Comments on “Procedural justice training reduces police use of force and complaints against officers”

Jonathan Roth[*] Pedro H.C. Sant’Anna[†]

July 14, 2020

In light of recent events, including the prominent police killings of George Floyd and Breonna Taylor, improving relationships between police and community is an extremely important topic. There is clearly a large need for good research on effective ways of actually achieving this goal. We were therefore very intrigued by your recent paper, Wood, Tyler and Papachristos (2020), which finds large reductions in complaints against police and police use of force from a procedural justice training program. We also appreciate your efforts to post the data along with well-documented code that allowed us to replicate and extend your analysis.

After reviewing your analysis, however, we have some concerns about the statistical methodology used to analyze the impacts of the training program that you study. Given the extreme policy importance of this topic, we wanted to write to describe to you what we view as the potential issues, and to make sure that we are properly understanding what has been done in your analysis (apologies in advance if we have misunderstood!). In what follows, we provide a more detailed discussion of the potential issues we have uncovered.

The main dataset used for analyzing causal effects in Wood et al. (2020) has data at the month-by-cohort level, where a cohort is a group of officers who are each trained on the same date. There are $N = 328$ cohorts and $T = 63$ months of data. The outcome used is Y_{it}, which is the *total* number of distinct complaints for officers in cohort i. Importantly, as we understand it, the outcome Y_{it} is not normalized by the number of officers in cohort i. Causal effects are then estimated using the interacted fixed effects method of Xu (2017), which infers $Y_{it}(0)$ for treated cohorts using a factor model fitted to Y_{it} for not-yet-treated cohorts.

A possible issue with this approach, however, is that if the sizes of the cohorts differ,

[*] Microsoft Research. Jonathan.Roth@microsoft.com

[†] Vanderbilt University. pedro.h.santanna@vanderbilt.edu

then cohort-level totals for not-yet-treated cohorts may not serve as a good counterfactual for already-treated cohorts even if officers are randomly assigned to cohorts. As can be seen in Figure 1, cohorts treated earlier tended to be larger. Since the authors use a specification that residualizes the outcome against time and unit fixed effects, what will matter is the trends in the counterfactual outcomes. Figure 2 shows that the total number of complaints per officer is decreasing over time. Therefore, even if there were no causal effect of treatment, we would expect the total number of complaints to decrease more for larger cohorts. To make this concrete, if complaints per officer decreases by 0.02 between early and late periods, we would expect the change in total complaints to be 1 in a cohort of 50 but only 0.6 in a cohort of 30 if officers are randomly assigned to cohorts. However, the specification used by the authors estimates treatment effects by effectively comparing the trends in the totals for already-treated and not-yet-treated cohorts. If we expect the totals to decrease more for already-treated cohorts absent a treatment effect, this may produce spurious negative estimates.

To analyze how much these potential issues matter in practice, we compare the results in the paper to analogous results on synthetic data where we would not expect to find a treatment effect. In Figure 3, we plot the estimated average treatment effects on the treated (ATT) for complaints at each event-time using the authors' replication code. These ATTs are calculated using the felc package in R to estimate the interactive fixed effects model. These point-in-time ATTs are then aggregated to form a cumulative ATT, which is presented in Figure 4 in Wood et al. (2020). The figure shows negative treatment effects that are growing in magnitude over time. We then conduct the following falsification exercise. We create a new dataset in which the number of complaints per officer is the same for all cohorts in all periods. The number of complaints per officer in each period is chosen so that the total number of complaints (summing over all cohorts) in each period matches that in the original data. We then re-estimate the ATTs using the same methodology except replacing the actual complaint counts with the synthetic counts. We should not expect to find a causal effect, since the number of complaints per officer is the same for all cohorts in all periods. However, as can be seen in Figure 4, the estimated treatment effects on the synthetic data follow a very similar pattern to those estimated on the real data.

