Reply to Data Colada #101

Anuj K. Shah, University of Chicago Booth School of Business
Michael LaForest, Pennsylvania State University

First, we thank Data Colada for reviewing our paper. We appreciate this discussion and Data Colada's role in creating a constructive conversation among a community of people who care about getting empirical work right. We also appreciate the chance to provide our own thinking about some of the issues raised in this post since there appears to be some confusion.

1) We think the authors of this post overstate how much our analysis deviates from the pre-registration, and they understate the value of the robustness checks.

2) They also seem to confuse one of our robustness checks with multiple-hypothesis testing. As part of this, they report a misleading and irrelevant analysis.

3) They argue there is an unreported confound in the field experiment. But we in fact discuss it at length in our paper and supplement.

Before proceeding to these points, it might be helpful for readers to have context that this post doesn't quite make clear. Below is Figure 1 from our paper, which includes more than the data-point highlighted by this post's own Figure 1.

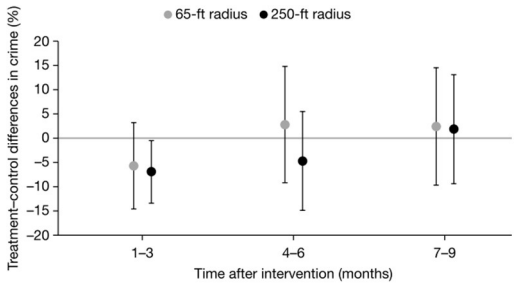


Fig. 1 | Treatment-control differences in crime after policing intervention. Providing residents with information about their neighbourhood officers reduced crime near housing developments in the first 3 months after the intervention. Changes in on-campus (grey) and near-campus (black) crimes reported post-intervention are shown, where $n = 39$ treatment developments and $n = 30$ control developments. Point estimates represent per cent change attributed to the intervention in crimes per resident per month (error bars represent 95% confidence intervals). Estimates are based on our primary specification (Supplementary Information C.3) for 1–3 months (on-campus, $P = 0.21$; near-campus, $P = 0.04$), 4–6 months (on-campus, $P = 0.68$; near-campus, $P = 0.36$), and 7–9 months (on-campus, $P = 0.70$; near-campus, $P = 0.74$) post-intervention. P values are from two-tailed tests.

And, of course, there is more in our paper than our Figure 1. Throughout our paper, we present our results as comprehensively as possible, with measured and appropriate interpretation.

1) Pre-registration and robustness checks

The two analytical choices the authors of this post focus on are (i) the distance and (ii) the duration over which to measure effects.

Regarding distance: We only pre-registered two distances and we report both side-by-side throughout the paper (see, e.g., Figure 1).

In fact, our pre-registration says: *"The primary research questions for this study, related to the impact of the field intervention on administrative crime outcomes, are: 1. What are the effects of the field intervention on crime on and near NYCHA developments? …"* (p. 2, underlining added here).

Confusingly, we later put the near-campus/250-foot analysis under the "exploratory" section. This is a contradiction we wish we had caught. While this was an editing error, it's an important one. So we presented results for both distances side-by-side throughout the paper, and we conducted additional robustness checks. Nevertheless, throughout the paper, there is an obvious, pre-registered focus on both on-campus (65-foot radius) and near-campus crime (250-foot radius).

The authors of this post are aware of this. We appreciate the chance to draw readers' attention to this, since the version of the post we reviewed does not mention this. And we disagree with how the authors of this post frame it as if analyzing crime at the two pre-registered distances over the same durations is as arbitrary as analyzing broken property rates or 311 calls at random intervals.

Regarding duration: In writing our pre-registration, we explain the *maximum* timeframe for the data, but we did not state that the primary outcome period was nine months (or that we expected the effects to persist for nine months) as might be implied in this post. Our estimating equation includes a time subscript that shows we will examine effects by time.

To assess fadeout, we divided the nine months of data into three, 3-month buckets. Although this is not in the pre-registration, we made the decision before analyzing the data.

Both of these issues (with outcome distance and duration) highlight the importance of robustness tests. Readers should ask: How sensitive are results to different analytical choices?

In our field setting, there are plausible countervailing forces for both time and distance from the intervention. One would expect stronger effects closer to the intervention (in time and distance), but less precision (there are fewer crimes observed in shorter durations and distances). Meanwhile, one would expect smaller effects farther from the intervention in time and distance, but more precision (because there are more crime observations).

