Revised findings for "Procedural justice training reduces police use of force and complaints against officers"

George Wood* New York University
Tom R. Tyler Yale University
Andrew V. Papachristos Northwestern University
Jonathan Roth Microsoft
Pedro H.C. Sant'Anna Vanderbilt University

October 8, 2020

Abstract

Wood et al. (2020) studied the rollout of a procedural justice training program in the Chicago Police Department and found large and statistically significant impacts on complaints and sustained complaints against police officers and police use of force. This document describes a subtle statistical problem that led the magnitude of those estimates to be inflated. We then re-analyze the data using a methodology that corrects for this problem. The re-analysis provides less strong conclusions about the effectiveness of the training than the original study: although the point estimates for most outcomes and specifications are negative and of a meaningful magnitude, the confidence intervals typically include zero or very small effects. On the whole, we interpret the data as providing suggestive evidence that procedural justice training reduced the use of force, but no statistically significant evidence for a reduction in complaints or sustained complaints.

Wood et al. (2020) studied a large-scale procedural justice training program in the Chicago Police Department. The study focused on 8,480 officers trained from January 2012 to March 2016. The study evaluated whether the training program affected three outcomes: complaints filed against officers, complaints that were sustained or resulted in a settlement payout, and use of force as measured by mandatory tactical response

*Corresponding author: george.wood@nyu.edu

reports. The paper reported that training reduced complaints by 11.6 per 100 officers, sustained and settled complaints by 1.7 per 100 officers, and use of force by 7.5 per 100 officers in the 24 months following training.

The study has a staggered adoption design in which officers are assigned to training at different times and remain trained thereafter. The 8,480 officers were clustered ($N = 327$) by the date on which they received training. The three outcomes were measured for each cluster in each month from January 2011 to March 2016 ($T = 63$). Wood et al. used the resulting $N * T$ panel data and an interactive fixed effects (IFE) estimator (Xu, 2017) to analyze the training effects using the specification:

$$Y_{it} = a_i + \lambda_i' f_t + e_{it} \tag{1}$$

where Y_{it} is the total outcome (e.g. the count of complaints received) in cluster i and period t, a_i is an intercept for cluster i, f_t is a vector of factors representing the adoption of training across time periods, λ_i is a vector of factor loadings which represent unobserved characteristics of the officer clusters, and e_{it} is an error term.

1. Specification error

On July 14 2020, Roth and Sant'Anna alerted us to an issue with the above specification. In their comment, Roth and Sant'Anna highlighted two aspects of the data: (1) the size of the clusters differ, with later-trained clusters containing fewer officers on average than early-trained clusters; and (2) the outcomes are in secular decline (that is, the outcomes declined among trained and not-yet-trained clusters) during the study period. Figure 1 shows the size of each cluster by training date. Figure 2 shows the secular decline in each outcome among not-yet-trained officers during the study period. Roth and Sant'Anna noted that, in the presence of a secular decline, smaller later-trained clusters would be expected to experience a smaller total reduction in the outcomes, regardless of any training effects. Being informed by the smaller total reductions in later-trained clusters, the magnitude of the secular decline among early-trained clusters will be underestimated, resulting in the counterfactual outcomes for early-trained clusters being unduly large.

To the extent that the counterfactual is underestimated, this would impact the training estimates because the average treatment effect on the treated (ATT) is calculated by:

$$ATT = \frac{1}{N_{trained}} \sum_{i \in \tau} [Y_{it}(1) - Y_{it}(0)], \quad \text{for } t > T_{0,i} \tag{2}$$

where $N_{trained}$ is the number of trained clusters, τ is the set of trained clusters, $T_{0,i}$ is the training period for cluster i, $Y_{it}(1)$ is the observed outcome for cluster i in period t, and $Y_{it}(0)$ is the counterfactual estimate for cluster i in period t. As can be seen in Equation 2, inflated estimates of the counterfactual values $Y_{it}(0)$ results in the magnitude of the ATT being overestimated in the original analysis.

We investigated this issue immediately. In this note, we report revised estimates of the training effects under an alternative specification that does not suffer from the above issue.

2. Revised estimates

The general empirical strategy in Wood et al, based on multiple period difference-in-differences, remains appropriate for the staggered adoption of procedural justice training as implemented in the Chicago Police Department and the panel structure of the data. From a policy perspective, it also remains important to allow for dynamic treatment effects, whereby effects can vary with respect to the time since training occurred. The model in the original analysis specifies the outcome Y_{it} as the total cluster-level outcome in each period. In the presence of the secular decline in the outcomes and differing cluster sizes, it is this specification that underlies the issue with the original model.

