CEPR

CENTER FOR ECONOMIC AND POLICY RESEARCH

# Response to Vanderbilt University's LAPOP Critique of CEPR Report, "Have US-Funded CARSI Programs Reduced Crime and Violence in Central America?"

By David Rosnick*

January 2017

Center for Economic and Policy Research
1611 Connecticut Ave. NW
Suite 400
Washington, DC 20009

tel: 202-293-5380
fax: 202-588-1356
www.cepr.net

* David Rosnick is an Economist at the Center for Economic and Policy Research (CEPR), in Washington DC.

# Contents



## Acknowledgements

The author would like to thank Mark Weisbrot, Dan Beeton, Alex Main, and Rebecca Watts for helpful comments and editorial assistance.

# Executive Summary

This report is a response to Vanderbilt University's Latin American Public Opinion Project (LAPOP) critique of our report, "Have US-Funded CARSI Programs Reduced Crime and Violence in Central America?" released in September 2016. That September report was an examination of the only publicly accessible impact assessment of USAID-funded anticrime and community-based violence prevention programs carried out under the umbrella of the US State Department's Central American Regional Security Initiative (CARSI). LAPOP took issue with our illustration of certain methodological flaws in LAPOP's study, as well as with the manner in which we presented our conclusions. LAPOP's criticisms appear to be largely based on misunderstanding and misinterpretation of our arguments and fail to address our main findings. The problems with the LAPOP study that we identified still stand, as does the validity of our conclusion: LAPOP's study cannot support the conclusion that intervention caused the areas subject to treatment in the CARSI programs to improve relative to those areas where no intervention took place.

# Response to Vanderbilt University's LAPOP Critique

We are pleased to see that — *in the interest of dialogue* — Vanderbilt University's Latin American Public Opinion Project (LAPOP) chose to reply to our critique of their 2014 study, "Impact Evaluation of USAID's Community-Based Crime and Violence Prevention Approach in Central America: Regional Report for El Salvador, Guatemala, Honduras and Panama." Given the fact that the criticism doesn't directly address our main findings, it is possible that our piece was not entirely clear. LAPOP nowhere makes clear reference to the respondent-level analysis that forms the basis of our conclusions; instead, LAPOP seems focused on our *illustration* of the problem of mean reversion.

We will respond here to every concern we could identify in LAPOP's response. First, however, we will summarize two important and related concerns regarding LAPOP's study.

Primarily, there exists in the data an unfortunate noise pattern whereby indicators in each arm of the study (treatment and control) do appear — pretreatment — to be dissimilar.[1] While this imbalance does not imply that neighborhoods were assigned nonrandomly, it does mean that one may not take the observed treatment effects at face value.[2]

Critically, treatment *appears* to be most effective in municipalities where treated neighborhoods were — relative to their corresponding control neighborhoods — particularly unhealthy in their pretreated state. Likewise, treated neighborhoods are observed to be made worse off in municipalities where treatment neighborhoods were particularly healthy pretreatment — relative to corresponding control neighborhoods. Where treatment and control neighborhoods were initially most similar, treatment appears to have no discernable effect whatsoever.

This is to say:

1. LAPOP's method may be *internally* valid, though the treatment "effect" they so label may be indistinguishable from mean reversion. *Given* the sample and the assignment, the model determined that the *given* "treatment" neighborhoods grew healthier on a range of indicators when compared to the changes in the given control neighborhoods. But — qualitative

1 See Appendix for details
2 Deaton and Cartwright (2016).

analysis notwithstanding — it is highly dubious to attribute the differences in differences primarily to the Central America Regional Security Initiative (CARSI) interventions.

2. LAPOP's approach lacks *external* validity because the results depend critically on study imbalance. The data strongly indicate that repeated experiments based on new samples — or at least on new assignments — would show that the reported effects are vastly overstated, if at all distinguishable from zero.

Over 14 municipalities and 16 main indicators of interest, we have a total of 288 pretreatment differences. These indicators are not independent, but overwhelmingly the pretreatment differences between treatment and control neighborhoods of a given municipality moved toward zero during treatment. In the 28 cases in which the difference grew more pronounced, the pretreatment differences were comparatively smaller by about a factor of four.

**TABLE 1**

**Pretreatment Differences and Mean Reversion[3]**

| Pretreatment Difference | | Average Change in Difference | Number of Instances |
|---|---|---|---|
| Sign | Average Difference | | |
| Treatment<control | -2.1 (1.0) | -7.1 (1.7) | 13 |
| | -9.4 (1.1) | 9.2 (0.9) | 75 |
| Treatment>control | 11.1 (0.8) | -11.7 (0.9) | 121 |
| | 3.3 (0.7) | 3.7 (0.6) | 15 |

Source and notes: Standard errors in mean in parentheses. Author's calculations based on data supplied by the LAPOP Project at Vanderbilt University.

**Table 1** strongly suggests that over the treatment period, neighborhoods regress toward the mean.[4] If the study concerned a sufficiently large number of clusters, then this *might* not be a serious problem. The apparent hypereffectiveness of treatment in relatively unhealthy neighborhoods might be balanced by hypoeffectiveness of treatment in relatively healthy ones. However, within each municipality, the variance in differences between each arm is large, and therefore requires study of many municipalities to safely put aside such concerns. The reported treatment effect may even depend more upon the random assignment of very few outlying neighborhoods than on the actual treatment.[5]

3 Note that we are reporting here average differences in the averages for the 16 indicators. Some indicators rise with health, while others fall. Thus, treatment less than control (pretreatment) neither means that treatment neighborhoods were initially healthier nor means that they were less healthy. We merely observe that changes in the differences correlate negatively with the initial differences.

4 Regression to the mean is utterly predictable. Taking again the robberies model as an example, if we have two neighborhoods that are otherwise indistinguishable, but there happens to be an unusual pretreatment crime wave in one, then we should expect respondents to report acute awareness of such crimes. However, as the crime wave passes, we would expect — on average — for the neighborhoods to become more similar and for respondents to report appropriately.

5 Deaton and Cartwright (2016).

Indeed, if the study had been balanced, regression to the mean would not necessarily be a problem. Of course, it would not necessarily require a difference-in-difference (DID) approach — one could measure the effectiveness of treatment directly based on the average post-treatment difference between the arms. But the *apparent* varying effectiveness of treatment with respect to the pretreatment differences confounds even the DID approach to an unbalanced sample. The data tells us that the critical parallel paths assumption is not valid — that LAPOP was unlucky in assigning (on balance) healthier neighborhoods to the control arm.