A simple way to address the differing cohort sizes is to use cohort-level averages, rather than totals, as the outcome. Figure 5 shows the results of re-estimating the model used by the authors replacing total complaints with complaints-per-officer.[1] The re-estimated model shows point estimates close to zero and statistically insignificant for nearly all event-times.

[1] We remove all observations for a single cohort that has no observations in some periods to avoid dividing by 0.

As a sanity check, we also re-estimate the model using complaints-per-officer on the synthetic data described above; reassuringly, the point estimates are all zero using the synthetic data (Figure 6).

We are thus led to conclude that the negative significant effects on the number of complaints is spuriously driven by differing cohort sizes. We note that this concords with a simple plot of the mean number of complaints by officer over time (Figure 7), which shows little systematic difference in complaints per officer between the already-treated and not-yet-treated cohorts. Finally, we note that we re-did the same analysis excluding the officers who resigned during the sample period, as in Figure S7 in the SI Appendix to Wood et al. (2020), and found similar results to those presented here.

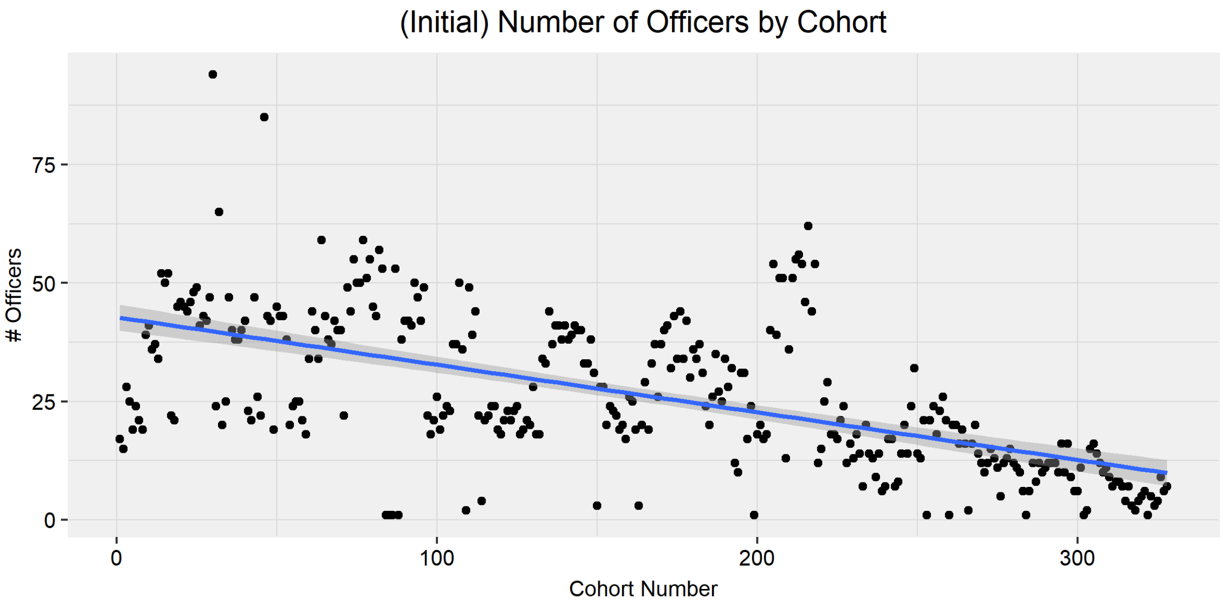
We would be very interested to hear from you whether we have properly understood the analysis in your paper, as well as your thoughts on the potential issues raise in this note. Thanks!

References

Wood, George, Tom R. Tyler, and Andrew V. Papachristos, "Procedural justice training reduces police use of force and complaints against officers," *Proceedings of the National Academy of Sciences*, May 2020, *117* (18), 9815–9821.

Xu, Yiqing, "Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models," *Political Analysis*, January 2017, *25* (1), 57–76. Publisher: Cambridge University Press.