Revising the estimates based on Roth and Sant'Anna's critique involves using a specification that accounts for cluster size. We re-estimated the effects using officer-level outcomes rather than cluster-level outcomes, resulting in a panel of outcomes Y_{it} for officer i in month t. Using the officer-level outcomes means that the counterfactual for already-trained officers is informed by the trend in the outcomes for not-yet-trained officers, rather than not-yet-trained clusters.

For the revised estimates reported here, we update the outcome data to cover the entire staggered adoption period of the training data up to December 2016. These additional 10 months of data from March 2016 to December 2016 were unavailable at the time of the original publication. We incorporated the updated data, which cover a further 137 officers who were trained from March 2016 onward, to provide the strongest possible assessment of the training program. In the Appendix, we present estimates using the data up to March 2016 as in Wood et al (2020). As for the original analysis, all data and code for reproducing the estimates is available on GitHub.[1]

We use the generalized difference-in-differences procedure in Callaway and Sant'Anna (2020) to estimate the impact of the training program. We calculate four types of effect aggregations to evaluate the heterogeneity of the training effects: (i) the simple weighted average effect; (ii) dynamic effects; (iii) calendar effects; and (iv) cohort-specific time averaged effects. The simple weighted average is the average of the estimated ATTs for each cohort-month pair, weighted by the cohort size and number of post-treatment periods, where a cohort is a group of officers who were trained in the same month. The dynamic aggregation evaluates the effects relative to the time since training occurred. The calendar aggregation provides the average monthly effect of undertaking training for cohorts treated by a particular period of the training roll-out. The calendar aggregation provides the average monthly effect of undertaking training for cohorts treated by a particular period of the training roll-out. We present the average of the dynamic, cohort-specific, and calendar effect aggregations to summarize the overall effect of training.

We fit the models using the did package in R.[2] The panel contains 8,614 officers across 72 months from January, 2011 to December, 2016. All 8,614 officers were assigned to training by December, 2016. We present estimates using a balanced panel in which 829 officers who resigned after undertaking training and before the end of 2016 were excluded, leaving a total of 7,785 officers.

[1] https://github.com/george-wood/procedural_justice_revisited

[2] https://github.com/bcallaway11/did

2.1. Average effects

Figure 3 shows the average effects of training on complaints, sustained and settled complaints, and use of force using the updated data from January, 2011 to December, 2016. For complaints, the simple weighted average is -0.005 complaints per officer-month (95% CI: -0.013, 0.003), the average of the dynamic effects is -0.011 (95% CI: -0.028, 0.005), the average of the calendar effects is -0.014 (95% CI: -0.029, 0.002), and the average of the cohort-specific effects is -0.004 (95% CI: -0.012, 0.003).

For sustained and settled complaints, the simple weighted average, the average of the dynamic effects, and the average of the cohort-specific effects are near zero. The average of the calendar effects is -0.004 (95% CI: -0.007, -0.001).

For use of force, the simple weighted average is -0.011 uses of force per officer-month (95% CI: -0.020, -0.001), the average of the dynamic effects is -0.015 (95% CI: -0.037, 0.007), the average of the calendar effects is -0.020 (95% CI: -0.036, -0.005), and the average of the cohort-specific effects is -0.011 (95% CI: -0.020, -0.001). The confidence interval for the dynamic effect aggregation includes zero. However, the intervals for the simple weighted average, calendar, and cohort-specific aggregations do not include zero.

In terms of the magnitude of the average effects, officers received 0.044 complaints per month, on average, in the 12 months before undertaking training (or 4.4 per 100 officers). A reduction of -0.004 complaints per month, the smallest average estimate under the cohort-specific aggregation, would imply a 9% reduction in complaints per month. The largest average estimate of -0.014 under the dynamic aggregation would imply a 32% reduction. However, the confidence interval for the dynamic effect aggregation is wide and the intervals include zero for the four average effect estimates. With the average effects on complaints being marginally non-significant for the four aggregations, it would be unwarranted to rule out the possibility that training had no effect on complaints.