It is *no surprise*, then, that in a study where treated neighborhoods were — on average — less healthy than the corresponding control neighborhoods, one finds that — on average — treated areas improved more than control areas. The estimated average treatment effect is negative because the specific assignment introduced a negative bias.[6] But the result has no external validity because we would likely find the opposite result if all the neighborhood assignments had been switched.

In the same way, we should not infer that a scale has inherent imbalance if we randomly place a one-ounce weight on the left and a two-ounce weight on the right. We know that if the weights had been switched, the result would have been the opposite.

What can be done to establish external validity? One approach, following, e.g., Deaton and Cartwright, is to perform postexperimental stratification. That is, we may estimate entirely separate treatment effects by municipality and appropriately average the disaggregated estimates consistent with the external environment. In our case, we are interested in the expected average effect of treatment given perfect balance in the pretreatment indicator. Essentially, we're interested in estimating not the average of the effects in, say, Figure 5 of our original report, but the intercept.[7]

Why the intercept? The point of a randomized controlled trial is that one wishes — but is unable — to both treat and not treat the population of interest and therefore randomization assists in establishing balance between the arms. In expectation over many samples and assignments, it would be as if we had both treated and not treated the entire population. By construction, this would achieve perfect balance. But the best we can do is estimate the average effect on a balanced sample

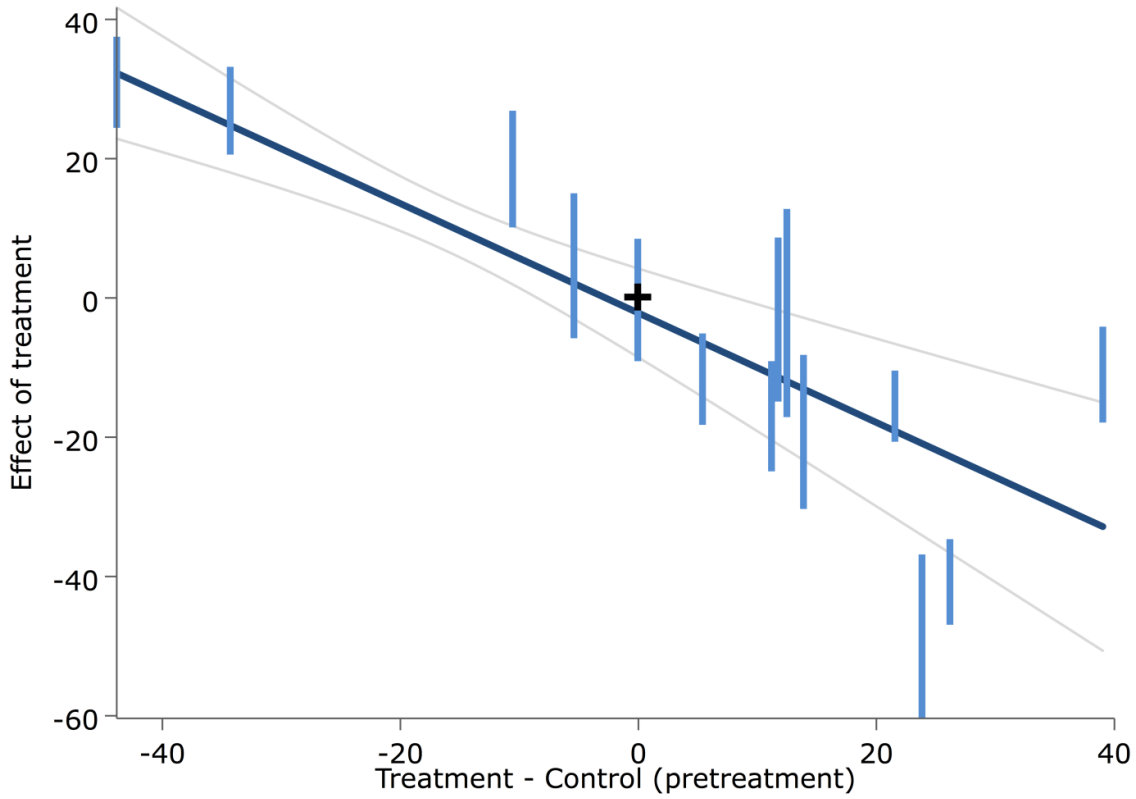
6 The sample average treatment effect (SATE) is an unbiased estimator of the population average treatment effect (PATE) when considered preselection and preassignment. We could therefore take the measured effect to be unbiased, but this would require us to add to the original uncertainty estimate the variance due to selection and assignment. That is, in putting forth the SATE as an estimate of the PATE either bias or variance must be added. It is certainly possible that for a given experiment either adjustment would turn out to be small, but as we illustrate here the data suggests the required adjustment critically affects LAPOP's conclusions.

7 Note that the municipal-level estimates in Figures 5 and 6 of our original report are not independent of each other, being based on a single regression for each indicator. In the stratification approach we use here, we independently estimate each effect.

where pretreatment differences are zero. Hence, for purposes of external validity, we are interested in the estimated intercept — not the *sample* average treatment effect.

In **Figure 1**, we show, graphically, this approach. The 14 municipal-level 95 percent confidence intervals are shown in light blue. The regression line through these intervals — inversely weighted by square of the interval width — is shown in dark blue with grey lines indicating 95 percent robust intervals. The black plus marks the origin. If the origin had lain outside the regression bounds, this would have provided evidence in favor of an average treatment effect on a balanced assignment.

**FIGURE 1**
**Stratification Estimation (Robberies Model)**




Source: Author's calculations based on data supplied by the LAPOP Project at Vanderbilt University.

In **Table 2**, we show the results of using stratification to estimate the balanced-sample average treatment effects.