Figure 1



Note: this shows the inital number of officers trained in each treatment cohort (or cluster) as well as a best-fit line.

Figure 2

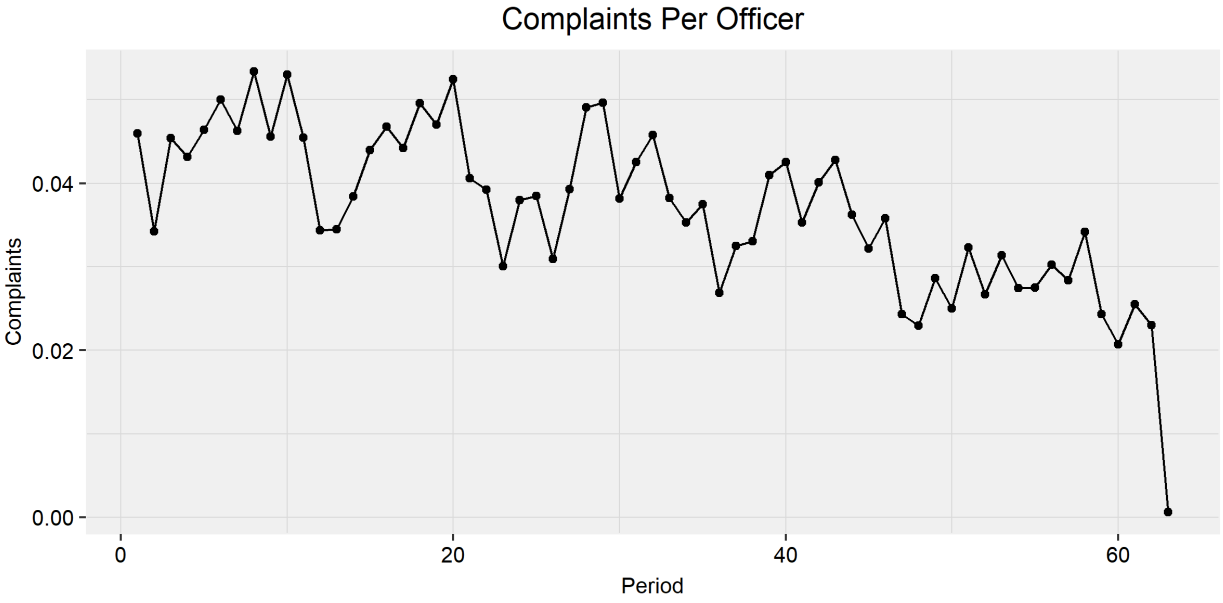
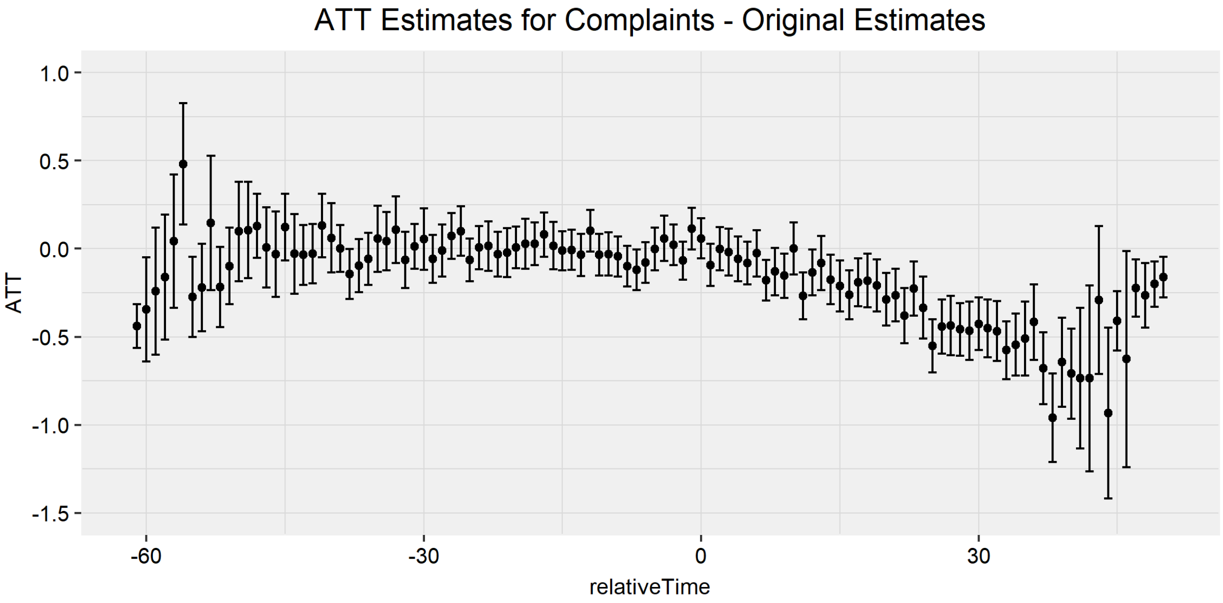
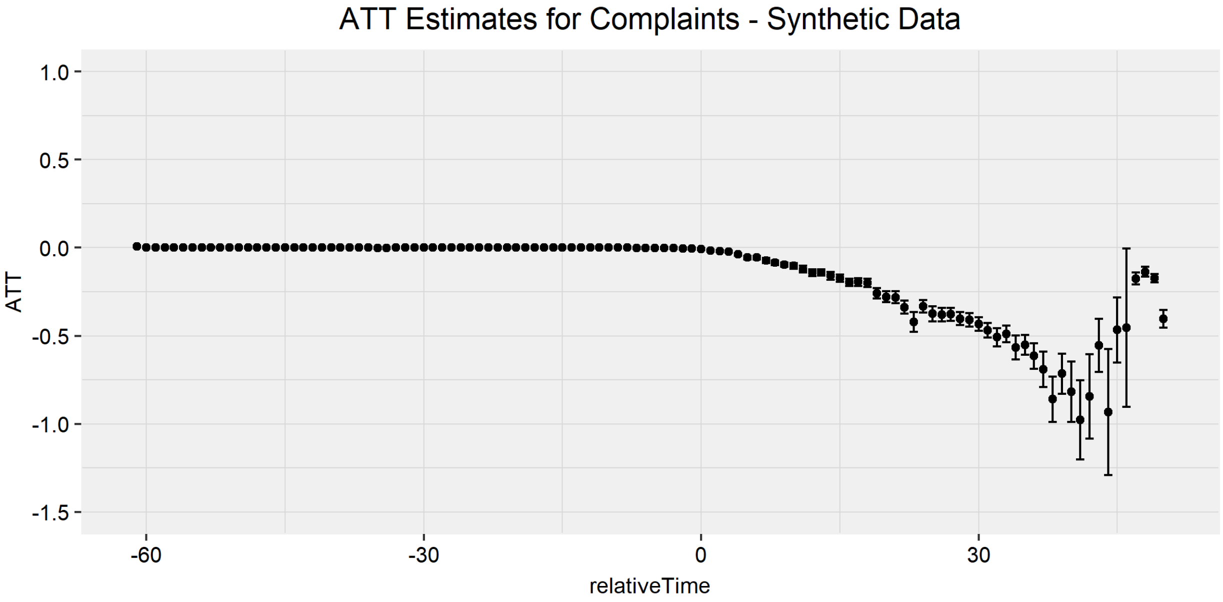