Officers reported using force 0.047 times per month in the 12 months before training (or 4.7 per 100 officers). The smallest estimated reduction in use of force of -0.010 per officer for the cohort-specific aggregation corresponds to a 21% reduction, whereas the largest estimated average reduction of -0.020 uses of force for the calendar aggregation

corresponds to a 43% reduction. Again, the confidence intervals are wide and cover -0.020 to -0.001 for the simple weighted average effect, containing effect magnitudes ranging from very small at the low end to very large at the high end. At the low-end, a reduction of -0.001 uses of force per officer per month would result in 8 fewer uses of force per month (a 2% reduction) across the 7,785 trained officers, while at the-high end an average reduction of -0.020 would result in 156 fewer uses of force per month. The confidence interval for the average of the dynamic effects is particularly wide and includes very large negative values and small positive values.

In the Appendix, we show the average effect estimates for the four aggregations using the data up to March 2016, which is the time frame used in the original study.

2.2. Dynamic effects

Figure 4 shows the dynamic treatment effects on complaints, sustained and settled complaints, and use of force, and can be interpreted similarly to an event-study plot where the effects are relative to the time since training. For complaints, 35 of the point estimates in the 36 months after training are negative and relatively noisy. For context, in the year before training, the mean complaints per officer was 0.044 per month. The point estimates generally vary in the range -0.005 to -0.01, which is equivalent to a reduction of between 0.5 and 1 complaints per 100 officers per month. However, the pointwise confidence intervals and simultaneous confidence intervals are wide and include zero which, along with the average ATT estimates in Figure 3, does not represent compelling evidence for a training effect on complaints. The simultaneous confidence intervals, in particular, indicate that no firm conclusion about the effect of training on complaints can be drawn.

The middle panel of Figure 4 shows the dynamic treatment effects on sustained and settled complaints. The majority of the point estimates are near zero and indicate no effect of training on sustained and settled complaints.

The right panel of Figure 4 shows the dynamic treatment effect on use of force. The point estimates in the 36 months after training are all negative and most are reasonably

large in magnitude, with all but 5 estimates varying between approximately -0.004 to -0.017 uses of force per officer-month. The estimates are relatively large in the context of a mean of 0.047 uses of force per officer-month in the year before training, suggesting a reduction between 31 and 132 uses of force across the 7,785 trained officers per month. The pointwise confidence intervals, which are akin to those used in the original analysis, the simple weighted average ATT, calendar ATT, and cohort ATT are consistent with an effect of training on use of force. However, the simultaneous confidence intervals for the dynamic effects, which account for the entire pathway of the dynamic effects, are very wide and exclude zero in only one post-training period. Consequently, while the dynamic effect estimates suggest a possibly large effect on use of force, we cannot rule out that the dynamic estimates could be zero in all but one period. This is reflected in the confidence intervals for the average of the dynamic effect ATTs in Figure 3, which includes zero.

2.3. Calendar effects

The procedure in Callaway and Sant'Anna (2020) allows us to learn more about the impact of training by calculating calendar effects. The calendar effects summarize the average effect among already-trained officers in each month of the study. Figure 5 shows the calendar effects on each outcome. The estimates for complaints are negative in 16 of the 24 months from month 17, when the training roll-out began in full, to month 40. After month 40, the effects are rarely substantively below zero. In two periods (21 and 32), the effect on complaints is positive. These effects provide suggestive evidence that the training program may have impacted complaints among earlier-adopters at the beginning of the observation period, with any reduction diminishing as the roll-out continued. Again, the simultaneous confidence intervals are wide and do not include zero in only two months, being marginally non-significant in four other months.

The middle panel of Figure 5 shows calendar effects on sustained and settled complaints. The estimates are negative until month 31, after which there is a positive estimate in month 32 followed by near-zero estimates. The calendar estimates early in the roll-out period are large in magnitude relative to the mean of 0.004 sustained or settled

complaints per month, with four ATTs between months 19 and 29 having simultaneous confidence intervals that do not include zero. However, any impact on sustained and settled complaints appears to have diminished after approximately the first year of the training roll-out and again we caution that the confidence intervals are wide.

The right panel of Figure 5 shows the calendar effects on use of force. The estimates are negative in 53 of the 60 periods. As in the case of sustained complaints, the effect of training on use of force seems to have been greater in the early-period of the training roll-out until approximately month 40, 24 months after the roll-out began in full. The pattern of these calendar effects suggests that training was more impactful for officers trained in 2012 and 2013, the first two years of the staggered adoption period.

