.[2] This places a special burden on researchers to rule out location-based sorting on observed or unobserved variables as a threat to the internal validity of treatment effects (Keele and Titiunik, 2015, 2016; Keele et al., 2016).

[1] Cattaneo, Frandsen, and Titiunik (2015) push this interpretation further by implementing randomization inference in samples near the cutoff.

[2] In rare cases, boundaries might be suddenly (and quasi-randomly) redrawn, leading to a more credible "re-randomization" of households before endogenous sorting begins anew (e.g., Billings, Deming, & Rockoff, 2014).

In addition to internal validity, one might assess the external validity of estimates obtained from geographic designs.[3] Suppose that a treatment is (non-randomly) assigned to 10 states, and that policy-makers are interested in the average treatment effect in this population. Yet, for identification purposes, geographic designs must exclude treated individuals that are (1) far from a state border or (2) near a state border with no cross-state variation in the treatment (as often occurs when treated states are contiguous). If excluded individuals have a different distribution of variables that moderate treatment effects—such as income, race, or even distance-to-border—then the geographic design will not recover the average treatment effect in the policy-relevant population. The challenge is well-understood in the context of non-representative convenience samples often used in randomized experiments (Hotz, Imbens, and Mortimer, 2005; Cole and Stuart, 2010; Muller, 2015). We note that it also applies to geographic designs.

This paper assesses both validity concerns in a geographic quasi-experiment (GQE) that is embedded in a cluster-randomized experiment conducted in Honduras.[4] In the original experiment, 70 malnourished municipalities were identified, and 40 were randomly awarded conditional cash transfers (IFPRI, 2000). Using the 2001 census, Galiani and McEwan (2013) found that the treatment increased the probability that children enrolled in school and reduced their probabilities of working outside and inside the home. The effects were especially large in

[3] External validity exists when causal relationships "[hold] over variations in persons, settings, treatment variables, and measurement variables" (Shadish, Cook, and Campbell, 2002, p. 507). Some authors assess the importance of variation in treatments (particularly implementer characteristics) and measurement variables (Allcott, 2015; Bold et al., 2012; Lucas et al., 2014). This paper focuses on the potentially moderating role of independent variables related to persons and settings.

[4] We refer to it as a geographic quasi-experiment because assignment is not transparently random (even local to borders), and because we must proxy the location of children using the coordinates of their caserío (a cluster of dwellings). The latter introduces mass points in the putative assignment variables of latitude and longitude in a geographic discontinuity design. For related explanations, see Keele et al. (2016) and later sections of this paper.

two of five strata (or blocks) with the highest rates of malnutrition. In this subsample, the effects on enrollment, work outside the home, and work inside the home were, respectively, 15, -6.8, and -6.2 percentage points. Relative to the control group, these represented changes of 25%, -54%, and -41%.

We compare these results to those of a geographic quasi-experiment using the same census data. Specifically, we identify a sample of treated children that are close to municipal borders shared with untreated, non-experimental municipalities. Children on the opposite side of the border constitute the quasi-experimental control group. Using the same covariates as Galiani and McEwan (2013), we show that treatment-control balance for nearly all covariates improves in samples that are increasingly close to the border (our preferred distance buffer is 2 kilometers). We can also rule out that households sorted across municipal borders in direct response to the treatment, addressing a common internal validity concern in geographic designs. Nevertheless, we find that treated children are more likely to self-identify as Lenca—an indigenous group—even very close to municipal borders. Ultimately, our analysis of the GQE sample finds that point estimates for two of the three dependent variables are attenuated relative to the experimental benchmarks.

Is this because of imbalance in unobserved variables (i.e., a threat to internal validity) or simply because the GQE sample has a different distribution of observed or unobserved variables that moderate treatment effects? We separately assess each explanation using subsamples of the randomized experiment. First, we re-estimate experimental effects in the subsample of treatment and control children within 2 kilometers of *any* municipal border (we refer to this as experiment 1). These estimates are slightly (but consistently) *stronger* than full-sample results. There are no

mean differences in census covariates between the two samples, suggesting that distance-to-border proxies unobserved moderators of treatment effects.

Second, we further restrict the sample in experiment 1 to treatment and control children residing near the border of an untreated, non-experimental municipality (we refer to this sample as experiment 2). Note that it includes exactly the same treated children as the GQE. However, it uses experimental rather than the quasi-experimental controls. This permits us to (momentarily) abstract from internal validity. The point estimates for school enrollment and work-in-home are attenuated relative to experiment 1. Descriptive statistics suggests that the sample in experiment 2 is better-off than that of experiment 1, given higher rates of electricity use, asset ownership, and other income proxies. This plausibly explains the attenuated effects, since the literature on conditional cash transfers finds smaller effects in when children are less poor (Fiszbein and Schady, 2009; Galiani and McEwan, 2013).

Third, we assess internal validity by comparing the unbiased estimates from experiment 2 to those of the GQE (noting again that both include the same treatment group but different control groups). Particularly for school enrollment, the GQE estimates are attenuated relative those of experiment 2. It suggests that imbalance in unobserved variables results in downward biases in the GQE enrollment estimates. This is perhaps consistent with the higher proportion of indigenous children in the GQE treatment group, relative to its quasi-experimental control group.

In summary, we find that the GQE cannot fully replicate the policy-relevant experimental benchmark in Galiani and McEwan (2013) for reasons related to both validity concerns. Based on these results, we make two concrete recommendations. First, it is essential that researchers using a geographic design carefully assess treatment-control balance on a wide range of observed covariates that are plausibly correlated with dependent variables (echoing the recommendations

of Keele et al., 2016 in this volume). Our GQE is an especially cautionary tale, since it had very good (but not perfect) balance in observed variables, but still could not replicate school enrollment estimates using the same treatment group and an experimental control group.

Second, we recommend that researchers assess the external validity of their geographic design by comparing the distributions of observed moderators of treatment effects—such as household income—to those of a well-defined, policy-relevant population. Aided by theory or prior empirical evidence on the relevance of moderators, this can be used to speculate about the generalizability of a GQE. More concretely, Cole and Stuart (2010) describe how one might construct inverse-probability weights and re-weight a convenience sample—whether experimental or quasi-experimental—to resemble a well-defined population. We conduct and report a similar analysis, re-weighting the sample of experiment 2 to resemble that of experiment 1. The weights are estimated using a wide range of covariates that are plausible moderators of treatment effects. Ultimately, however, the weighted estimates in experiment 2 are still attenuated relative to experiment 1, suggesting that some relevant moderators are unobserved.

Our results contribute to a growing literature that compares regression-discontinuity designs with a single assignment variable to experimental benchmarks (see Shadish et al., 2011 and the citations therein). In this literature, several papers analyze conditional cash transfer experiments in which eligibility was determined by a poverty proxy. Oosterbeek, Ponce, and Schady (2008) found that experimental enrollment effects in Ecuador were large for the poorest households, but that RDD effects were zero for less-poor households in the vicinity of the eligibility cutoff. Similarly, Galiani and McEwan (2013) found no RDD effects on enrollment and child labor in the vicinity of the cutoff used to determine assignment to the experimental sample, but large experimental effects among households residing in municipalities with the lowest levels of the

assignment variable. In the absence of an experiment, both papers caution against generalizing "away" from cutoffs when the assignment variable is a plausible or well-documented moderator of treatment effects.[5] A recent strand of methodological literature has considered situations in which such generalizations might still be possible.[6]

2. The PRAF-II Experiment

A. Design and treatment

In the late 1990s, the International Food Policy Research Institute (IFPRI) designed a cluster-randomized experiment to estimate the impact of conditional cash transfers (CCTs) on the poverty, education, and health outcomes of households in poor Honduran municipalities. In the absence of a national poverty map, researchers identified poor municipalities with a nutrition-related proxy from a 1997 census of first-graders' heights (Secretaría de Educación, 1997). IFPRI (2000) ordered 298 municipalities by their mean municipal height-for-age z-scores. Seventy-three municipalities with the lowest scores were eligible (the implied cutoff was -2.3, highlighting the extremely high rates of stunting). Three were excluded due to accessibility, leaving an experimental sample of 70.

In 1999, IFPRI divided the sample into 5 quintiles of mean municipal height-for-age. Within quintiles, municipalities were randomly assigned to three treatment arms and a control group (in

[5] Buddelmeyer and Skoufias (2004) analyzed Mexico's well-known Progresa experiment (and a proxy means test and cutoff used to determine eligibility). In contrast to other results, they found that experimental enrollment estimates in samples "close" to the eligibility cutoff were roughly similar or slightly larger than full-sample estimates.

[6] Angrist and Rokkanen (2015) show how RDD effects might be estimated "away" from the cutoff if the assignment variable is ignorable, conditional on a set of covariates unaffected by the treatment. Dong and Lewbel (2015) note that the relative slopes of lines fit to data within bandwidths on either side of the cutoff provide insights into how modest changes in the assignment cutoff could affect the magnitude of estimated effects.

a ratio of 4:4:2:4). Arms 1 and 2 received conditional cash transfers, while arms 2 and 3 were to receive grants to schools and health centers. Moore (2008) suggests the grants were sparsely implemented as late as 2002. Using this paper's census data, Galiani and McEwan (2013) failed to reject the null that average treatment effects in arms 1 and 2 were equal (relative to the control). Arm 3 had small and statistically insignificant effects relative to the control (but its effect was statistically different from arm 2). Following Galiani and McEwan (2013), we compare 40 municipalities in a pooled CCT treatment arm and 30 in a pooled control group.

In the CCT treatment, households were eligible for an annual per-child cash transfer of L 800 (about US$50) if a child between 6 and 12 enrolled in primary school grades 1 to 4.[7] Children with higher attainment were not eligible, and households could receive up to 3 per-child transfers. During the experiment, transfers were distributed in November 2000, May-June 2001, October 2001, and late 2002 (Galiani and McEwan, 2013; Morris et al., 2004). The average household in experimental municipalities would have been eligible for transfers equal to about 5% of median per capita expenditure (Galiani and McEwan, 2013). This is smaller than most Latin American CCTs such as Progresa (Fiszbein and Schady, 2009). Indeed, payments were only intended to cover the out-of-pocket and opportunity costs of enrolling a child in school (IFPRI, 2000).

[7] A school attendance condition was apparently not enforced (Glewwe and Olinto, 2004).

B. Data and replication

Galiani and McEwan (2013) used the 2001 Honduran census—collected in late July 2001—to estimate the short-run effects of offering CCTs to eligible children.[8] Their sample contained 120,411 6-12 year-olds eligible for the education transfer, residing in the 70 experimental municipalities. The census includes three dummy dependent variables: (1) whether a child was enrolled on the census date, (2) whether a child worked outside the home in the week preceding the census, and (3) whether a child worked exclusively in the home during the preceding week (Appendix Table A.1 provides variable definitions).[9]

Their preferred specification regressed each dependent variable on a treatment dummy, dummy variables indicating randomization blocks, and a set of individual and household covariates unlikely to have been affected by the treatment. Table 1 replicates the main results. The regressions in this and subsequent tables control for block dummy variables, the 21 covariates described in Table A.1, squared terms for continuous variables, and dummy variables indicating missing values of any covariate. In addition to analytic standard errors clustered by municipality, we report symmetric p-values from the wild cluster bootstrap percentile-t that imposes the null hypothesis (Cameron, Gelbach, and Miller, 2008).