An omniscient researcher (who understands these countervailing forces perfectly) could pre-register a cherry-picked sweet spot in the data (*a priori* cherry-picked, but cherry-picked nonetheless). A naïve researcher's pre-registration would be a pure guess. In either extreme, and everywhere in between, readers should expect to see robustness checks because any analysis over some time and distance is a snapshot of these countervailing forces.

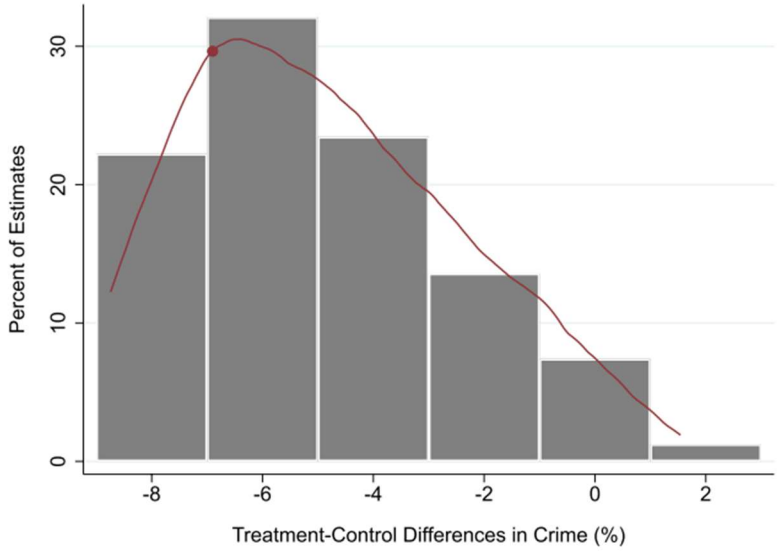
There is genuine disagreement in the social sciences about whether to only consider pre-registered analyses and about the value of additional analyses and robustness checks (see, e.g., Olken, 2015). In our paper, we present main results based on decisions made prior to the analysis

that hew closely to our pre-registration. We also believe it is useful to let readers see more of the data and how sensitive the analyses are to different specifications. In our paper and supplement, we thoroughly discuss these robustness analyses and try to convey a measured and accurate summary of the effects of the intervention.

2) Robustness checks versus multiple-hypothesis testing

As part of our robustness checks, we asked how sensitive results are to our specification choice of outcome period after the intervention and distance around housing developments. For instance, is the 3-month, 250-feet result an anomaly or is it in line with other analytic choices that one could have made?

For this exercise, we show two sets of robustness checks. First, Extended Data Figure 3 (reproduced below) shows that the effect size for one of our two main specifications is not unusual. Second, our P-value heat map (Extended Data Figure 4) further suggests that this result is not unusual. We discuss this at length in our supplement (Section C.3, pp. 17-20).



Extended Data Fig. 3 | Distribution of point estimates for treatment effect. As a robustness check, we conducted analyses for various radii ranging up to three blocks around developments: 65 ft., 100 ft., 150 ft., 200 ft., 250 ft., 300 ft., 400 ft., 500 ft., and 750 ft. And, for each radius, we conducted analyses for cumulative time intervals ranging from one month after the intervention (i.e., February 2018) to the first nine months after the intervention (i.e., February through October 2018). Varying both of these dimensions produced 81 sets of results, based on our primary specification applied to each radius and time interval (see Supplementary Information C.3). This figure shows the distribution of point estimates for the crime reductions across these analyses, along with an Epanechnikov kernel density function over the distribution. The red dot highlights where the 250-ft, 3-month result falls in the distribution, suggesting it is in line with the central estimates across all 81 analyses.

The robustness checks (and Extended Data Figures 3 and 4) simply highlight that the analyses in the paper are consistent with a variety of other specifications one could run.

This is what our robustness checks tell the reader: If our analyses are a snapshot of those countervailing forces between increasing and decreasing time and distance, this snapshot is consistent with other similar snapshots, which vary in their overlap.

We note this in our paper: "We also conducted analyses that vary the radius around the housing developments and the time interval after the intervention. Across these specifications, the point estimates of the treatment effect are quite consistent with our main results (Extended Data Figs. 3, 4), which appear to slightly understate the duration (over time) and reach (over distance) of the effects of the intervention" (p. 300). That is, we show that our main specifications are consistent with other plausible specifications one might run.

We <u>do not</u> present our robustness checks as saying: "Because you see 16 P's < .05 out of 81, you should be reassured that the effect is real." We neither interpret nor present the robustness checks that way. That would be more like multiple-hypothesis testing (e.g., H1: "If the intervention is effective, it will reduce crime at time A and distance B"; H2: "If the intervention is effective, it will reduce crime at time X and distance Y"; "Look, there is support for 16 of these hypotheses, we can reject the null").