**TABLE 2**

**Poststratification Estimates of Treatment Effects**

| Indicator | Mean Pretreatment Difference | Mean Change in Difference | Estimated Treatment Effect[1] LAPOP[2] | Stratified[3] |
|---|---|---|---|---|
| vicbar1arr | 5.1 | -5.9 | -7.9 (1.4)*** | -2.2 (2.9) |
| vicbar3arr | 0.6 | -3.8 | -7.3 (1.2)*** | -1.8 (2.6) |
| vicbar4arr | 9.5 | -9.1 | -10.2 (1.1)*** | -0.4 (1.3) |
| vicbar7arr | 8.3 | -13.9 | -17.3 (1.3)*** | -7.8 (4.4)# |
| pese0r | -0.8 | -1.7 | -2.0 (0.7)** | -2.6 (0.9)* |
| fear4r | 4.8 | -4.6 | -6.0 (0.9)*** | -0.6 (1.2) |
| diso7r | 2.5 | -1.8 | -5.4 (0.9)*** | 0.2 (1.2) |
| diso8r | 3.2 | -5.2 | -8.8 (0.9)*** | -2.3 (0.7)** |
| diso18r | 1.7 | -3.5 | -6.9 (1.0)*** | -1.6 (1.2) |
| fear10rr | 13.3 | -12.1 | -19.9 (1.4)*** | 0.3 (1.8) |
| soco9r | 0.2 | 2.9 | 6.7 (0.9)*** | 3.3 (1.5)* |
| b18r | 0.2 | 0.3 | 3.8 (0.9)*** | 1.1 (0.9) |
| pole2r | -0.3 | 1.1 | 2.1 (0.7)** | 0.7 (0.9) |
| it1r | -0.3 | 2.3 | 1.9 (0.8)* | 1.1 (0.9) |
| pn4r | -2.6 | 2.1 | 3.0 (0.7)*** | -0.3 (0.9) |
| n11r | 2.0 | -1.4 | -0.5 (0.8) | -0.2 (0.6) |

Source and notes: Estimates are disaggregated by municipality-level signed difference between treatment and control in pretreatment indicator means. The unweighted averages of these municipal-level differences are shown in column 1.

1 Standard errors in parentheses.

2 "Published" model.

3 Municipal regressions include "published" individual-level controls and random intercepts for each arm. Poststratification effect is estimate of intercept based on regression of municipal-level estimates of the effect against pretreatment differences in the indicator of interest. Municipal-level observations are weighted by the inverse of the square of the standard error of each effect estimated. The final-stage regression uses robust standard errors.

# Significant at 10 percent

* Significant at 5 percent

** Significant at 1 percent

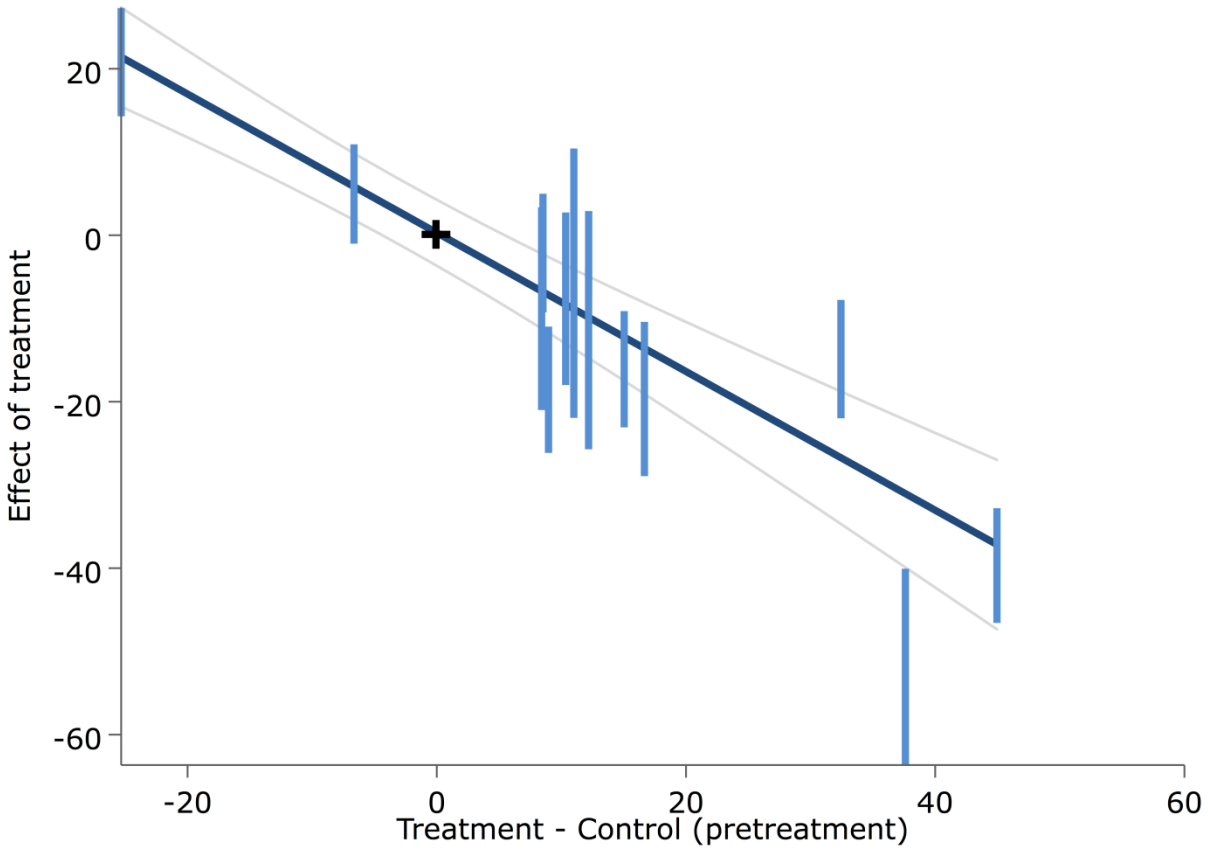
*** Significant at 0.1percent

Author's calculations based on data supplied by the LAPOP Project at Vanderbilt University.

The estimated effects are much smaller than those from LAPOP's models. The starkest difference is for fear10rr (avoid walking in dangerous areas) as we see in **Figure 2**.