In the full sample, the treatment increases the probability of enrollment by 8.1 percentage points (a 13% increase relative to the control group). The treatment reduces the probability of work outside the home by 3.1 p.p. (32%) and work only inside the home by 4.1 p.p. (30%). The effects are larger in the two poorest blocks (1 and 2), and closer to zero and not statistically

[8] A related literature uses a panel household survey—collected in 2000 and 2002—to estimate effects on child health and nutrition (Morris et al., 2004), education (Glewwe and Olinto, 2004), and adult labor supply (Alzúa et al., 2013).
[9] The interpretation of work-only-inside-home variable is governed by the flow of survey questions.

significant in three less-poor blocks (3, 4, and 5). In blocks 1 and 2, the effect on enrollment, work outside the home, and work inside the home are, respectively, 15 p.p. (25%), -6.8 p.p. (-54%), and -6.2 p.p. (-41%). The magnitude of these effects is notable given the comparatively small size of the transfer.[10]

3. A Geographic Quasi-Experiment

A. Sample

The Honduran census does not record the precise location of dwellings. To proxy location, we use the latitude and longitude of caseríos. In Honduras, 18 departments contain 298 municipalities and over 3700 aldeas (villages). Within the boundaries of villages, points identify the center of over 24,000 caseríos ("hamlets") that are contiguous groups of dwellings. We calculated the straight-line distance between each caserío and its nearest municipal border.[11]

We then identified a sample of 801 experimental, treated caseríos (with 23,974 children) that share a municipal border with 794 non-experimental, untreated caseríos (with 25,025 children). In the pooled sample of children, the 10^{th}, 50^{th}, and 90^{th} percentiles of the distance-to-border distribution are 0.37, 2.02, and 4.22 kilometers, respectively. We henceforth refer to this as the GQE (or geographic quasi-experiment) sample.

[10] Benedetti, Ibarrarán, and McEwan (2016) analyze a later Honduran CCT experiment—also conducted in a sample of poor municipalities—which offered much larger transfers. They found smaller effects on both enrollment and child labor, which they attributed to a weaker application of the education enrollment condition.

[11] We identified the caserío coordinates for 93% of all 6-12 year-olds in the census (and 95% of the full experimental sample). The missing coordinates are due to incomplete geocoding of caseríos in an ArcGIS file obtained from the Infotecnología unit of the Secretaría de Educación in 2008.

The map in Figure 1 illustrates the subsample of GQE caseríos that fall within 2 kilometers of a municipal border (in the next section, we provide a rationale for using this distance buffer). It highlights that treated caseríos are a non-random sample of all treated caseríos. In particular, treated caseríos are excluded when their municipalities are fully circumscribed by other treatment or control municipalities in the experimental sample.

B. Covariate balance near municipal borders

By focusing on treated and untreated children that reside near municipal borders, there may be fewer differences in observed and, perhaps, unobserved variables that affect child outcomes. We assess this in the top-left panel of Figure 2, using 21 covariates from the experimental analysis. Dots indicate the absolute values of standardized treatment-control differences for each covariate. The left-most dots pertain to the full GQE sample just described, while others refer to GQE samples restricted by increasingly narrow distance buffers.

Balance in the GQE sample is sensitive to distance-to-border, markedly improving when caseríos fall within 2 kilometers of the border. In the 2-kilometer subsample, it is notable that 20 of 21 covariates show treatment-control differences of less than 8% of a standard deviation, and none are statistically different from zero at conventional levels (see Table 2). This is despite the fact that there are large mechanical differences in mean municipal height-for-age z-scores due to the selection rule for the experimental sample (see Figure 3.) The bottom-left panel of Figure 2 shows F-statistics from omnibus tests of covariate balance.[12] In each subsample, we regress the treatment dummy on the 21 covariates, squared terms for continuous variables, and dummies

[12] A simulation in Hansen and Bowers (2008) shows that a similar test using logistic regression leads to over-rejection of the null with a modest number of assigned units (100). In the present case, we are concerned with comparing balance across subsamples.

indicating missing values. The F-statistic declines as distance-to-border restrictions are applied, consistent with prior evidence.

In the right-hand panels of Figure 2, we can compare balance in GQE samples to balance in experimental samples with similar distance-to-border restrictions. As anticipated, given randomized assignment, covariate balance in the experiment does not depend on the distance of caseríos to municipal borders. The top-right panel shows that absolute values of treatment-control differences for 21 covariates are rarely larger than 10% of a standard deviation, regardless of distance. The bottom-right panel of Figure 2 shows relatively stable F-statistics of around 2 from the omnibus test.

In fact, the smallest F-statistics in the GQE are approximately twice as large as those in the experimental sample. In the 2 kilometer GQE sample, this is driven by imbalance in a single covariate (Lenca). Children on the treated side of borders are about 10 percentage points more likely to self-identify as a member of the indigenous Lenca group (see Table 2). National poverty headcounts are higher among Lenca than non-indigenous Hondurans.[13] Nevertheless, this does not necessarily imply imbalance in unobserved socioeconomic outcomes such as poverty, because there is demonstrable balance in many poverty proxies such as schooling and household assets.

C. Potential threats to internal validity

Why might imbalance persist close to borders? One explanation is that Lenca households manipulated treatment status by moving after experimental assignment, but prior to the census.

[13] In a national sample from 2004, the poverty headcount is 49% among non-indigenous individuals and 71% among ethnic and racial minorities (World Bank, 2006).

We regard this as unlikely for three reasons. First, the cash transfer for a typical household was extremely small (less than 5% of a household's consumption) and unlikely to provide sufficient liquidity for poor households to move. Second, 91% of children in the 2-kilometer GQE sample were born in their municipality of residence, and only 4% lived in a different caserío, aldea, or city in 1996 (five years before the census date). Both variables are among those with the smallest cross-border differences in the GQE, and neither is statistically different from zero (see Table 2).[14] Third, there is a similar pattern of imbalance in the 2-kilometer sample of untreated, experimental caseríos that share a border with untreated, non-experimental caseríos (see Table B.2; we later use this sample to conduct a placebo test). In other words, imbalance persists even when there is no CCT treatment to create incentives for cross-border sorting.

Thus, a second explanation is that some households chose dwellings years or decades before treatment assignment, but were not indifferent to attributes of communities on opposite sides of the border. Figure 4 illustrates the proportion of Lenca children in aldeas (a sub-territory of municipalities), along with the GQE caseríos. Cross-border imbalance is most evident on the eastern-most municipal borders of the experimental sample, but is certainly not a feature of all border segments. Most notably, one interior segment cleaves the "twin cities" of La Esperanza and Intibucá, both rich centers of Lenca culture. Though in separately-governed municipalities, the towns are commonly referred to by a single name (and treated as such by locals).

In summary, there is evidence of balance in the GQE sample (with a buffer of 2 kilometers) on 20 of 21 covariates that are typically correlated with child education and labor outcomes. From a design perspective, it is notable that covariate selection was imposed by an earlier paper

[14] It is possible that households somehow misreported their answers, but seems unlikely given the fact that census data collection was independent of PRAF, IFPRI, and the original impact evaluation's data collection schedule.

(Galiani and McEwan, 2013). However, there is persistent imbalance in one covariate (Lenca) that is plausibly correlated with unobserved determinants of child outcomes. In this volume, Keele et al. (2016) report similar covariate imbalances close to borders, leading them to invoke an assumption of conditional geographic mean independence (also see Keele et al., 2015). That is, potential outcomes are assumed to be mean independent of treatment assignment within a specified buffer, conditional on observed covariates.[15] We make a similar ignorability assumption within a 2-kilometer buffer and refer to the design as a geographic quasi-experiment.

We do not assume local continuity in potential outcomes at municipal borders, as one might in a "pure" geographic discontinuity design (Imbens and Zajonc, 2009; Keele and Titiunik, 2015). First, our evidence suggests that assignment is not locally randomized, given long-standing municipal borders and households' ability to sort around them. Second, we are forced to proxy the location of dwellings using the latitude and longitude of caseríos. This leads to mass points in the putative assignment variables,[16] even though standard methods for analyzing discontinuity designs rely on continuous assignment variables (Calonico et al., 2014; Keele et al., 2016).

D. GQE estimates

[15] The standardized differences are within thresholds beyond which regression adjustment is particularly sensitive to specification (Imbens and Wooldridge, 2009; Rubin, 2001).

[16] In a histogram of distance-to-border distribution in the GQE sample—available from the authors—there is a puzzlingly large spike on the untreated side of the border, between 3 and 4 kilometers away. In fact, this is the town of Santa Bárbara (identified as a single caserío in Honduran data). It stretches about 2 kilometers across at its widest point and contains 1178 dwellings with eligible children. If distance-to-border had been measured without error for each dwelling, it might have "filled" an apparent notch in the histogram. This is the most severe example in the GQE sample of mis-measurement of the assignment variable, given the use of caserío rather than dwelling location. In the full GQE sample of caseríos, the mean (median) number of dwellings is 16.6 (8), and the 90th and 95th percentiles are 33 and 47, respectively.

Using the 2-kilometer GQE sample, Table 3 reports estimates from ordinary least squares regressions that control for the same covariates as Table 1, but excluding experimental block dummy variables. The analytic standard errors apply multi-way clustering on both municipality and border segments, given the spatial proximity of treatment and control caseríos (Cameron, Gelbach, and Miller, 2011). A separate border segments exists for every unique combination of bordering municipalities. Overall, there are 81 municipalities and 65 non-nested border segments.[17] As in Table 1, we also report symmetric p-values from the wild cluster bootstrap percentile-t that imposes the null hypothesis, clustering by municipality (Cameron, Gelbach, and Miller, 2008).

In the 2-kilometer GQE sample, the treatment increases enrollment by a marginally significant 5.7 percentage points, or 2.4 p.p. smaller than the experimental estimate in Table 1. In blocks 1-2, the enrollment effect is 10.6 p.p., or 4.4 p.p. lower than the experimental estimate. In blocks 3-5, both the GQE and the experiment find similarly small and statistically insignificant estimates.

The results are mixed for the two work-related variables. In the GQE sample, the treatment reduces work outside the home by 2.4 p.p. in all blocks and 8.6 p.p. in blocks 1-2 (only the latter is marginally statistically significant). The experimental estimates are roughly similar. Neither the GQE nor the experiment suggest any effects in blocks 3-5. In contrast, the GQE estimates are attenuated for work inside the home, relative to the experimental estimates. There is never a large or statistically significant effect for this dependent variable in the GQE sample. Yet, in the experiment, there were reductions of 4.1 and 6.2 p.p., respectively, in all blocks and blocks 1-2.

[17] When GQE estimates are reported within subsamples of blocks 1-2 and 3-5, control observations in untreated, non-experimental municipalities are assigned to the block corresponding to the treated observations on the opposite side of the border segment.

Despite these differences between the GQE and the experiment, bootstrapped p-values in Table 3 suggest that the estimates are not statistically distinguishable from one another. Even so, one can pose a practical question: would a reasonable policy analyst—relying on the GQE point estimates and blinded to the experimental ones—have reached conclusions as optimistic as those of Galiani and McEwan (2013)? Most likely the attenuated GQE estimates would have yielded more guarded conclusions.

4. Empirical Strategy

A. External and internal validity of the GQE

We consider two explanations for the divergence in point estimates of the GQE and the experiment, related to external and internal validity. Recall that treated caseríos in the GQE are a non-random subset of all treated caseríos. First, they are close to municipal borders. Second, they share a border with untreated, non-experimental caseríos. This naturally excludes caseríos in the spatial core of the experimental sample. As Figures 3 and 4 suggest, excluded caseríos might exhibit higher rates of child stunting or greater concentrations of indigenous children.