2.4. Cohort-specific effects

Figure 6 shows cohort-specific effects on each outcome. These represent the effect for each cohort of trained officers, i.e. officers who were trained in the same month, averaged over all time periods. The cohort-specific effects are highly variable for all three outcomes. Additionally, the confidence intervals are extremely wide in many periods relative to the mean outcome in the 12 months before training. The average of these cohort-specific ATTs is reported in Figure 3.

3. Summary

Wood et al. (2020) examines an important question: can procedural justice training reduce misconduct complaints against officers and officer use of force? Police misconduct and use of force are critical issues, affecting public health, public safety, and the legitimacy of the criminal justice system. Any evaluation of procedural justice training has direct policy implications, contributing to an evidence base that should we weighed when determining the content and structure of training programs in police departments, how to reduce harmful policing practices, and the broader options for police reform.

We re-evaluated our analyses based on a critique offered by Roth and Sant'Anna using a different methodology that addresses the issue associated with varying cluster

sizes and the secular decline. This methodology allowed us re-investigate our original analyses while considering calendar and cohort-specific effects, in addition to the dynamic effects in the original analysis.

On the balance of the evidence, the new analysis provides suggestive evidence that training reduced officer use of force. The estimated magnitude of the simple weighted average effect and average of the dynamic effects, calendar effects, and cohort-specific effects on use of force is substantively meaningful. The 7,785 trained officers used force approximately 363 times per month in the year before training, on average. The simple weighted average estimate implies a reduction of approximately 70 uses of force per month across the trained officers, although the confidence intervals are wide and could imply smaller or larger effects. The confidence intervals for three of the four aggregations—the simple weighted average, calendar, and cohort-specific effects—do not include zero, and the dynamic point estimates are negative in nearly all post-training periods. Overall, it is clear both that there are a variety of possible effect sizes depending on the aggregation used and that for all of the estimates the level of uncertainty is high. We also note that a Bonferroni-type correction for multiple hypothesis testing would lead to further uncertainty about the magnitude of the estimates.

However, when using this new modeling approach, we do not find a statistically significant reduction in complaints or sustained and settled complaints, although again the confidence intervals are wide and include meaningful effects.

This analysis using a different methodology brings our findings into alignment with the findings of Owens et al. (2018), who estimated that officers receiving a procedural justice intervention were 16% to 50% less likely to be involved in a use of force incident in the six-week period after training but were no more or less likely to receive complaints. The combination of the evidence in this paper and Owens et al (2018) suggests that procedural justice training is likely to have reduced officer use of force, but we cannot conclude that training reduced complaints or sustained and settled complaints.

References

B Callaway and PHC Sant'Anna. Difference-in-differences with multiple time periods. *Available at SSRN: https://ssrn.com/abstract=3148250*, 2020.

Emily Owens, David Weisburd, Karen L. Amendola, and Geoffrey P. Alpert. Can you build a better cop? *Criminology & Public Policy*, 17(1):41–87, 2018.

Jonathan Roth and PHC Sant'Anna. Comments on "procedural justice training reduces police use of force and complaints against officers".

George Wood, 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*, 117(18):9815–9821, 2020.

Yiqing Xu. Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, 25:57–76, 2017.