**FIGURE 2**

**Stratification Estimate (Avoid Walking in Dangerous Areas Model)**

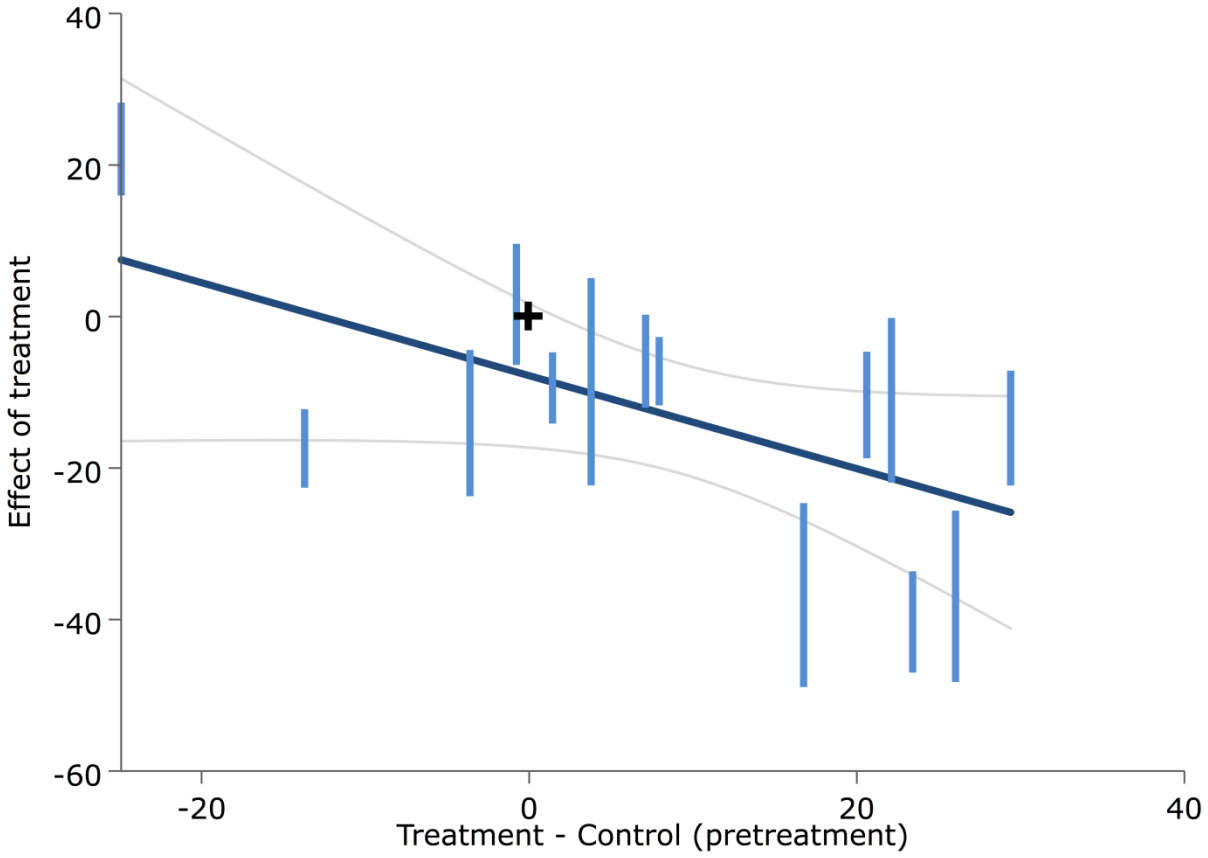



Source: Author's calculations based on data supplied by the LAPOP Project at Vanderbilt University.

Rather than a large negative effect both economically and statistically (with a coefficient 13.83 standard errors below zero of -13.83), stratification results in a very small and statistically insignificant effect (p-value of 0.87). This is no surprise, given that the largest imbalance lay there; 12 of the 14 municipalities had average fear10rr elevated in treatment relative to control. Despite the large sample average treatment "effect," we cannot expect the result to generalize.

The largest effect was for vicbar7arr (murder). However, the uncertainty in the estimate was also large.

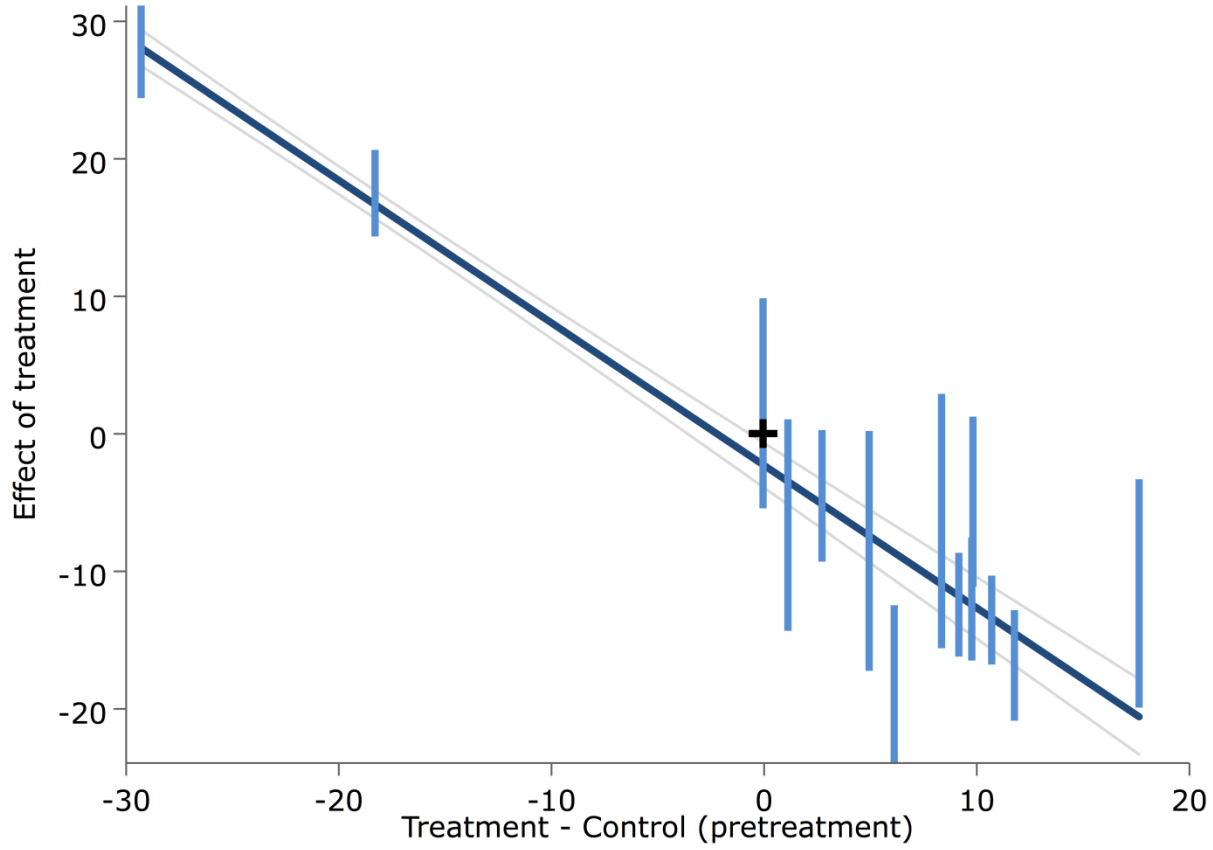
**FIGURE 3**
**Stratification Estimation (Murder Model)**




Source: Author's calculations based on data supplied by the LAPOP Project at Vanderbilt University.