Thus, treated children in the GQE are plausibly different in variables observed by the econometrician—such as distance-to-border, height-for-age, and ethnicity—and perhaps in unobserved variables, such as income. If these variables moderate treatment effects, then the GQE estimates—even internally valid ones—will differ from the experimental benchmarks in Table 1. In the present application, it is plausible that GQE caseríos are "less poor" than other treated ones. Since treatment effects are much larger in the poorest municipalities (see Table 1), this provides a plausible explanation for the attenuated GQE treatment effects.

An alternative explanation for the divergence of point estimates is related to internal validity. Suppose that treated children in the GQE differ in unobserved ways from their bordering control group, even after conditioning on a rich set of covariates (e.g., they are more likely to be poor, an unmeasured variable). This too could explain attenuated treatment effects, assuming that poorer children are less likely to enroll in school and more likely to work.

B. Experimental samples used to assess validity

Table 4 summarizes the experimental samples that we use to assess the external and, then, the internal validity of the GQE. The full experimental sample was already used, in Table 1, to obtain estimates of the average treatment effect (ATE). Given the design of the experiment, the estimates are generalizable to a well-defined, policy-relevant population of Honduran children residing in malnourished municipalities.

We next limit the experimental sample to children residing in caseríos no more than 2 kilometers from *any* municipal border. This sample—denoted experiment 1—is used to obtain estimates of ATE_1. If distance-to-border does not moderate treatment effects, then ATE and ATE_1 (and estimates thereof) should be similar.

We further limit the sample of experiment 1 to children residing in caseríos no more than 2 kilometers from a municipal border shared with untreated, non-experimental caseríos. This sample—denoted experiment 2—includes exactly the same sample of treated caseríos (and children) as the GQE. However, its control group consists of the experimental control group subject to the same sample restriction (illustrated in Figure 5). By using an experimental control group instead of a quasi-experimental one, we can abstract from the internal validity of the GQE and focus on external validity. If observed and unobserved moderators are similarly distributed

across the samples of experiments 1 and 2, then ATE_1 and ATE_2 (and estimates thereof) should be similar.

If they differ, then it weakens the GQE's claim on external validity. More constructively, one can further diagnose whether the samples of experiment 1 and 2 differ in their distributions of observed moderators of treatment effects. If they do, then one can re-weight the sample in experiment 2 to resemble that of experiment 1, and re-estimate effects (Cole and Stuart, 2010). To the extent that relevant moderators are observed and contribute to the estimation of the weights, then weighted estimates should be similar to estimates of ATE_1. If they still diverge, then it suggests that a relevant moderator is unobserved. The next subsection will further elaborate the assumptions and method.

Lastly, we assess the internal validity of the GQE in two ways. We first compare GQE estimates from Table 3 to those of ATE_2, which uses an experimental rather than quasi-experimental control group for the same sample of treated children. Any divergence is indicative of bias in the GQE. Second, we implement the placebo test alluded to in an earlier section. We construct a placebo sample of untreated, experimental caseríos no more than 2 kilometers from a municipal border shared with untreated, non-experimental caseríos. We anticipate finding zero effects in the "GQE" placebo sample, conditional on covariates. A positive or negative effect is likely the result of imbalance in unobserved variables.

C. Inverse-probability weighting and external validity

Using the potential outcomes framework, let the outcome Y for individual i be a function of a randomly-assigned treatment T_i. The difference in potential outcomes under treated (1) and

untreated (0) conditions is $Y_i(1) - Y_i(0)$. Table 1 reported estimates of the average treatment effect (ATE) in a well-defined, policy-relevant sample, denoted S.

We will further report estimates of ATE_2 in the non-random subsample of experiment 2, denoted S_2. If the treatment has heterogeneous effects on individuals, moderated by a set of variables X, then effects may differ across samples depending on the distribution of X. An intuitive method of correcting for this difference is to re-weight S_2 so that its distribution of moderators is similar to S (Cole and Stuart, 2010; Stuart et al., 2015).

Closely following Hotz, Imbens, and Mortimer (2005), we specify three assumptions under which the procedure can recover ATE from S_2.[18] Assumption 1 is that the treatment is randomly assigned in S_2:

$$T_i \perp [Y_i(1), Y_i(0)]|S_2.$$

The assumption is satisfied because there was random assignment in S, and S_2 is a subsample of treatment and control groups obtained by imposing exogenous sample restrictions.

Assumption 2 asserts that one's presence in the subsample does not depend on potential outcomes, given the moderators:

$$(i \in S_2\ ) \perp [Y_i(1), Y_i(0)]|X_i.$$

Hotz et al. (2005) refer to this as unconfounded location. Stuart et al. (2011) invoke a similar assumption and refer to it as unconfounded sample selection. Both papers highlight the need to measure all relevant moderators in order to satisfy the assumption. Lastly, Assumption 3 imposes a requirement of common support for the moderators between the two samples. For each moderator, it must be the case that:

[18] Given the assumptions described below, Hotz et al. (2005) show that $ATE = E[Y_i(1) - Y_i(0)|S] = E\{E[Y_i|T_i = 1, S_2, X_i] - E[Y_i|T_i = 0, S_2, X_i]|S\}$.

$$0 < P(i \in S_2 | X_i) < 1.$$

In the present context, assumption 3 does not hold for one moderator, namely distance-to-border (since observations more than 2 kilometers from the border have zero probability of contributing to S_2). However, we can reframe the task as generalizing from the sample S_2 to S_1, which also imposes the 2-kilometer distance restriction (Stuart et al., 2011). Common support holds for all other covariates.

Assumption 2 may not hold if there are unobserved moderators. A typical study cannot verify this, just as a typical observational study cannot directly test for selection-on-unobservables into a treatment or control group. In contrast, we can compare weighted estimates in experiment 2 to estimates of ATE_1. Any difference suggests that relevant moderators are unobserved.

To implement the method, we estimate inverse probability weights (Cole and Stuart, 2010; Stuart et al., 2015).[19] In S_1—which imposes the 2-kilometer distance buffer—we estimate a logit regression in which the dependent variable indicates observations in S_2. The regressors include 21 covariates, 6 squared terms, and dummy variables indicating missing values.[20] They further include block dummy variables, mean municipal height-for-age, distance-to-border, and squared terms for the latter two. Lastly, we calculate inverse probability weights for observations in S_2 as $w_i(X_i) = 1/\hat{p}(X_i)$, where $\hat{p}$ is the estimated probability that an observation is selected for S_2.

5. Results

A. External validity: experiment 1

[19] Cole and Stuart (2011) prove that the method yields consistent estimates—in this paper, of ATE_1—under assumptions similar to those just described.
[20] When weighted estimates are reported in subsamples (e.g., blocks 1-2), we separately estimate weights in that subsample.

Tables 5 to 7 reports results for the three dependent variables. Experiment 1 includes children residing in caseríos within 2 kilometers of *any* municipal border. Imposing this restriction slightly increases the positive enrollment estimates and makes the work-related estimates slightly more negative. In Table 4, for example, the enrollment estimate is 8.9 percentage points inside the buffer (versus 8.1 in Table 1). Further limiting the sample to blocks 1 and 2, the estimate in experiment 1 is 16.3 p.p. (versus 15 p.p. in Table 1). There is no ready explanation for the slight increases in enrollment effects in Table 5 (and slightly more negative work effects reported in Tables 6 and 7). The mean covariate differences between the full experimental sample and experiment 1 are small and statistically insignificant (full results are available from the authors). This suggests that distance-to-border is a proxy for other, unobserved moderators.

B. External validity: experiment 2

Tables 5 to 7 also report estimates for experiment 2. Recall that it includes the same treated observations as the GQE sample, but with an experimental control group. Imposing this sample restriction reduces the enrollment estimates by 2.6 p.p. relative to experiment 1, apparently driven by a 3 p.p. decline in the blocks 1-2 subsample. A similar pattern of attenuation is evident for work-in-home estimates (Table 7), but not for work-outside-home (Table 6). For the latter variable, the coefficient in the blocks 1-2 sample is slightly *more* negative.

Did the sample restriction in experiment 2 change the distribution of plausible moderating variables? In fact, observations in the experiment 2 sample are 8 percentage points less likely to belong to blocks 1-2 (see Appendix Table C.1). There are also substantial differences for specific covariates, especially within blocks 1-2 (see Figure 6). For example, households of children in experiment 2 are 12 p.p. more likely to have electric light, 10 p.p. more likely to own a

television, and their mothers have two-thirds of a year more schooling, on average. In addition to Galiani and McEwan (2013), the literature on Latin American CCTs usually finds that effects on school enrollment are larger among poorer households (Fiszbein and Schady, 2009). This implies that sample selection on observed (and perhaps unobserved) moderating variables—all common proxies for poverty—is responsible for the pattern of attenuated estimates in experiment 2.

C. External validity: weighted estimates in experiment 2

To further examine this issue, we estimated the probability that each observation in experiment 1 was selected for experiment 2 (using the logit specification described earlier). The mean difference in the estimated propensity score between the samples of experiment 2 and experiment 1 is 0.045 (or 37% of the standard deviation in the experiment 1 sample). For each of the 21 covariates, we then estimated the standardized difference between the weighted mean in experiment 2—applying the inverse-probability-weights described earlier—and the unweighted mean in experiment 1. As Figure 6 illustrates, re-weighting nearly eliminates observed differences between the two samples.

Finally, Tables 5 to 7 report weighted estimates in the experiment 2 sample. We anticipate that the weighted estimates will more closely resemble those from experiment 1. On the contrary, we find that the point estimates from unweighted and weighted specifications in experiment 2 are quite similar (and both exhibit similar patterns of attenuation relative to experiment 1). One interpretation is that sample selection into experiment 2 altered the distribution of unobserved variables that moderate treatment effects, leading to a violation of assumption 2.

What else might be done? One possibility is to implement a two-stage correction for sample selection, à la Heckman (1979). In the sample of experiment 1, one estimates a first-stage probit

with a dependent variable indicating selection into experiment 2. It includes the same independent variables as the logit used to estimate the inverse-probability weights, in addition to variable(s) that affects selection into experiment 1, but not child outcomes. Of course, compelling exclusion restrictions are usually hard to come by (and no obvious candidates exist in our application). Finally, in the second-stage regression, one includes the inverse Mills ratio as a regressor (along with other covariates) and examines its sign and significance for evidence of sample selection bias.

D. Internal validity

Recall the estimates in experiment 2 use the same group of treated children as the GQE, but with an experimental control group. How do they compare to the quasi-experimental GQE estimates reported in Table 3? For enrollment, the GQE estimates are attenuated relative to those of experiment 2 (which themselves were attenuated relative to those of experiment 1). This is especially evident in the blocks 1-2 subsample. The enrollment effect is 13.3 p.p. in experiment 2 and 10.6 in the GQE. The work-related variables provide less obvious conclusions because the estimates—in the experiment 2 and GQE samples—are small and not significant at 5% in blocks 1-5. In blocks 1-2, however, the point estimates for work-outside-home have a similar magnitude in both samples. In summary, the evidence is suggestive the GQE enrollment estimates are downward biased relative to the unbiased estimates from experiment 2.