Figure 3



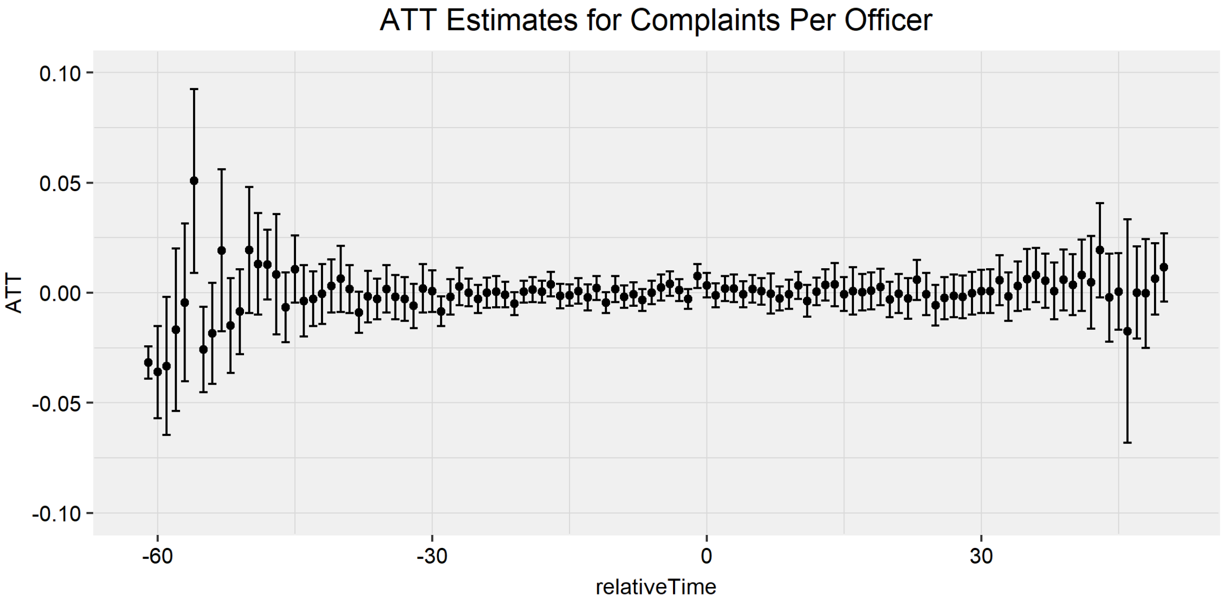
Note: this figure shows the ATT estimates from the felc package using the original specification used by the authors.

Figure 4



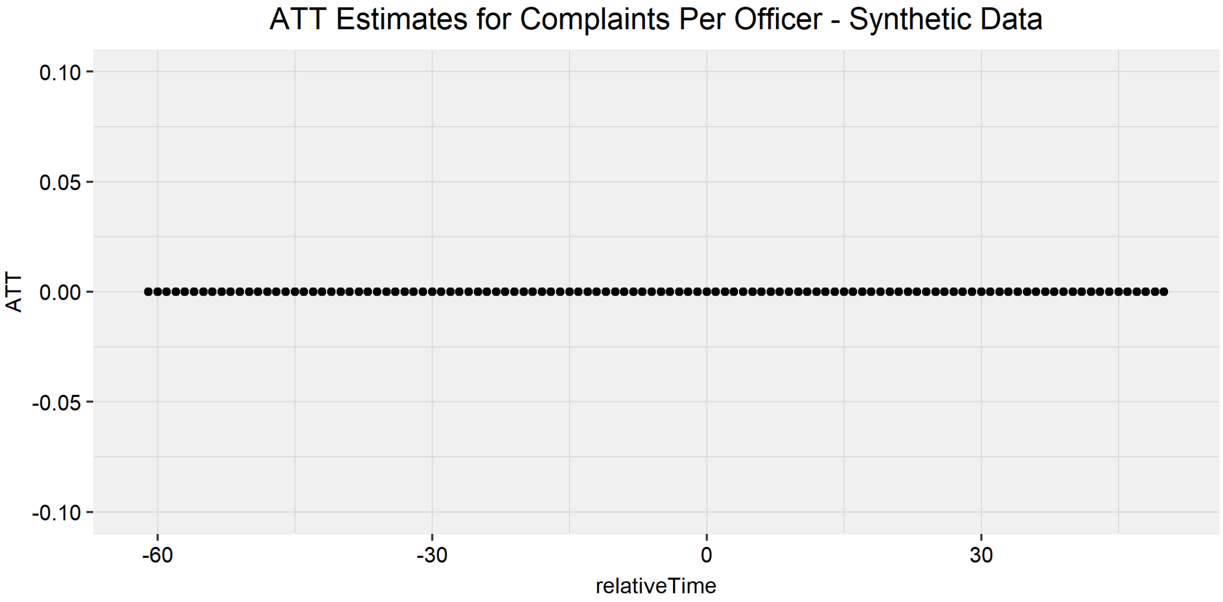
Note: this figure shows ATT estimates from the felc package using a synthetic dataset where complaints-per-officer is the same for all cohorts in each period.