On the other extreme, the municipal-level effects for diso8r (youth in gangs) provide the narrowest range among the regression estimates. This provides some evidence that we might expect treatment to be effective even if the effect is only a quarter of that suggested by LAPOP.

**FIGURE 4**
**Stratification Estimation (Youth in Gangs Model)**




Source: Author's calculations based on data supplied by the LAPOP Project at Vanderbilt University.

However, even if treatment is ineffective, we should — nearly 20 percent of the time — find spurious results in at least 2 of 16 indicators. This does not mean that treatment did not influence youth in gangs. It means only that we can cannot ex post point to the isolated results as convincing. Across LAPOP's chosen indicators, there is no reliable evidence that interventions improved community health. Instead, assignment imbalance and reversion to the mean conspired to make it appear that treatment was — on average — effective.

We now turn our attention to the specifics of LAPOP's response to our paper. To begin:

> Based on his email, we expected to engage in further dialogue with Rosnick and his team.

Back in February 2015, we raised with LAPOP our concern that their method might not suffice to distinguish temporary pretreatment differences from effective treatment. It was not obvious in July that our report would almost entirely focus on this very matter; given that LAPOP had already responded privately it seemed unnecessary to then ask for additional comment.

Obviously, we are still engaged in dialogue; rather, LAPOP seems concerned about maintaining *private* dialogue. To wit:

> We are disappointed, therefore, that Rosnick and his team shortly thereafter published their critique without giving us the opportunity to comment on their work. Such a dialogue would have provided us an opportunity ….

Obviously, LAPOP does not need us to provide an opportunity to comment. Public dialogue is ongoing. Again, LAPOP means here an opportunity to discuss further in private, which was never lost and has in fact informed this reply. Had LAPOP chosen — prior to publication — to avail themselves of that opportunity to better understand, we believe we would have succeeded in clearing up much confusion.

> RMJ [the CEPR study authors] misrepresent our research …. This error is not trivial, for two reasons. First, our study was longitudinal, aiming to examine impact over a range of years rather than the typical one-shot evaluations that dominate the field. We were looking for impact *over time* as the central driving element of our research design …. It is hard to imagine how RMJ could have missed or discounted this design feature. [emphasis in original]

It is "hard to imagine" that we missed this point precisely because we did not miss it. Though we might have better described LAPOP's period of study, they seem to read far too much into it. Theirs is not a criticism of our work based on a fair reading of our report. It would be very strange indeed for us to discuss pretreatment responses if we did not understand that surveys were conducted both before and after treatment. It is not clear how LAPOP inferred from our work that we discounted this element.

> [W]hile RMJ do note that we based our conclusions on a study of 127 communities in four countries, they fail to note that the conclusions were based on an unusually large sample involving 29,000 respondents.

This point is, first, incorrect. We laid out in Table 1 of our report the number of pretreatment respondents — a fair reflection of LAPOP's survey size.

More importantly, the survey size is misleading. If out of necessity imperfectly executed, LAPOP performed not simply a randomized controlled trial (RCT) but a longitudinal clustered randomized controlled trial. That is, individuals were not selected independently from a broad population, but from select communities in select municipalities in select countries. Rather than assigning individuals at random to treatment or control, assignments were geographic; respondents in the same neighborhood were all placed in the same arm of the study. Some of these neighbors were surveyed before — and some after — treatment.

Because neighbors are not independent, clustering reduces the effective sample size of the study. The more that clusters account for residual variation in the data, the more the sample size reflects clusters and not individuals. This makes sense: if individuals within any neighborhood are all basically the same, then we are sampling neighborhoods and not individuals.

Accounting for this "design effect" can be very complicated.[8] Stata add-ons are available to estimate these effects for relatively simple experiments, but these do not incorporate into Stata's built-in routines.[9]

An examination of 14 municipalities in 4 countries is likely to yield a false positive even if thousands of people are surveyed in each municipality. If in nine municipalities, LAPOP just happens to assign treatment to areas that improved more than control, this may be bad luck. The "unusually large sample" greatly increases the chances that we find a false positive.

In short, the above stratification exercise suggests LAPOP's "unusually large sample" is misleading, because LAPOP used it to precisely measure the confounding imbalance in assignment.

Moving on, LAPOP proceeds with several criticisms of style rather than substance:

> A further problem with the RMJ study is its lack of transparency. We pride ourselves on the transparency of our research and posted online the study's datasets and corresponding computer code for anyone to download and analyze, which is precisely what RMJ did *[sic]*. In contrast, when RMJ published their report they provided no link to their analysis and replication code for us and others to examine.

It is true that we have not published our code. However, neither had LAPOP made any request for code. In fact, LAPOP asked nothing at all of us prior to publishing their response.

Apart from that, LAPOP's code availability was — by their own standard — not at all transparent. For a year following the release of their report, they denied requests for data. This all but ensured that any criticism would be less than timely. This is not to say that they should have released data at the time they published. Rather, that (by their standard) they should have delayed publication of their report until the data was ready for release.

---

8 See, for example Cunningham (2010).

9 See Hemming and Marsh (2013). To get an idea what kind of "design effect" might be in play in LAPOP's study, we note that they report a standard error of 1.4 on the treatment effect in the robberies model. This means they ought to be able to detect a 3 percentage point difference in the post-treatment rates. If we — for simplicity — assume 9,000 individuals surveyed post-treatment divided among 125 neighborhoods, we have an average cluster size of 72. If the intracluster correlation (ICC) is only 10 percent, then detecting a difference in reported robbery between 40 percent and 37 percent requires 466 neighborhoods per arm — more than 67,000 total respondents.

Even after criticizing CEPR for not posting code on our website, LAPOP had not yet posted their code. Five days after their reply, and approximately 90 minutes after we pointed out this oversight to LAPOP, they quietly uploaded some of their code on September 21, 2016.

As for our data, we used none beyond the censored set eventually provided by LAPOP. Importantly, by the terms laid out by LAPOP for data access, we were "agreeing" to "not independently distribute, electronically post, or otherwise make the data available to any third party ..." making it impermissible for us to meet their proposed data-sharing standard.