Finally, Table 8 reports the placebo test described earlier. In the first column, the coefficients are small and statistically insignificant, providing some evidence that the GQE estimates are not explained by selection-on-unobservables. The pattern is not as clear in blocks 1-2, likely due to the much smaller samples of municipalities (recalling the experimental control group contained

fewer municipalities than the treatment group). Overall, the imprecision prevents us from ruling out some bias in the GQE estimates.

E. Compound treatment irrelevance

The GQE must assume that the conditional cash transfer is the only treatment that varies across borders (or, if there is another, that is does not affect potential outcomes). Keele and Titiunik (2015, 2016) describe this assumption as compound treatment irrelevance. In the Honduran context, the most likely violation occurs when a municipal border is also a department border. Although the management and financing of Honduran public schools is still highly centralized, each department controls some functions, especially related to personnel management. This leaves open the possibility that the assumption is violated, and so we assess robustness to the dropping of observations near municipal border segments that also happen to be department borders.

Of course, this occasions further non-random sample restrictions, which may affect external validity. Thus, Tables D.1 to D.3 repeat all experimental analyses from Tables 5 to 7 after excluding municipal border segments that are also a department border. The immediate result is a reduction in the number of municipalities in experiment 2 (from 52 to 43). Despite the reduced precision, the substantive findings are similar to earlier ones, focusing especially on blocks 1-2. The large effects for enrollment and work-at-home (in experiment 1) are attenuated in experiment 2, while the effects are more robust across samples for work-outside-home. Applying inverse probability weights to experiment 2 again has little effect on the point estimates. Table D.4 then replicates the GQE estimates. Here too, the substantive conclusions are similar. The only obvious difference is an attenuation of the enrollment estimate in blocks 1-2 (from 10.6 to

7.3 percentage points), though the full-sample estimates are within 0.2 percentage points of one another.

6. Conclusions

This paper analyzed a geographic quasi-experiment embedded in a cluster-randomized experiment in Honduras. In the experiment, average treatment effects on school enrollment and child labor were large—especially in the poorest blocks—and could be generalized to a policy-relevant population given the original sample selection criteria (Galiani and McEwan, 2013; IFPRI, 2000). In contrast, the geographic quasi-experiment yielded point estimates that, for two of three dependent variables, were attenuated. A judicious policy analyst without access to the experimental results might have provided misleading advice based on the magnitude of point estimates.

We assessed two main explanations for the difference in point estimates, related to external and internal validity. The GQE sample is necessarily restricted to children residing close to a municipal border with cross-border variation in the treatment. Sample selection modifies the distribution of some observed and (perhaps) unobserved variables that moderate treatment effects, relative to the original experiment. We find that this explains some, but not all of the attenuation, especially for school enrollment effects. The remainder is plausibly explained by imbalance in unobserved variables between treatment and control groups in the 2-kilometer GQE sample. While there is treatment-control balance along a wide range of pre-specified covariates, the GQE enrollment estimates are still attenuated relative to the benchmark estimates of experiment 2.

Both findings suggest that researchers using geographic designs should carefully describe how their geographically-imposed convenience sample differs from that of a well-defined, policy relevant population. If feasible, they might further apply inverse-probability weighting as a robustness check (following Cole and Stuart, 2010 and the analysis herein). Moreover, they should carefully assess treatment-control balance in the geographic sample. In this volume, Keele et al. (2016) discuss related consideration when units of analysis (such as households) cannot be precisely geo-located.

The findings on external validity have broader implications for the design and interpretation of randomized field experiments, which often rely on convenience samples defined by observed and unobserved moderators of treatment effects (such as poverty, distance to urban centers, agents' willingness to submit to randomization, and so on). At a minimum, experiments should specifically describe the criteria for sample selection (e.g., Campbell et al, 2012), and whether these variables are plausible moderators of treatment effects. Our paper suggests that authors can push this exercise further and assess robustness after re-weighting the experimental convenience sample to resemble a policy-relevant population, with appropriate caveats about selection-on-unobservables into the convenience sample (Hotz et al., 2005; Cole and Stuart, 2010; Stuart et al., 2015).

References

Allcott, H. (2015). Site Selection Bias in Program Evaluation. *The Quarterly Journal of Economics, 130*(3), 1117-1165. doi:10.1093/qje/qjv015

Alzúa, M. L., Cruces, G., & Ripani, L. (2013). Welfare Programs and Labor Supply in Developing Countries: Experimental Evidence from Latin America. *Journal of Population Economics, 26*(4), 1255-1284. doi:10.1007/s00148-012-0458-0

Angrist, J. D., & Rokkanen, M. (2015). Wanna Get Away? Regression Discontinuity Estimation of Exam School Effects Away from the Cutoff. *Journal of the American Statistical Association, 110,* 1331-1344.

Benedetti, F., Ibarrarán, P., & McEwan, P. J. (2016). Do Education and Health Conditions Matter in a Large Cash Transfer? Evidence from a Honduran Experiment. *Economic Development and Cultural Change*. doi:10.1086/686583

Billings, S. B., Deming, D. J., & Rockoff, J. (2013). School Segregation, Educational Attainment, and Crime: Evidence from the End of Busing in Charlotte-Mecklenburg. *The Quarterly Journal of Economics, 129*(1), 435-476. doi:10.1093/qje/qjt026

Black, S. (1999). Do Better Schools Matter? Parental Valuation of Elementary Education. *Quarterly Journal of Economics, 114,* 577–599.

Bold, T., Kimenyi, M., Mwabu, G., Ng'ang'a, A., & Sandefur, J. (2013). *Scaling-up What Works: Experimental Evidence on External Validity in Kenyan Education.* Centre for the Study of African Economies WPS/2013-04.

Buddelmeyer, H., & Skoufias, E. (2004). An Evaluation of the Performance of Regression Discontinuity Design on PROGRESA. Policy Research Working Paper 3386. World Bank, Washington, DC.

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014). Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs. *Econometrica, 82*(6), 2295-2326. doi:10.3982/ecta11757

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-Based Improvements for Inference with Clustered Errors. *Review of Economics and Statistics, 90*(3), 414-427. doi:10.1162/rest.90.3.414

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2011). Robust Inference With Multiway Clustering. *Journal of Business & Economic Statistics, 29*(2), 238-249. doi:10.1198/jbes.2010.07136

Campbell, M. K., Piaggio, G., Elbourne, D. R., & Altman, D. G. (2012). Consort 2010 Statement: Extension to Cluster Randomised Trials. BMJ, 345, e5661. doi:10.1136/bmj.e5661

Cattaneo, M. D., Frandsen, B. R., & Titiunik, R. (2015). Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the U.S. Senate. *Journal of Causal Inference, 3*(1), 1-24. doi:10.1515/jci-2013-0010

Cole, S. R., & Stuart, E. A. (2010). Generalizing Evidence From Randomized Clinical Trials to Target Populations: The ACTG 320 Trial. *American Journal of Epidemiology, 172*(1), 107-115. doi:10.1093/aje/kwq084

Dell, M. (2010). The Persistent Effects of Peru's Mining *Mita. Econometrica, 78*(6), 1863-1903.

Dong, Y., & Lewbel, A. (2015). Identifying the Effect of Changing the Policy Threshold in Regression Discontinuity Models. *Review of Economics and Statistics, 97*(5), 1081-1092.

Fiszbein, A., & Schady, N. (2009). *Conditional Cash Transfers: Reducing Present and Future Poverty.* Washington, DC: World Bank.

Galiani, S., & McEwan, P. J. (2013). The Heterogeneous Impact of Conditional Cash Transfers. *Journal of Public Economics, 103,* 85-96. doi:10.1016/j.jpubeco.2013.04.004

Glewwe, P., & Olinto, P. (2004). *Evaluating of the Impact of Conditional Cash Transfers on Schooling: An Experimental Analysis of Honduras' PRAF Program, Final Report for USAID.*

Hahn, J., Todd, P., & Van der Klaauw, W. (2001). Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design. *Econometrica, 69*(1), 201-209.

Hansen, B. B., & Bowers, J. (2008). Covariate Balance in Simple, Stratified and Clustered Comparative Studies. *Statistical Science, 23*(2), 219-236. doi:10.1214/08-sts254

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica, 47,* 153-161.

Hotz, V. J., Imbens, G. W., & Mortimer, J. H. (2005). Predicting the Efficacy of Future Training Programs Using Past Experiences at Other Locations. *Journal of Econometrics, 125*(1-2), 241-270. doi:10.1016/j.jeconom.2004.04.009

Imbens, G. W., & Wooldridge, J. M. (2009). Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature, 47*(1), 5-86. doi:10.1257/jel.47.1.5

Imbens, G., & Zajonc, T. (2009). *Regression Discontinuity Design with Vector-Argument Assignment Rules.* Mimeo.

International Food Policy Research Institute (IFPRI). (2000). *Second Report: Implementation Proposal for the PRAF/IDB Project—Phase II.* Washington, DC: International Food Policy Research Institute.

Keele, L., Lorch, S., Passarella, M., Small, D., & Titiunik, R. (2016). An Overview of Geographically Discontinuous Treatment Assignments with an Application to Children's Health Insurance. In M. D. Cattaneo & J. C. Escanciano (Eds.), *Advances in Econometrics, 38.*

Keele, L. J., & Titiunik, R. (2015). Geographic Boundaries as Regression Discontinuities. *Political Analysis, 23*(1), 127-155. doi:10.1093/pan/mpu014

Keele, L., & Titiunik, R. (2016). Natural Experiments Based on Geography. *Political Science Research and Methods, 4*(1), 65-95. doi:10.1017/psrm.2015.4

Lee, D. S. (2008). Randomized Experiments from Non-random Selection in U.S. House Elections. *Journal of Econometrics, 142*(2), 675-697. doi:10.1016/j.jeconom.2007.05.004

Lee, D. S., & Lemieux, T. (2010). Regression Discontinuity Designs in Economics. *Journal of Economic Literature, 48*(2), 281-355. doi:10.1257/jel.48.2.281

Lucas, A. M., McEwan, P. J., Ngware, M., & Oketch, M. (2014). Improving Early-Grade Literacy in East Africa: Experimental Evidence from Kenya and Uganda. *Journal of Policy Analysis and Management, 33*(4), 950-976. doi:10.1002/pam.21782

Moore, C. (2008). Assessing Honduras' CCT Programme PRAF, Programa de Asignación Familiar: Expected and Unexpected Realities. *Country Study No. 15.* International Poverty Center.

Morris, S. S., Flores, R., Olinto, P., & Medina, J. M. (2004). Monetary Incentives in Primary Health Care and Effects on Use and Coverage of Preventive Health Care Interventions in Rural Honduras: Cluster Randomised Trial. *The Lancet, 364*(9450), 2030-2037. doi:10.1016/s0140-6736(04)17515-6

Muller, S. M. (2015). Causal Interaction and External Validity: Obstacles to the Policy Relevance of Randomized Evaluations. *The World Bank Economic Review, 29,* S217-S225. doi:10.1093/wber/lhv027

Oosterbeek, H., Ponce, J., & Schady, N. (2008). The Impact of Cash Transfers on School Enrollment: Evidence from Ecuador. Policy Research Working Paper 4645. World Bank, Washington, DC.

Rubin, D. B. (2001). Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation. *Health Services and Outcomes Research Methodology, 2*(3/4), 169-188. doi:10.1023/a:1020363010465

Secretaría de Educación. (1997). *VII Censo Nacional de Talla, Informe 1997.* Tegucigalpa: Secretaría de Educación, Programa de Asignación Familiar.