Figure 5



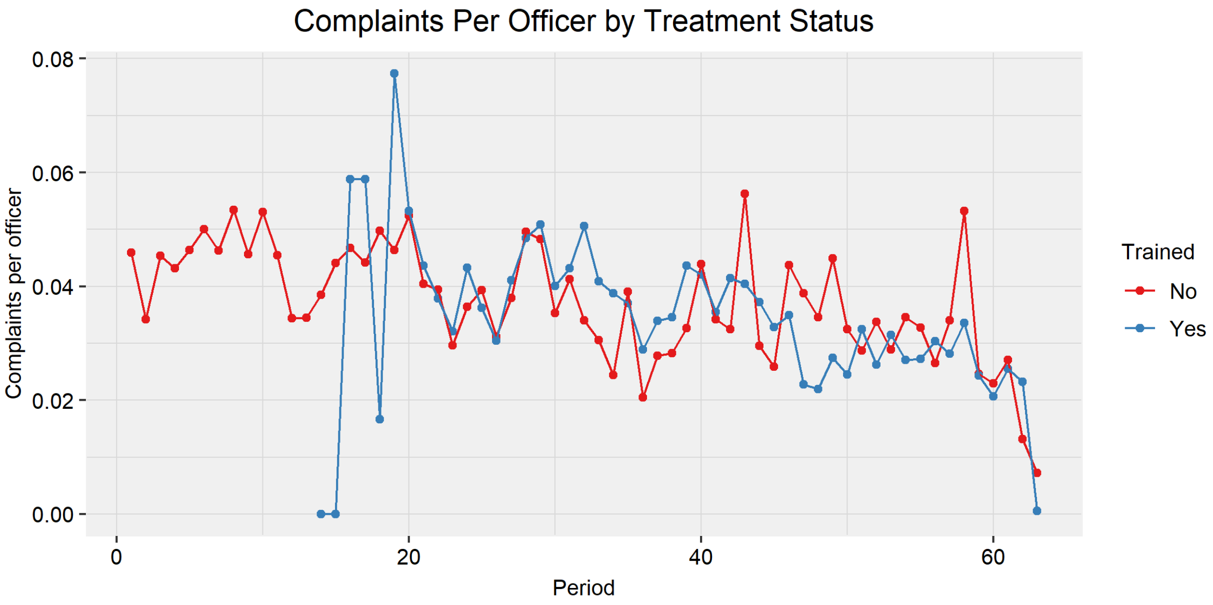
Note: this figure shows ATT estimates from the felc package using complaints-per-officer instead of complaints.

Figure 6



Note: this figure shows ATT estimates from the felc package using complaints-per-officer on the synthetic data described above.

Figure 7


Complaints Per Officer by Treatment Status
Complaints per officer
0.08
0.06
0.04
0.02
0.00
0
20
40
60
Period
Trained
No
Yes