The authors of this post evaluate our robustness checks using a permutation test. **We think this permutation test is misapplied.** It may be relevant for multiple-hypothesis testing. It does not seem relevant for the robustness checks.

Their test asks: How often would you see a similar count of P-values <.05 from chance alone? Again, this test is irrelevant because we never argue that the count of P-values implies the existence of an effect.

Still, we will try to offer some constructive thoughts in passing on how we think this test should be applied to situations where it is appropriate.

In psychology and the social sciences, there is a common problem of trying to think about how likely a pattern of outcomes might be from chance alone. The authors of this post conduct a test that shows the probability of getting a heat map "at least as reassuring as" ours *conditional* on the result from the main text. It is not clear to us exactly what setting such a test would be relevant for.

Instead, if the heat map were presented as a set of 81 outcomes and we pointed to the 16 significant P-values as signs of a true effect (again, we do not do this, but this is to illustrate the point about multiple-hypothesis testing), then it seems like a more straightforward test would be: "How often do you get data like ours *out of all 50,000 draws*?" In the authors' test, 841/50,000 resamples produce results like ours (i.e., a P-value for the focal test between .036-.07 and a heat map "at least as reassuring as" ours).

That is, "under the null of no effects whatsoever," the probability that you would see these results is .017. Also note that is not a P-value we would attach to our heat map because, again, we don't view that as a family of outcomes to be tested jointly.

The version of the Data Colada post that we reviewed states, "Indeed, our conservative simulation shows that*, under the null of no effects whatsoever*, the probability of observing a heat map that contains as many significant results as the authors observed is p = .43." (italicized in the original) This seems misleading. We think the usual way to describe such a permutation test would be to state what proportion of resamples have *an overall pattern of this type*. That probability is .017. It is not clear what type of test the .43 corresponds to, and it is difficult to see

how it corresponds to "under the null of no effects whatsoever" because the authors of this post have made a choice to condition on a subset of their resamples.

Alternative explanations for the field results

The final point the authors of this post make is that there is an unreported confound that doesn't come through clearly in the paper. We were surprised to read this because we pay considerable attention to confounds and alternative explanations in our paper and supplement. And we include the field experiment materials in our paper itself, since that is the clearest way to show what is manipulated. Indeed, in the very sentence from our paper quoted in the post, we direct readers to the figures showing the materials in the paper itself: "…their hobbies or why they became an officer (Extended Data Figs. 1, 2)" (p. 299).

We also begin with a series of lab experiments that offer clearer tests of mechanism. Testing mechanism in a field experiment is inevitably more difficult. There are space constraints in the paper, but in the main text we refer to tests of our proposed mechanism and tests of alternative explanations and confounds, which we describe in more detail in Supplementary Information Section C.6. To briefly summarize these tests, we find:

- The intervention shifts residents' perceptions of how much officers know about their illegal activity, and this predicts changes in crime. This effect may also be specific to what residents think officers know about *them* in particular, rather than what officers know about crime in general (Main text Fig. 2, SI Figs. S5, S6, SI p. 29).
- We do not find support for the "attentive cop" hypotheses that the authors of this post discuss. For instance, in our post-intervention survey, residents did not think that officers were around more or patrolling more often or were responding more quickly to calls (Main text p. 300, ED Table 3, SI pp. 29-31, Fig. S7, Table A5).
- The effect also does not appear to be driven by changes in officer behavior (SI p. 32, Table A6).
- Finally, it is also worth noting that even in the control condition, flyers are regularly posted in housing developments with the names and contact information of neighborhood officers (but not the personal information included in our intervention). This is something that NYPD does separately from the intervention we evaluated.

Field experiments are typically much better at answering the "what" than "why." And reasonable people can disagree about whether our tests have successfully isolated the mechanism. We think more work on these questions is warranted, and we hope our paper inspires such work. But we strongly disagree with the suggestion that we do not consider these alternative explanations or confounds. We consider these carefully in our paper and supplement.

We hope readers of Data Colada will actually read our paper. Skepticism is inherent in the scientific process and it drives work forward. But such skepticism is less informative when it is based on a skewed description of the work or misapplied analyses. We hope our reply adds some transparency to this post.

Reference

Olken, B. A. (2015). Promises and perils of pre-analysis plans. *Journal of Economic Perspectives, 29*, 61-80.