Shadish, J. W., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference.* Houghton Mifflin.

Shadish, W. R., Galindo, R., Wong, V. C., Steiner, P. M., & Cook, T. D. (2011). A randomized experiment comparing random and cutoff-based assignment. *Psychological Methods, 16*(2), 179-191.

Stuart, E. A., Bradshaw, C. P., & Leaf, P. J. (2015). Assessing the Generalizability of Randomized Trial Results to Target Populations. *Prevention Science, 16*(3), 475-485. doi:10.1007/s11121-014-0513-z

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The Use of Propensity Scores to Assess the Generalizability of Results from Randomized Trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 174*(2), 369-386. doi:10.1111/j.1467-985x.2010.00673.x

World Bank. (2006). *Honduras Poverty Assessment: Attaining Poverty Reductions, Report No. 35622-HN.* Washington: World Bank.

Table 1: Replication of experimental estimates in Galiani and McEwan (2013)

	All blocks	Blocks 1-2	Blocks 3-5
Enrolled in school	0.081***	0.150***	0.035
	(0.024)	(0.035)	(0.025)
N	120,411	44,358	76,053
Control mean	0.646	0.600	0.680
BS p(sym)	0.007	0.005	0.198
Works outside home	-0.031***	-0.068***	-0.008
	(0.012)	(0.017)	(0.013)
N	98,783	36,261	62,522
Control mean	0.097	0.126	0.075
BS p(sym)	0.013	0.005	0.585
Works in home	-0.041***	-0.062***	-0.027
	(0.013)	(0.018)	(0.017)
N	98,783	36,261	62,522
Control mean	0.136	0.150	0.126
BS p(sym)	0.010	0.005	0.145

Note: *** indicates statistical significance at 1%, ** at 5%, and * at 10%. Robust standard errors are in parentheses, clustered by municipality (70, 28, and 42, respectively, in all blocks, blocks 1-2, and blocks 3-5). BS p(sym) is the symmetric p-value from a wild cluster-bootstrap percentile-t procedure with 399 replications. The regressions in this and all subsequent tables control for block dummy variables, the 21 covariates in Table A.1, squared terms for 6 continuous variables, and dummy variables indicating missing values of variables. Sample sizes are lower for the work-related variables because the census restricts the question to children 7 and older.

Table 2: Balance in the geographic quasi-experiment (≤ 2 kilometers from border)

	T/C differences	p-value
Age	-0.002/-0.001	0.979
Female	0.007/0.013	0.383
Born in municipality	-0.004/-0.013	0.849
Lenca	0.094/0.231	0.166
Moved	0.006/0.032	0.505
Father is literate	0.040/0.081	0.470
Mother is literate	0.030/0.060	0.531
Father's schooling	0.231/0.076	0.689
Mother's schooling	0.147/0.050	0.778
Dirt floor	0.008/0.017	0.915
Piped water	0.004/0.009	0.949
Electricity	0.022/0.053	0.827
Rooms in dwelling	0.041/0.054	0.747
Sewer/septic	0.030/0.063	0.683
Auto	0.006/0.029	0.780
Refrigerator	-0.014/-0.051	0.791
Computer	-0.001/-0.020	0.746
Television	0.012/0.035	0.890
Mitch	-0.002/-0.009	0.890
Household members	0.054/0.022	0.692
Household members, 0-17	0.013/0.007	0.920
Predicted mean municipal child height-for-age z-score	-0.464/-1.538	0.001

Note: In the difference column, the first number is the mean difference and the second number if mean difference divided by the full-sample standard deviation. p-values account for clustering by municipality.

Table 3: Estimates of the geographic quasi-experiment (≤ 2 kilometers from border)

	All blocks	Blocks 1-2	Blocks 3-5
Enrolled in school	0.057*	0.106**	0.035
	(0.032)	(0.054)	(0.041)
N	24360/81/65	6888/26/19	17472/61/46
BS p(sym)	0.095	0.072	0.463
Diff RCT p(sym)	0.537	0.580	0.998
Works outside home	-0.024	-0.086*	-0.005
	(0.015)	(0.044)	(0.016)
N	20009/81/65	5591/26/19	14418/61/46
BS p(sym)	0.122	0.025	0.743
Diff RCT p(sym)	0.677	0.715	0.885
Works only inside home	0.014	-0.013	0.026
	(0.017)	(0.026)	(0.022)
N	20009/81/65	5591/26/19	14418/61/46
BS p(sym)	0.468	0.613	0.330
Diff RCT p(sym)	0.010	0.105	0.048

Note: *** indicates statistical significance at 1%, ** at 5%, and * at 10%. Robust standard errors are in parentheses, clustered by municipality and border segment (see text for details). N indicates the number of eligible children, municipalities, and border segments. BS p(sym) is the symmetric p-value from a wild cluster- bootstrap percentile-t procedure (clustering on municipalities) with 399 replications. Diff RCT p(sym) is the p-value from the test of equality between a GQE estimate and the corresponding experimental estimate from table 1, computed using 399 replications. All regressions control for the variables described in the note to Table 1 (but excluding block dummy variables).

Table 4: Experimental samples used to assess external and internal validity of the GQE

	Sample restriction on full experimental sample	Parameter(s)
Full experimental sample	—	ATE
Experiment 1	2 km from any municipal border	ATE_1
Experiment 2	2 km from municipal borders shared with untreated, non-experimental caseríos	Unweighted: ATE_2 *Weighted: ATE_1

Note: ATE is the average treatment effect in the full experimental sample (Galiani & McEwan, 2013), and subscripts indicate average treatment effects in subsamples of the experiment. * indicates that identification relies on a selection-on-observables assumption described in the text.

Table 5: Estimates for experiments 1 and 2 (dependent variable: enrolled in school)

	All blocks	Blocks 1-2	Blocks 3-5
Experiment 1	0.089***	0.163***	0.041
	(0.025)	(0.042)	(0.025)
N	65310/70	26122/28	39188/42
BS p(sym)	0.013	0.010	0.180
Experiment 2	0.063*	0.133**	0.035
	(0.035)	(0.059)	(0.040)
N	21703/52	6996/17	14707/35
BS p(sym)	0.133	0.095	0.455
Experiment 2 (weighted)	0.064*	0.134**	0.022
	(0.035)	(0.054)	(0.037)
N	21703/52	6996/17	14707/35
BS p(sym)	0.107	0.055	0.560

Note: *** indicates statistical significance at 1%, ** at 5%, and * at 10%. Robust standard errors are in parentheses, clustered by municipality. N indicates the number of eligible children and the number of municipalities. BS p(sym) is the symmetric p-value from a wild cluster-bootstrap percentile-t procedure with 399 replications. All regressions control for the variables described in the note to Table 1.

Table 6: Estimates for experiments 1 and 2 (dependent variable: works outside home)

	All blocks	Blocks 1-2	Blocks 3-5
Experiment 1	-0.042***	-0.082***	-0.017
	(0.012)	(0.022)	(0.012)
N	53703/70	21387/28	32316/42
BS p(sym)	0.003	0.003	0.217
Experiment 2	-0.032*	-0.090**	-0.009
	(0.017)	(0.034)	(0.017)
N	17883/52	5691/17	12192/35
BS p(sym)	0.095	0.013	0.632
Experiment 2 (weighted)	-0.034*	-0.087**	-0.009
	(0.017)	(0.033)	(0.018)
N	17883/52	5691/17	12192/35
BS p(sym)	0.070	0.007	0.637

Note: *** indicates statistical significance at 1%, ** at 5%, and * at 10%. Robust standard errors are in parentheses, clustered by municipality. N indicates the number of eligible children and the number of municipalities. BS p(sym) is the symmetric p-value from a wild cluster-bootstrap percentile-t procedure with 399 replications. All regressions control for the variables described in the note to Table 1.

Table 7: Estimates for experiments 1 and 2 (dependent variable: works only in home)

	All blocks	Blocks 1-2	Blocks 3-5
Experiment 1	-0.041***	-0.067***	-0.024
	(0.013)	(0.021)	(0.016)
N	53703/70	21387/28	32316/42
BS p(sym)	0.007	0.025	0.188
Experiment 2	-0.024	-0.026	-0.023
	(0.019)	(0.025)	(0.023)
N	17883/52	5691/17	12192/35
BS p(sym)	0.320	0.352	0.448
Experiment 2 (weighted)	-0.018	-0.020	-0.013
	(0.019)	(0.025)	(0.022)
N	17883/52	5691/17	12192/35
BS p(sym)	0.438	0.515	0.630

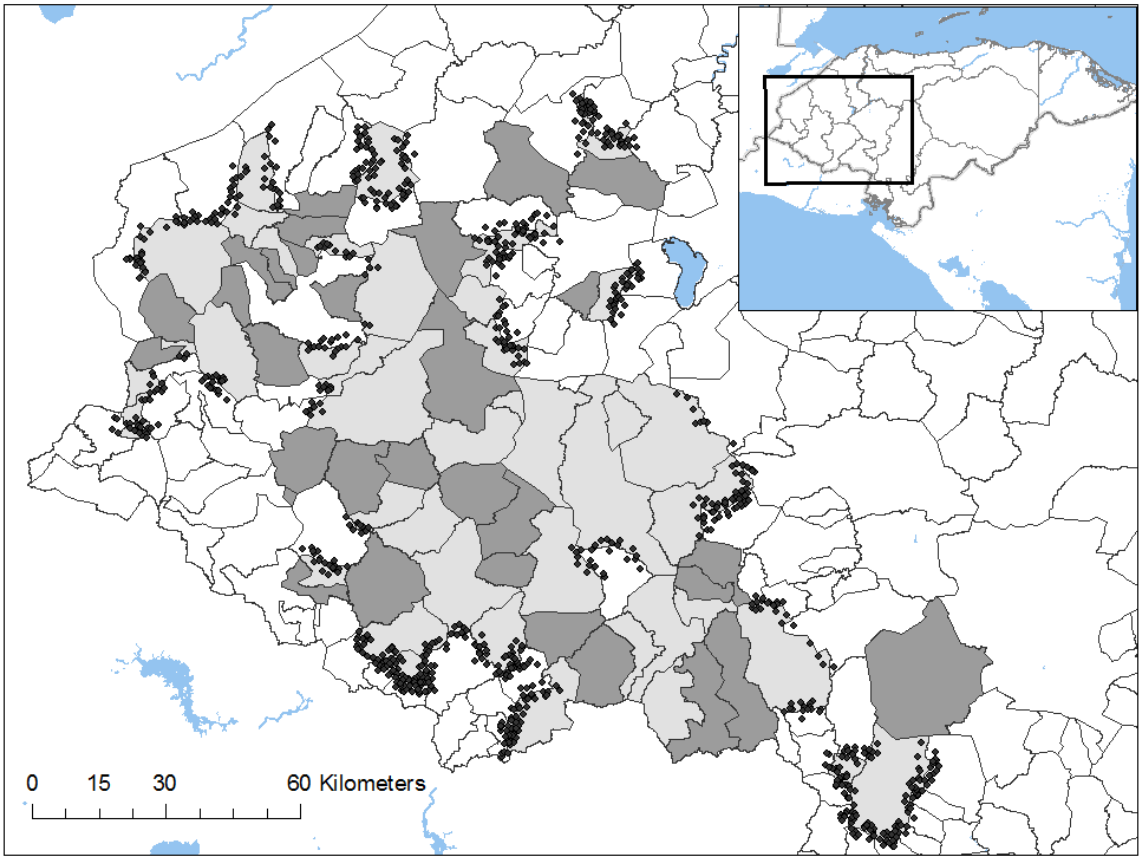
Note: *** indicates statistical significance at 1%, ** at 5%, and * at 10%. Robust standard errors are in parentheses, clustered by municipality. N indicates the number of eligible children and the number of municipalities. BS p(sym) is the symmetric p-value from a wild cluster-bootstrap percentile-t procedure with 399 replications. All regressions control for the variables described in the note to Table 1.