We therefore find LAPOP's concerns regarding transparency extremely odd:

> Without this information, their claims cannot be fully verified.

If LAPOP had asked for the information — as we had in February 2015 — they might have found verification possible in a timely manner. We have shared our code, but as far as we can tell, LAPOP has made no serious effort to verify our results. Regardless, LAPOP sets for others a strangely high standard (full verification) given that the data they shared permitted only a partial replication of their results.

> RMJ also fail to note in their executive summary, conclusions, or press release the mitigating language used in their more detailed analysis. For example, in reference to their own analyses they state (p. 9):
>
> > *Though this analysis cannot rule out the possibility that there is no effect from intervention, the sample size has been reduced greatly from the thousands surveyed. The test may simply lack the power to detect a small effect.*
>
> In other words, using their approach, which reduced the sample size dramatically, they are unable to conclude that there is no effect (positive or negative) of the CARSI community-based crime and violence prevention approach. We are troubled that RMJ do not highlight these limitations to their approach in their summary conclusions.

LAPOP takes our quote utterly out of context. They take as our argument the early low-power tests. We performed these tests to explain more clearly our high-powered analysis that followed (with sample size identical to LAPOP's). By shifting focus away from the analysis that leads to our conclusions and toward the descriptive, LAPOP misleads.

LAPOP also takes issue with, as they put it:

> qualifying language was buried in the accompanying CEPR press release:

> *The paper notes that in some treatment areas, "Statistically, the possibility that intervention had no effect on reported robberies cannot be ruled out."*
>
> In other words, rather than critique our overall conclusions, CEPR's report delimits sharply its broader, headline-grabbing statement, "Study Doesn't Show that Areas Subject to Treatment in CARSI Programs Have Better Results," to a comment on a subset of treatment areas and, principally, one particular outcome indicator.

First, LAPOP makes a very strange argument here — that we buried language in a short press release. If we wanted to bury language, to put it in a press release would be self-defeating. Rather, this "qualifying language" is exactly the main finding of our report: LAPOP's quantitative approach is deficient and does not in fact rule out that intervention had no effect. Their data shows simply that some treated neighborhoods improved relative to control; others worsened. They provide no statistical evidence that the average effect of intervention was the result of anything but luck. Our analysis suggests that had the control areas been treated in lieu of those actually treated, LAPOP's approach may well have led to the (equally invalid) inference that treatment was harmful.

Second, we did indeed take considerable space "on one particular outcome indicator" just as LAPOP focuses on 16 out of a good many more than they might have. We chose robberies as an example simply because it was the first indicator LAPOP put forth. However, we found (and reported) that our criticism held broadly (see Figure 6 of our report). These indicators are obviously related; laying out a lengthy exposition of every indicator (rather than a graphical summary) would be of no interest to the general reader.

We have no idea what LAPOP means by reference to "a subset of treatment areas"; we employed the entire data set as provided us by LAPOP.

> Perhaps most telling about the weakness of the RMJ critique is that their conclusion makes only the weakest of claims. They do not refute our evidence, but merely claim that they cannot "rule out that the intervention had no effect." Given the low power that their statistical analyses have, as they themselves note in the report, they also cannot confidently rule out that it did have an effect.

LAPOP — citing low power — again appears to present background discussion as our final critique. Based on our "longitudinal" individual-level analysis — derived from LAPOP's model — we do safely conclude that LAPOP fails to provide reliable evidence of effective treatment. Of course, the interventions are bound to have some — if possibly tiny — effect one way or another, but our analyses tell us that it is not clear, based on the data, whether such interventions are helpful or harmful.

> The CEPR report repeatedly points to the "nonrandom" nature of the selection of treatment vs. control communities …. Vanderbilt *randomly* selected treatment and control communities, and told USAID where their community-based interventions should occur (treatment) and where they should not (control).

The basis for our claim was Honduras. As LAPOP states in their own response:

> *In the case of Honduras*, where our study was delayed by the 2009 coup, USAID had already selected the treatment communities by the time we were ready to begin, so *random selection was no longer possible*. [emphasis added]

So, per their own defense, LAPOP selected randomly in only three of the four countries. That LAPOP aimed to account for this in their analysis is irrelevant — they concede the point. Nevertheless, our conclusions do not depend on any assumption that selection was nonrandom.

Howsoever it came about that the majority of "an unusually large sample" of respondents lived in municipalities where treated areas were worse (prior to treatment) in comparison to control, this poses a problem for the LAPOP methodology. Whether such a bias came about because of simple bad luck (as our original Table 1 suggests possible) or as a consequence of the stipulated nonrandom selection is to a certain extent irrelevant. Even if the sampling of communities was of sufficient size to eliminate reasonably bad luck, and there was no bias in the nonrandom selection in Honduras, the fact is that the unit of analysis in LAPOP's test was clearly skewed. Our analysis suggests that insufficient accounting for this imbalance misled LAPOP to false certainty in their result.

In response, LAPOP cites Campbell and Stanley (1963) — arguing that their survey was well designed and that random assignment counters the problem of reversion to the mean. However, as we discussed above, random assignment works in this regard by creating balance — in expectation. If random assignment fails for whatever reason — luck included — to provide the necessary balance, then reversion remains a problem. The clustered assignment of respondents to control or treatment reduces the effective sample size, making imbalance more likely than in an experiment with a large number of independent observations.

> What the RMJ report fails to recognize in its discussion is that once a set of at-risk municipalities had been identified non-randomly, the selection of treatment and control communities *within* those municipalities was indeed conducted randomly. In other words, it seems RMJ confused "municipality" with "community."

We do not discuss at all the selection of municipalities except to note that there were not many; we indeed discuss the selection of areas for treatment within each municipality. We see no obvious basis for LAPOP's suggestion that we have been confused on this point.

> In sum, on the issue of this supposed "main problem," RMJ are simply wrong. We did randomize our selection of communities in Guatemala and El Salvador and did a propensity score matching in Honduras. In Panama, where we also did random selection, the number of communities that were eventually treated was too small for us to report individual country results, so we provided regional results with Panama included and excluded. The RMJ "main problem" critique on this point is without substance.