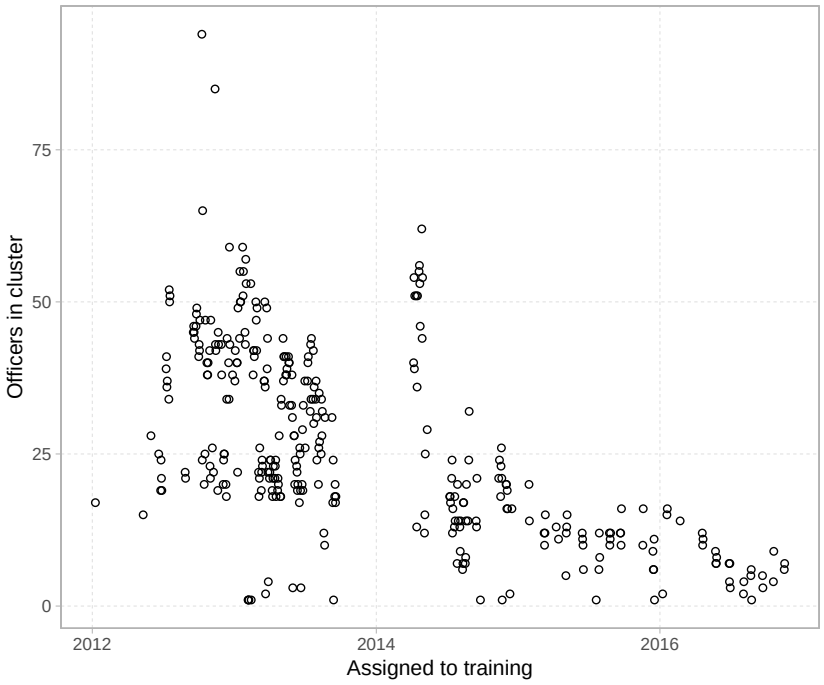

Figure 1: Officers in each cluster by date of training. Early-trained clusters contain more officers, on average, than later-trained clusters.

Figure 2: Mean complaints, sustained and settled complaints, and use of force per officer by month in the balanced panel. The original and updated data are shown. As in the original analysis, the sustained and settled complaints data ends in October, 2015.

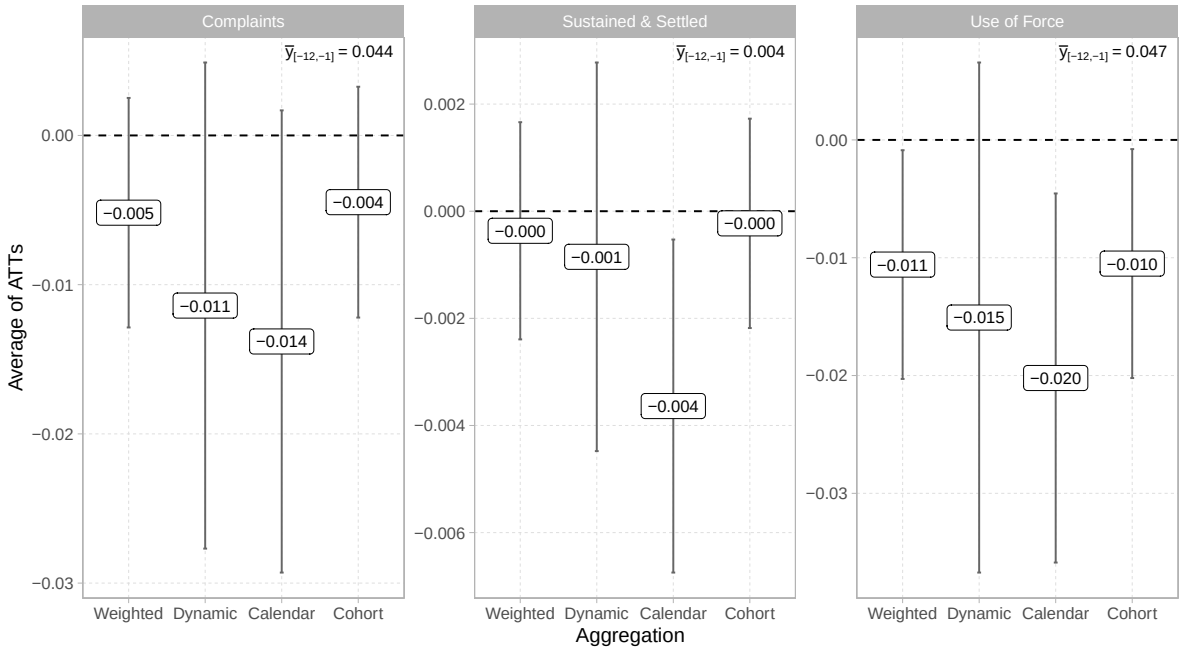

Figure 3: Estimated simple weighted average ATT, average of the dynamic effects, average of the calendar effects, and average of the cohort-specific effects on each outcome per officer by month. The mean outcome per officer in the 12 months before training ($\bar{y}_{[-12,-1]}$) is reported. 95% confidence intervals are shown. Data from January 2011 to December 2016.