Table 8: Placebo estimates (≤ 2 kilometers from border)

	All blocks	Blocks 1-2	Blocks 3-5
Enrolled in school	-0.025	0.073	-0.052
	(0.036)	(0.063)	(0.035)
N	13980/47/40	4064/17/12	9916/33/28
BS p(sym)	0.525	0.415	0.165
Works outside home	-0.014	-0.035	-0.008
	(0.018)	(0.054)	(0.015)
N	11573/47/40	3365/17/12	8208/33/28
BS p(sym)	0.542	0.705	0.623
Works only inside home	-0.001	-0.069**	0.024
	(0.019)	(0.032)	(0.018)
N	11573/47/40	3365/17/12	8208/33/28
BS p(sym)	0.930	0.135	0.188

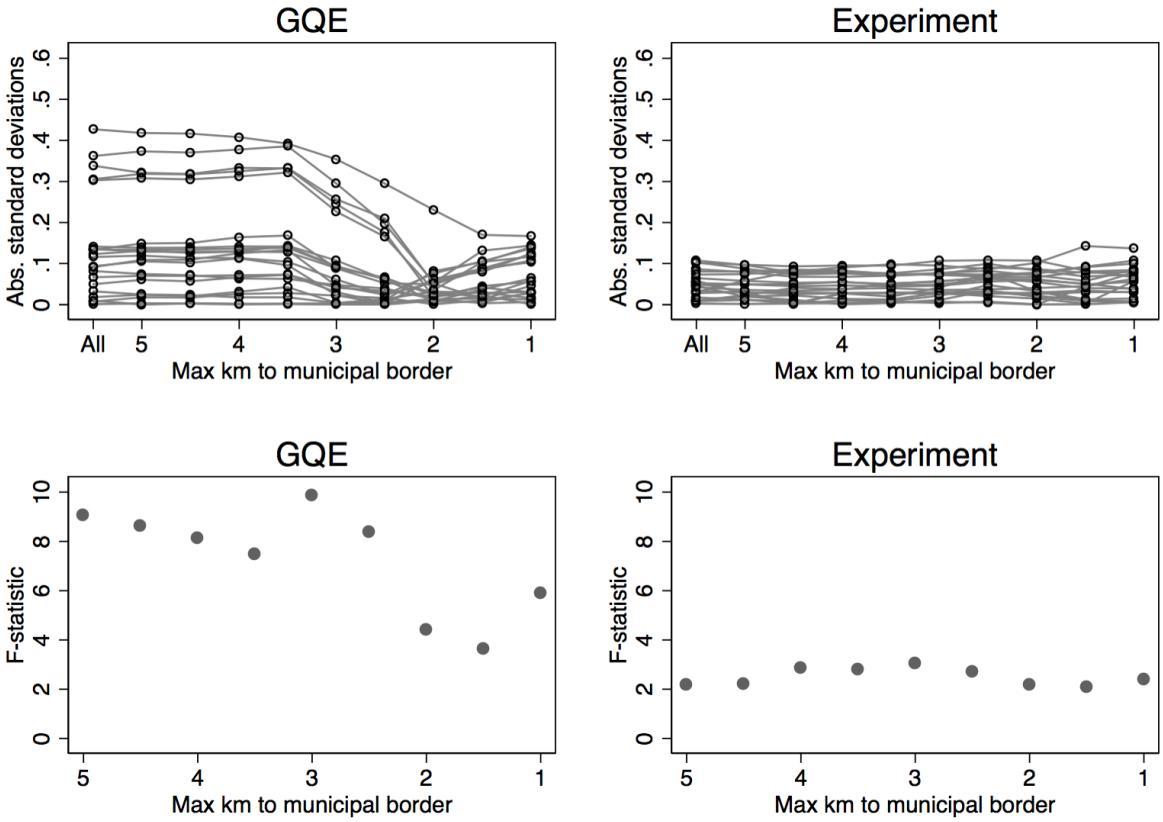
Note: *** indicates statistical significance at 1%, ** at 5%, and * at 10%. Robust standard errors are in parentheses, clustered by municipality and border segment (see text for details). N indicates the number of eligible children, municipalities, and border segments. BS p(sym) is the symmetric p-value from a wild cluster- bootstrap percentile-t procedure (clustering on municipalities) with 399 replications. All regressions control for the variables described in the note to Table 1 (but excluding block dummy variables).

Figure 1: Caseríos in the geographic quasi-experiment



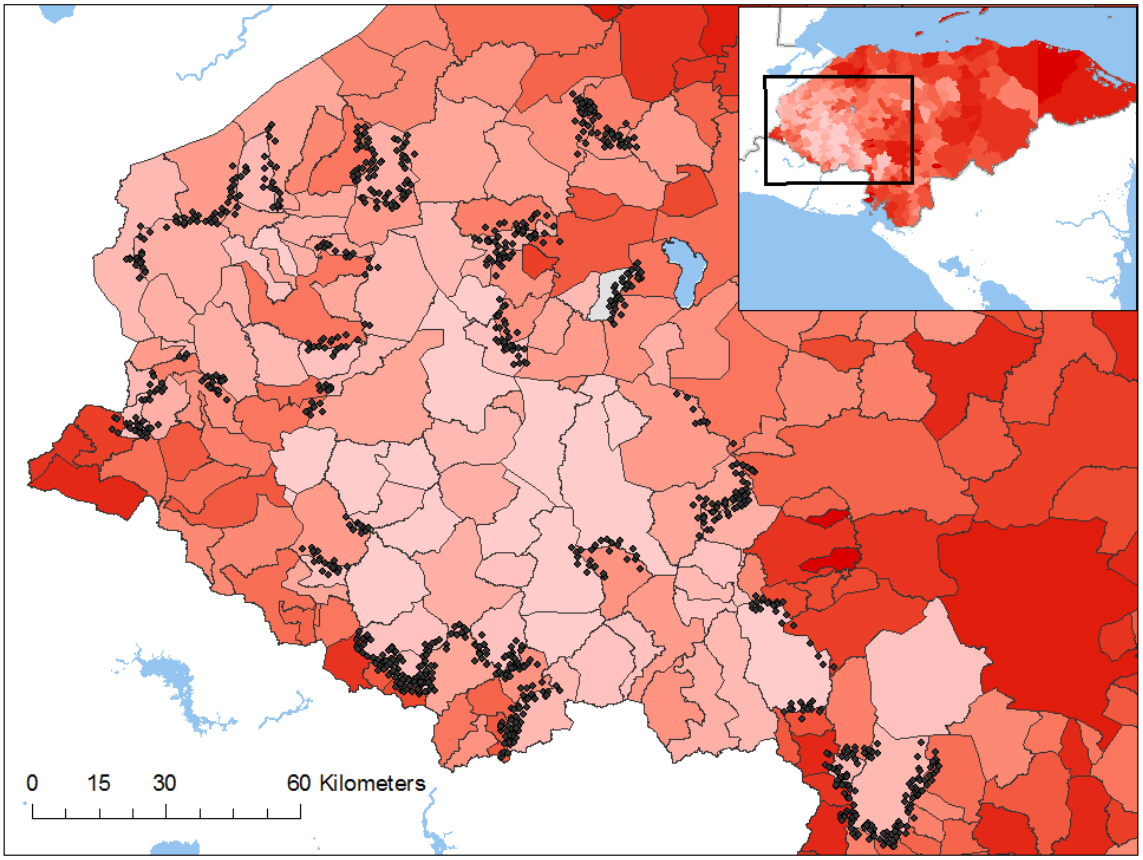
Note: Experimental treatment municipalities are lightly shaded; experimental control municipalities are darkly shaded. Unshaded areas are untreated, non-experimental municipalities. Dots indicate caseríos within 2 kilometers of municipal borders shared by experimental treatment municipalities and untreated non-experimental municipalities. The inset map indicates department borders.

Figure 2: Covariate balance increasingly close to municipal borders



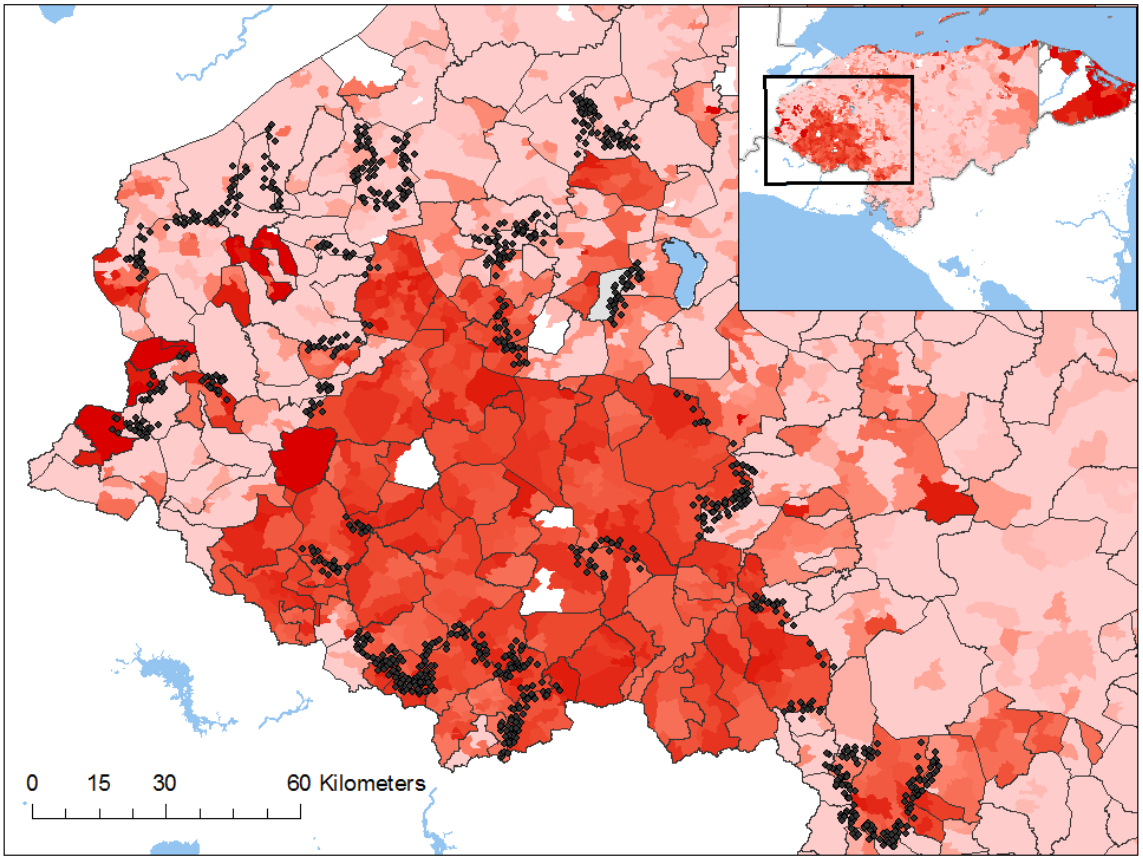
Note: In the upper panels, dots indicate the absolute value of the standardized mean difference (using the full-sample standard deviation) between the treatment and comparison group for 21 covariates in Table A.1, within the specified distance-to-border (Appendix B reports unstandardized differences, standardized differences, and cluster-adjusted p-values). In the lower panels, dots indicate F-statistics from regressions of the treatment dummy on the 21 covariates, squared terms for 6 continuous covariates, and dummy variables indicating missing values.