The word "propensity" never appears in LAPOP's regional report, and "score" appears only twice, and both in unrelated contexts.[10] However, in the Honduras report, LAPOP writes:

> Vanderbilt selected the control communities based upon a careful 'propensity score matchin' [*sic*] technique in order to minimize any important differences between treatment and control. The decision on the part of UAID [*sic*] not to allow random selection of treatment communities does open the door to the possibility that there may be something about those selected communities that is systematically different from a randomly selected control group. For example, the selected treatment municipalties [*sic*] could be ones in which USAID somehow knew were likely to more successfully absorb the interventions, or, on the contrary, were especially difficult "nuts to crack." Since we do not know if either or neither of these factors was used in the selection decision, all we can do is to note, here at the outset of our findings, *that non-random selection of treatment communities does create the possibility for selection bias that could weaken or undermine the results of the evaluation*. As noted, our selection via *propensity score matching minimized but did not eliminate this risk*, as is explained in the body of the report. [emphasis added]

LAPOP offered a pretty lengthy disclaimer for something of "no substance." Just because LAPOP noted that nonrandom selection was a problem that needed addressing, and they did what they could to address it, selection bias is by their own admission still a problem. However, as we have repeatedly pointed out, the LAPOP analysis fails even if we assume the assignments had been effectively — if not completely — random. Even if nonrandom assignment could be accounted for so as not to introduce any expected bias, luck remains an important factor. We made efforts to make this clear to LAPOP prior to the release of their response.

> The first issue is the level at which the data should be analyzed. RMJ chose to aggregate and analyze the results at the level of the municipality. There is no justification or foundation for doing this.

10 Berk-Seligson et al (2014a, 2014b).

Again, LAPOP ignores that we based our conclusions on extension of LAPOP's own model. Our discussion of municipal-level data was meant to illustrate the underlying problem of reversion, which may be difficult to describe at the individual level. Our conclusions relied on additional municipality controls, but our model was based on individual observations — the same as LAPOP's.

> The LAPOP study was specifically designed *not* to draw inferences at the sub-national level. The number of municipalities was simply too small to do so.

Again, we make a point of this in discussing municipality-level data, which is why we do not stop with illustration. However, there is nothing small about the samples within each municipality to bar separate analyses for each. This would be like what we did in the higher-powered analysis producing Figures 5 and 6. Just as LAPOP employed (constant) random geographical effects in their model, we employ a full set of interactions at the municipal level (the lowest level of geography available to us in the data).

> Subsetting the sample into municipal sub-units could be misleading because the small sample could produce "Type II errors" (i.e., failing to detect significant effects that actually exist in the data). We suspect this is part of the reason RMJ report impacts that are not statistically significant.

To the extent that LAPOP's regional sample is very large, the samples in each municipality are hardly small. We *do* detect many significant municipal-level effects based on LAPOP's approach. Our concern is not the significance of these effects, but the significance of the relationship between pretreatment differences and the estimated effects. If sample size is a problem for these municipalities, it is surely a problem for the regional whole as well.

Indeed, the LAPOP response states:

> With only a small number of communities sampled in each municipality (fewer than 10, on average), it is not surprising that the pretreatment averages for each group are different.

This is precisely so. The problem is that these differences are large and do not balance out over the regional sample given the small number of *municipalities*. LAPOP's conclusions appear to be a misinterpretation of the statistical results derived from failure to recognize the importance of these imbalances.

> The treatment is considered effective if outcomes in the treatment group improve over time significantly more (or decline less) than they did in the control group. One of the great advantages of this approach is that it does not require that the treatment and control groups

> have identical starting values, a basic fact that is not addressed by RMJ, since it is the *trend* in each group and the difference between those trends that is of interest.

This misrepresents our argument. Of course, DID does not require treatment and control to have identical starting values. But it must assume (implicit in the constructed counterfactual) that those differences persist from pretreatment to posttreatment — that is, that the parallel paths assumption holds. We find evidence that these differences on average tend to disappear over the period of treatment. Alternatively, one may argue that treatment has effects that vary based on the pretreatment differences, but a regional estimate of the average treatment effect require balance. Our above estimate of the average effect on a balanced sample is not statistically distinguishable from zero.

LAPOP supports their position in a footnote, saying:

> Another challenge to the DID is when the composition of treatment or control groups changes systematically over time. We analyzed the data for this, and found that the composition of the groups did not change on key observable variables (education, wealth, age, etc.) over the course of the experiment. In short, our study design minimized challenges to causal inference and we carefully examined the data and found no evidence of threats to inference.

We agree that the above control variables had little influence on the results. For us, the question is not whether the composition of the respondents changed over time, but something not modeled. Even if the population had been unchanged and the survey had been a complete census of the population, an unusually high number of actual robberies in one pretreatment area may lead to a fall in the reported robberies over the course of treatment relative to those in control neighborhoods. The observables listed above are less key than the pretreatment differences in the indicators of interest — precisely because those pretreatment differences do not balance out.

> Setting aside their erroneous assertion [read, LAPOP's declaration] of nonrandom assignment, this claim is based on a flawed approach …. The authors of the CEPR report do not establish an explicit counterfactual with which to compare outcomes in the treatment area. Our careful reading of their methodology suggests that they implicitly assume that the "normal" level of robberies is the average municipal level of robberies in control areas in the first time period (prior to the implementation of the CARSI intervention in the treatment areas). As we explain below, such a counterfactual lacks any basis in scientific rigor.

Setting aside that LAPOP's own report is the source of our "assertion," it is very clear that despite their "careful reading," LAPOP needs to revisit our report. In our full analysis, we assume there is a

common (counterfactual) change in both treated and untreated areas of each municipality. By employing full interactions, we effectively perform separate DIDs on each municipality using a single model. The underlying DID, however, is the same as that presented by LAPOP. If our counterfactual "lacks any basis in scientific rigor," then so does LAPOP's.

Furthermore, although it was not a fundamental point in our report, we did compare changes in treatment relative to control in our low-powered analysis as well. In Table A2, we saw that pretreatment rates did have effects on both treatment and control that *were* statistically significant, though columns 2 and 4 suggested that the effects were not statistically different between arms. That is, we failed to find that the arms converged or diverged when controlling for pretreatment effects. Again, this finding suffered from low power, but the point was to illustrate that the neighborhood data was not obviously consistent with zero mean reversion and therefore a potential problem. The extent of the reversion, combined with the extent of the imbalance in the study, confirmed our original suspicions.

Their Figure 2 reflects complete misunderstanding of our model; it is completely and totally wrong. Given that, there is no further point in addressing their arguments on the matter except to repeat that LAPOP seems — based on earlier comments — confused between our illustrative and final analyses.