Figure 3: Mean municipal height-for-age Z-scores in 1997



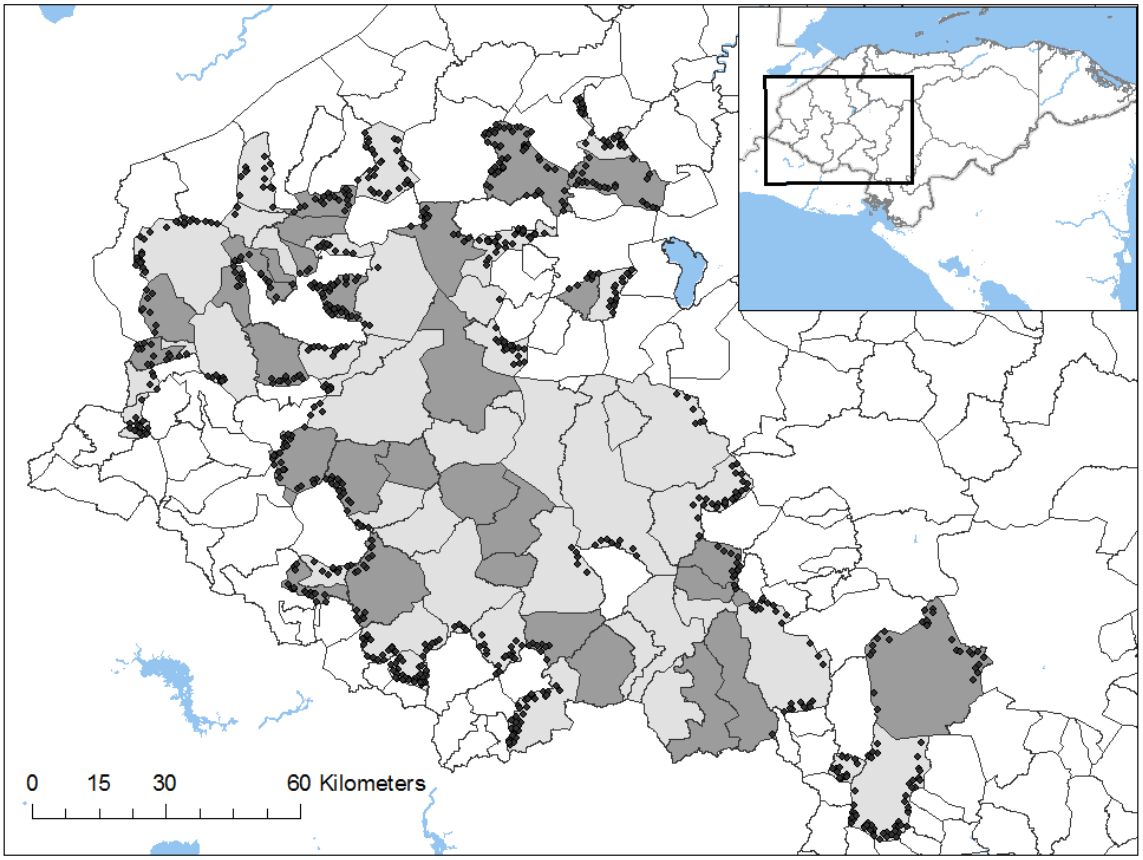
Note: Lighter shades indicate increasingly negative values of the predicted mean municipal height-for-age Z-scores (Galiani and McEwan, 2013), using 20 quantiles of the municipal distribution. Municipal borders are outlined. Dots indicate caserios in the geographic quasi-experimental sample within 2 kilometers of municipal borders.

Figure 4: Proportion of eligible children self-identifying as Lenca in 2001



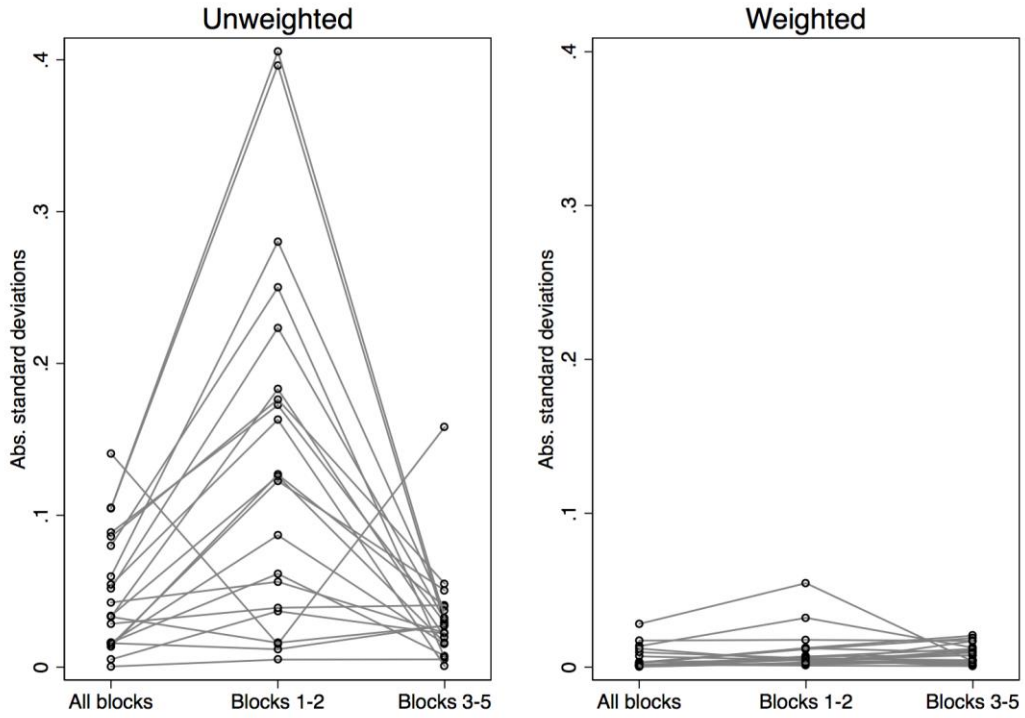
Note: Darker shades indicate higher proportions of children self-identify as Lenca (or another racial or ethnic minority), using 20 quantiles of the aldea (village) distribution. Municipal borders are outlined. Dots indicate caserios in the geographic quasi-experimental sample within 2 kilometers of municipal borders.

Figure 5: Caseríos in experiment 2



Note: Experimental treatment municipalities are lightly shaded; experimental control municipalities are darkly shaded. Unshaded areas are untreated, non-experimental municipalities. Dots indicate caseríos within 2 kilometers of a municipal border with untreated non-experimental municipalities. The inset map indicates department borders.

Figure 6: Comparing covariate means between samples in experiments 1 and 2



Note: Dots indicate the absolute value of the standardized mean difference (using the full-sample standard deviation) between the pooled experiment 2 sample and the pooled experiment 1 sample, for the 21 covariates in Table A.1. In all panels, the sample includes caseríos within 2 kilometers of municipal borders. In the right panel, the mean of the experiment 2 sample is weighted, as described in the text.

Appendix A

Table A.1: Census variable definitions

	Variable definition and census question(s) used to construct variable
Dependent variables	
Enrolled in school	1=Enrolled in school on census date; 0=not (F8).
Works outside home	1=Worked during past week, including self-employment, family business, and agricultural work; 0=not (F12, F13A01-04); only reported for ages 7 and up.
Works only in home	1=Worked during past week, exclusively on household chores; 0=not (F13B10). Variable only reported for ages 7 and up.
Independent variables	
Age	Integer age on census date (F3).
Female	1=Female; 0=Male (F2).
Born in municipality	1=Born in present municipality; 0=not (F4A).
Lenca	1=Lenca or other non-mestizo ethnicity/race; 0=not (F5).
Moved	1=Resided in a different caserío, aldea, or city in 1996;0=resided in current caserío, aldea, or city in 1996 ().
Father is literate	1=Father is literate; 0=not (F7, F1, F2).
Mother is literate	1=Mother is literate; 0=not (F7, F1, F2).
Father's schooling	Years of father's schooling (F9, F1, F2).
Mother's schooling	Years of mother's schooling (F9, F1, F2).
Dirt floor	1=Dwelling has dirt floor; 0=not (B5).
Piped water	1=Dwelling has piped water from public or private source; 0=not (B6).
Electricity	1=Electric light from private or public source; 0=light from another source (ocote, etc.) (B8).
Rooms in dwelling	Number of bedrooms used by household (C1).
Sewer/septic	1=Household has toilet connected to sewer or septic system; 0=not (C5).
Auto	1=Household has at least one auto; 0=not (C7).
Refrigerator	1=Household has refrigerator; 0=not (C8a).
Computer	1=Household has computer; 0=not (C8g).
Television	1=Household has television; 0=not (C8e).
Mitch	1= at least 1 household member emigrated after Hurricane Mitch in October 1998; 0=not (E1).
Household members	Total individuals residing in household.
Household members, 0-17	Total individuals, ages 0-17, residing in household.

Notes: The 2001 Honduran census form is available at http://unstats.un.org/unsd/demographic/sources/census/quest/HND2001es.pdf

Appendix B: Treatment-control balance

Table B.1: Balance in experiments 1 and 2 (≤ 2 kilometers from border)

	Experiment 1 sample		Experiment 2 sample	
	T/C differences	p-value	T/C differences	p-value
Age	-0.055/-0.029	0.124	-0.066/-0.036	0.333
Female	0.000/0.000	0.961	0.005/0.011	0.452
Born in municipality	-0.016/-0.059	0.325	-0.015/-0.055	0.482
Lenca	-0.007/-0.014	0.925	-0.057/-0.129	0.632
Moved	0.004/0.026	0.419	0.015/0.082	0.088
Father is literate	0.030/0.062	0.305	0.029/0.059	0.539
Mother is literate	0.027/0.055	0.277	0.030/0.061	0.467
Father's schooling	0.228/0.086	0.285	0.432/0.150	0.382
Mother's schooling	0.215/0.083	0.234	0.407/0.144	0.350
Dirt floor	0.046/0.102	0.312	0.078/0.167	0.347
Piped water	-0.000/-0.001	0.994	-0.001/-0.003	0.981
Electricity	0.012/0.032	0.802	0.085/0.217	0.322
Rooms in dwelling	0.050/0.071	0.308	0.104/0.139	0.342
Sewer/septic	0.049/0.107	0.171	0.050/0.105	0.434
Auto	-0.008/-0.040	0.455	0.012/0.058	0.512
Refrigerator	0.004/0.019	0.831	0.035/0.142	0.377
Computer	0.001/0.022	0.404	0.003/0.055	0.211
Television	0.009/0.033	0.779	0.068/0.219	0.341
Mitch	0.008/0.067	0.088	0.024/0.170	0.006
Household members	0.186/0.078	0.087	0.310/0.128	0.067
Household members, 0-17	0.146/0.076	0.115	0.274/0.141	0.075
Predicted mean municipal child height-for-age z-score	0.001/0.004	0.989	-0.029/-0.129	0.735

Note: See Table 4 for sample definitions of experiment 1 and 2. In the difference columns, the first number is the mean difference and the second number is the mean difference divided by the full-sample standard deviation. p-values account for clustering by municipality.

Table B.2: Balance in the placebo sample (≤ 2 kilometers from border)

	T/C differences	p-value
Age	0.057/0.030	0.191
Female	0.005/0.010	0.513
Born in municipality	0.012/0.045	0.537
Lenca	0.200/0.485	0.085
Moved	-0.013/-0.076	0.063
Father is literate	0.009/0.019	0.750
Mother is literate	-0.043/-0.087	0.111
Father's schooling	-0.069/-0.026	0.771
Mother's schooling	-0.239/-0.095	0.261
Dirt floor	-0.020/-0.041	0.800
Piped water	-0.037/-0.077	0.439
Electricity	-0.046/-0.127	0.382
Rooms in dwelling	-0.007/-0.010	0.867
Sewer/septic	0.023/0.050	0.632
Auto	-0.015/-0.075	0.177
Refrigerator	-0.008/-0.037	0.691
Computer	-0.000/-0.014	0.587
Television	-0.019/-0.073	0.495
Mitch	-0.008/-0.084	0.066
Household members	-1.036/-0.156	0.266
Household members, 0-17	-0.977/-0.161	0.254
Predicted mean municipal child height-for-age z-score	-0.495/-1.423	0.001

Note: See the text for definition of the placebo sample. In the difference column, the first number is the mean difference and the second number is the mean difference divided by the full-sample standard deviation. p-values account for clustering by municipality.

Appendix C: Covariate differences between experimental subsamples (≤ 2 kilometers from border)

Table C.1: All blocks

	Experiment 2 - Experiment 1	
	Unweighted	Weighted
Age	0.009/0.005	-0.002/-0.001
Female	0.000/0.000	-0.001/-0.001
Born in municipality	-0.004/-0.016	0.008/0.028
Lenca	-0.066/-0.141	0.002/0.003
Moved	0.003/0.016	0.000/0.001
Father is literate	-0.014/-0.029	-0.001/-0.003
Mother is literate	0.017/0.034	0.000/0.000
Father's schooling	0.087/0.033	-0.007/-0.003
Mother's schooling	0.208/0.080	-0.003/-0.001
Dirt floor	-0.040/-0.089	0.001/0.003
Piped water	0.007/0.015	-0.007/-0.014
Electricity	0.037/0.105	0.003/0.007
Rooms in dwelling	0.037/0.052	-0.001/-0.001
Sewer/septic	0.039/0.086	-0.008/-0.017
Auto	0.003/0.013	0.000/0.001
Refrigerator	0.013/0.060	0.002/0.010
Computer	0.001/0.016	0.000/0.002
Television	0.028/0.105	0.003/0.012
Mitch	0.006/0.054	0.000/0.001
Household members	-0.079/-0.033	0.003/0.001
Household members, 0-17	-0.082/-0.043	0.000/0.000
Predicted mean municipal child height-for-age z-score	0.052/0.204	0.015/0.058
Block 1 or 2 (proportion)	-0.078/-0.158	-0.007/-0.015
Estimated propensity score	0.045/0.369	0.005/0.039

Note: In each cell, the first number is the mean difference between the experiment 2 sample and experiment 1 sample; the second number is the mean difference divided by the standard deviation in the experiment 1 sample. In the weighted column, the means for experiment 2 apply inverse probability weights described in the text.

Table C.2: Blocks 1-2

	Experiment 2 - Experiment 1	
	Unweighted	Weighted
Age	-0.069/-0.037	-0.003/-0.002
Female	-0.002/-0.005	-0.003/-0.005
Born in municipality	0.003/0.012	0.015/0.055
Lenca	-0.008/-0.015	0.002/0.004
Moved	0.010/0.061	-0.001/-0.006
Father is literate	0.019/0.039	0.006/0.012
Mother is literate	0.063/0.126	0.001/0.003
Father's schooling	0.486/0.183	0.031/0.012
Mother's schooling	0.658/0.250	0.006/0.002
Dirt floor	-0.073/-0.173	-0.005/-0.012
Piped water	0.060/0.123	-0.016/-0.032
Electricity	0.122/0.396	0.001/0.004
Rooms in dwelling	0.166/0.223	-0.005/-0.007
Sewer/septic	0.083/0.176	-0.008/-0.018
Auto	0.023/0.127	-0.001/-0.007
Refrigerator	0.055/0.280	0.001/0.006
Computer	0.003/0.087	0.000/0.001
Television	0.103/0.405	0.001/0.004
Mitch	0.020/0.163	0.001/0.005
Household members	-0.038/-0.016	-0.006/-0.002
Household members, 0-17	-0.107/-0.056	-0.005/-0.003
Predicted mean municipal child height-for-age z-score	0.041/0.213	0.026/0.134
Estimated propensity score	0.077/0.535	0.005/0.035

Note: In each cell, the first number is the mean difference between the experiment 2 sample and experiment 1 sample; the second number is the mean difference divided by the standard deviation in the experiment 1 sample. In the weighted column, the means for experiment 2 apply inverse probability weights described in the text.

Table C.3: Blocks 3-5

	Experiment 2 - Experiment 1	
	Unweighted	Weighted
Age	0.041/0.022	0.008/0.004
Female	0.003/0.005	-0.000/-0.001
Born in municipality	-0.008/-0.027	0.001/0.004
Lenca	-0.068/-0.158	-0.003/-0.008
Moved	-0.001/-0.007	-0.001/-0.004
Father is literate	-0.020/-0.041	-0.010/-0.020
Mother is literate	-0.008/-0.016	-0.005/-0.011
Father's schooling	-0.052/-0.020	-0.049/-0.019
Mother's schooling	-0.016/-0.006	-0.044/-0.017
Dirt floor	-0.014/-0.029	0.004/0.010
Piped water	-0.024/-0.051	-0.006/-0.013
Electricity	-0.011/-0.029	0.001/0.002
Rooms in dwelling	-0.022/-0.031	-0.008/-0.012
Sewer/septic	0.025/0.055	-0.008/-0.017
Auto	-0.008/-0.040	-0.002/-0.008
Refrigerator	-0.009/-0.037	0.001/0.004
Computer	-0.001/-0.015	0.000/0.001
Television	-0.009/-0.032	0.003/0.009
Mitch	0.000/0.001	-0.000/-0.002
Household members	-0.065/-0.027	0.007/0.003
Household members, 0-17	-0.043/-0.023	0.009/0.005
Predicted mean municipal child height-for-age z-score	0.007/0.084	0.001/0.008
Estimated propensity score	0.048/0.358	0.001/0.007

Note: In each cell, the first number is the mean difference between the experiment 2 sample and experiment 1 sample; the second number is the mean difference divided by the standard deviation in the experiment 1 sample. In the weighted column, the means for experiment 2 apply inverse probability weights described in the text.

Appendix D: Experimental and quasi-experimental estimates that exclude department borders (≤ 2 kilometers from border)

Table D.1: Estimates for experiments 1 and 2 (dependent variable: enrolled in school)

	All blocks	Blocks 1-2	Blocks 3-5
Experiment 1	0.094***	0.167***	0.047*
	(0.026)	(0.045)	(0.024)
N	54089/70	22344/28	31745/42
BS p(sym)	0.007	0.013	0.092
Experiment 2	0.056	0.141*	0.029
	(0.036)	(0.074)	(0.039)
N	17156/43	5736/14	11420/29
BS p(sym)	0.210	0.150	0.575
Experiment 2, weighted	0.056	0.139*	0.012
	(0.039)	(0.071)	(0.039)
N	17156/43	5736/14	11420/29
BS p(sym)	0.253	0.135	0.777

Note: *** indicates statistical significance at 1%, ** at 5%, and * at 10%. Robust standard errors are in parentheses, clustered by municipality. N indicates the number of eligible children and the number of municipalities. BS p(sym) is the symmetric p-value from a wild cluster-bootstrap percentile-t procedure with 399 replications. All regressions control for the variables described in the note to Table 1.

Table D.2: Estimates for experiments 1 and 2 (dependent variable: works outside home)

	All blocks	Blocks 1-2	Blocks 3-5
Experiment 1	-0.047***	-0.086***	-0.020*
	(0.013)	(0.024)	(0.012)
N	44433/70	18277/28	26156/42
BS p(sym)	0.003	0.003	0.152
Experiment 2	-0.033	-0.107**	-0.004
	(0.020)	(0.043)	(0.017)
N	14139/43	4656/14	9483/29
BS p(sym)	0.182	0.058	0.858
Experiment 2, weighted	-0.040*	-0.104**	-0.004
	(0.022)	(0.041)	(0.017)
N	14139/43	4656/14	9483/29
BS p(sym)	0.120	0.043	0.855

Note: *** indicates statistical significance at 1%, ** at 5%, and * at 10%. Robust standard errors are in parentheses, clustered by municipality. N indicates the number of eligible children and the number of municipalities. BS p(sym) is the symmetric p-value from a wild cluster-bootstrap percentile-t procedure with 399 replications. All regressions control for the variables described in the note to Table 1.

Table D.3: Estimates for experiments 1 and 2 (dependent variable: works only in home)

	All blocks	Blocks 1-2	Blocks 3-5
Experiment 1	-0.042***	-0.061**	-0.029*
	(0.014)	(0.024)	(0.015)
N	44433/70	18277/28	26156/42
BS p(sym)	0.018	0.075	0.105
Experiment 2	-0.016	-0.023	-0.015
	(0.020)	(0.030)	(0.024)
N	14139/43	4656/14	9483/29
BS p(sym)	0.545	0.502	0.578
Experiment 2, weighted	-0.008	-0.022	-0.001
	(0.020)	(0.030)	(0.024)
N	14139/43	4656/14	9483/29
BS p(sym)	0.790	0.545	0.958

Note: *** indicates statistical significance at 1%, ** at 5%, and * at 10%. Robust standard errors are in parentheses, clustered by municipality. N indicates the number of eligible children and the number of municipalities. BS p(sym) is the symmetric p-value from a wild cluster-bootstrap percentile-t procedure with 399 replications. All regressions control for the variables described in the note to Table 1.

Table D.4: Estimates for geographic quasi-experiment (all dependent variables)

	All blocks	Blocks 1-2	Blocks 3-5
Enrolled in school	0.059*	0.073	0.047
	(0.030)	(0.055)	(0.041)
N	17288/56/43	5835/17/13	11453/41/30
BS p(sym)	0.062	0.205	0.300
Works outside home	-0.025	-0.085*	-0.001
	(0.017)	(0.052)	(0.014)
N	14200/56/43	4724/17/13	9476/41/30
BS p(sym)	0.133	0.058	0.968
Works only inside home	0.014	-0.005	0.023
	(0.017)	(0.028)	(0.024)
N	14200/56/43	4724/17/13	9476/41/30
BS p(sym)	0.468	0.887	0.398

Note: *** indicates statistical significance at 1%, ** at 5%, and * at 10%. Robust standard errors are in parentheses, clustered by municipality and border segment (see text for details). N indicates the number of eligible children, municipalities, and border segments. BS p(sym) is the symmetric p-value from a wild cluster-bootstrap percentile-t procedure (clustering on municipalities) with 399 replications. All regressions control for the variables described in the note to Table 1 (but excluding block dummy variables).