> While RMJ focus primarily on a single indicator of crime and security in their critique — robberies — it is important to note that several of the indicators reported in the LAPOP study have starting values in the treatment and control areas that are statistically indistinguishable, including perceptions of insecurity, sale of drugs, and interpersonal trust.

Even if some indicators had starting rates statistically indistinguishable when pooled, this would not mean that they are indistinguishable when disaggregated by municipality or any other factor. It would be more convincing if the study had been balanced, which would not preclude any mean reversion as a more reasonable explanation of their results.

Further, all the indicators LAPOP points to in the above quote show the same general problem as robberies.

> Focusing principally on a single indicator is consistent with the tendency in the RMJ report to overlook quantitative and qualitative analyses in the reports and datasets that comport with the general conclusion drawn in our report.

For some reason, LAPOP repeatedly objects to our employment of a single indicator as an example for purposes of discussion. We choose to believe that this represents a simple oversight. Figure 6 (of our report) shows how our critique applies broadly.

While LAPOP's qualitative work might lead them to believe the assumptions in their quantitative analyses are valid, the quantitative findings are not externally valid.

Regardless, we use the data they provided and performed analysis of all 16 indicators based on their DID approach. For LAPOP to respond that we have overlooked their "reports and datasets" is baseless.

# References


Berk-Seligson, Susan, Diana Orcés, Georgina Pizzolitto, Mitchell A. Seligson, and Carole J. Wilson. 2014a. "Impact Evaluation: Honduras Country Report." Nashville, TN: The Latin American Public Opinion Project (LAPOP), Vanderbilt University. http://www.vanderbilt.edu/lapop/carsi-study.php

———. 2014b. "Impact Evaluation of USAID's Crime and Violence Prevention Approach in Central America." Nashville, TN: The Latin American Public Opinion Project (LAPOP), Vanderbilt University. http://www.vanderbilt.edu/lapop/carsi-study.php

Cunningham, Tina. 2010. "Power and Sample Size for Three-Level Cluster Designs." PhD dissertation, Virginia Commonwealth University. http://scholarscompass.vcu.edu/cgi/viewcontent.cgi?article=1147&context=etd

Deaton, Angus and Nancy Cartwright. 2016. "Understanding and Misunderstanding Randomized Controlled Trials." National Bureau of Economic Research Working Paper 22595. http://www.nber.org/papers/w22595.pdf

Hemming, Karla and Jen Marsh. 2013. "A Menu-driven Facility for Sample-size Calculations in Cluster Randomized Controlled Trials." The Stata Journal 13(1):114–135. http://www.stata-journal.com/sjpdf.html?articlenum=st0286

Latin American Public Opinion Project (LAPOP). 2016. "LAPOP's Response to David Rosnick, Alexander Main, and Laura Jung." Nashville, TN: The Latin American Public Opinion Project (LAPOP), Vanderbilt University. http://www.vanderbilt.edu/lapop/carsi/CARSI_Panama_v3_FinalV_W_02.17.16.pdf

Rosnick, David, Alexander Main, and Laura Jung. 2016. "Have US-Funded CARSI Programs Reduced Crime and Violence in Central America?" Washington, DC: Center for Economic and Policy Research. http://cepr.net/publications/reports/have-us-funded-carsi-programs-reduced-crime-and-violence

# Appendix

## Examination of Study Imbalance

There are any number of ways to consider the extent of study imbalance. For our purposes, the simplest way to check is to consider not the treatment effect in the LAPOP's DID regressions, but the assignment effect (the coefficient on treat_contr_grps). Note that such a test is not subject to the general criticisms we present here and in our paper; we are interested explicitly in the sample differences and not in a regional average difference (which must be, by construction, zero).

However, we have a multiple comparisons problem as well. If we consider 16 indicators, each with a 5 percent chance of spurious difference between treatment and control, then there is a better than 50/50 chance that we will find at least one "significant" result. To reduce the chance of spurious findings, we consider each individual test at about the 0.32 percent level. That is, our 5 percent critical value will lie not at the usual 1.96, but rather 2.95 standard errors from zero.[11]

**TABLE A1**

**Assignment Effects on Pretreatment Variables in LAPOP Models**

| Indicator | "Published" Effect |
|---|---|
| vicbar1arr (robberies) | 7.8 (3.67)** |
| vicbar3arr (drug sales) | 3.3 (1.74) |
| vicbar4arr (extortion) | 8.8 (5.94)*** |
| vicbar7arr (murder) | 6.8 (3.00)* |
| pese0r (perception of insecurity) | 0.4 (0.3) |
| fear4r (perception of insecurity when walking alone) | 4.2 (3.31)* |
| diso7r (youth loitering) | 4.9 (3.82)** |
| diso8r (youth in gangs) | 4.5 (3.22)* |
| diso18r (gang fights) | 4.0 (2.70) |
| fear10rr (avoid walking in dangerous areas) | 16.3 (8.89)*** |
| soco9r (community organized to prevent crime) | -2.7 (-2.42) |
| b18r (trust in police) | -3.5 (-3.75)** |
| pole2r (satisfaction with police) | -1.8 (-2.29) |
| it1r (interpersonal trust) | -1.0 (-0.96) |
| pn4r (satisfaction with democracy) | -4.4 (-6.52)*** |
| n11r (government handling of security) | 0.7 (0.9) |

Source and notes:
# Significant at 10 percent
* Significant at 5 percent
** Significant at 1 percent
*** Significant at 0.1 percent
Z-statistics shown in parentheses. Critical values (2.718, 2.948, 3.419, 4.003) are adjusted for multiple comparisons.
Author's calculations based on data supplied by the LAPOP Project at Vanderbilt University.

11 Note that this adjustment is conservative. The indicators are correlated in that neighborhoods that are unhealthy with respect to one indicator are probably unhealthy with respect to many. The individual tests are not truly independent; in adjusting as if we were running 16 independent tests, we have set the critical values too high.

Nine of the 16 indicators have assignment effects estimated to be at least 3 standard errors away from 0 — more than sufficient to warrant concern regarding the possibility of an unfortunate assignment. Note that for 15 of the 16 indicators — including all 9 of adjusted statistical significance and all 12 unadjusted — the treated neighborhoods were estimated to be less healthy before treatment than were the control neighborhoods. For only n11r (government handling of security) does the balance run the other way, but this difference is far from statistically significant ($p=0.366$). Perhaps not coincidentally, this was the only indicator of the 16 that LAPOP did not claim had a statistically significant improvement because of CARSI intervention.