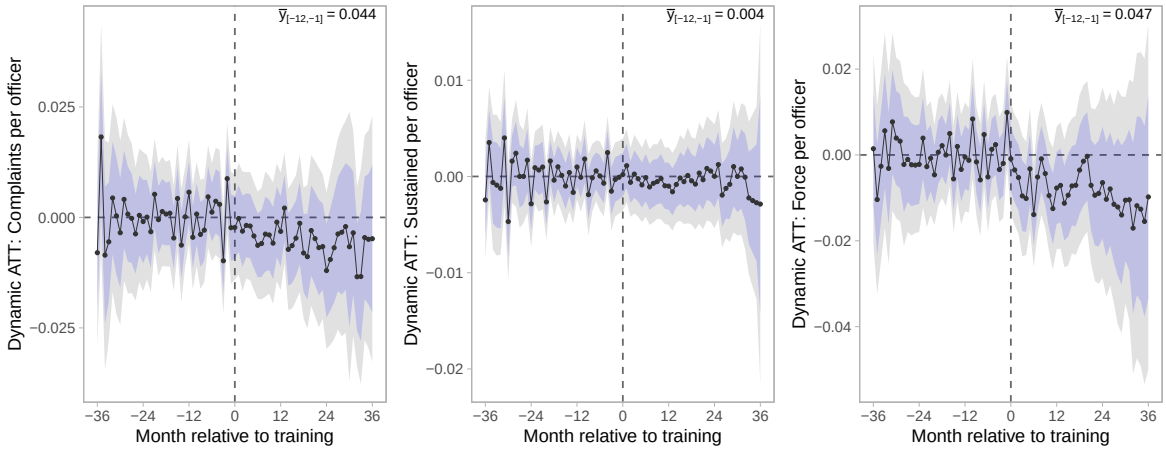

Figure 4: Estimated dynamic ATT per officer by month relative to training. 95% pointwise confidence intervals are shown in blue and 95% simultaneous confidence intervals, which account for multiple-hypothesis testing (Callaway and Sant'Anna, 2020), are shown in gray. The mean outcome per officer in the 12 months before training ($\bar{y}_{[-12,-1]}$) is reported.

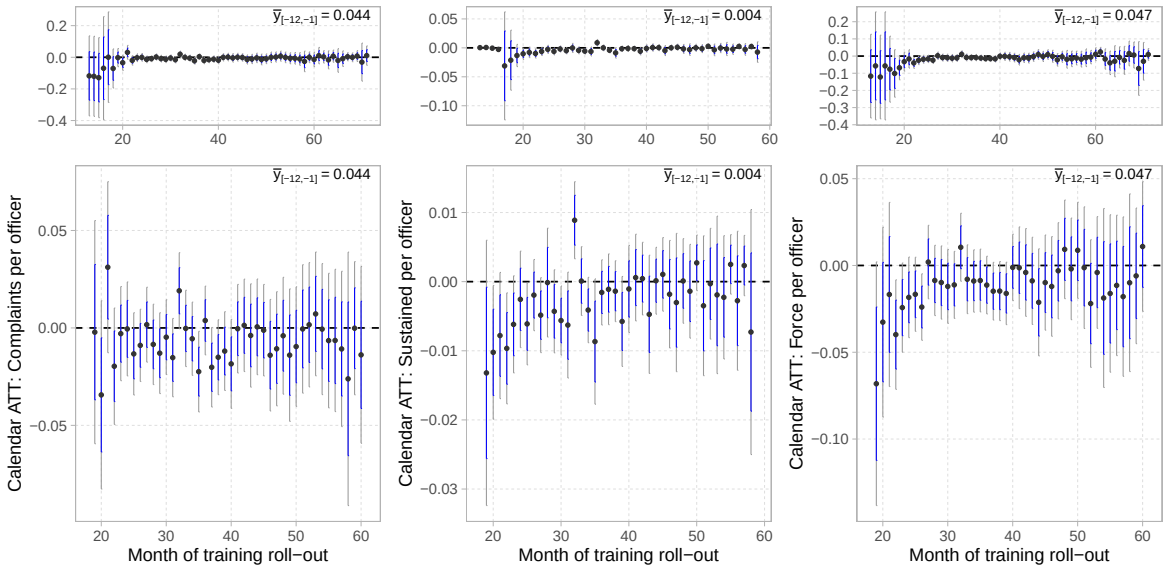

Figure 5: Estimated calendar effects per officer by month of the study. The top row shows the full range of the study period. The bottom row shows the period between month 19 and 60 to provide greater readability of the estimates and confidence intervals. 95% pointwise confidence intervals are shown in blue and 95% simultaneous confidence intervals are shown in gray. The mean outcome per officer in the 12 months before training ($\bar{y}_{[-12,-1]}$) is reported.

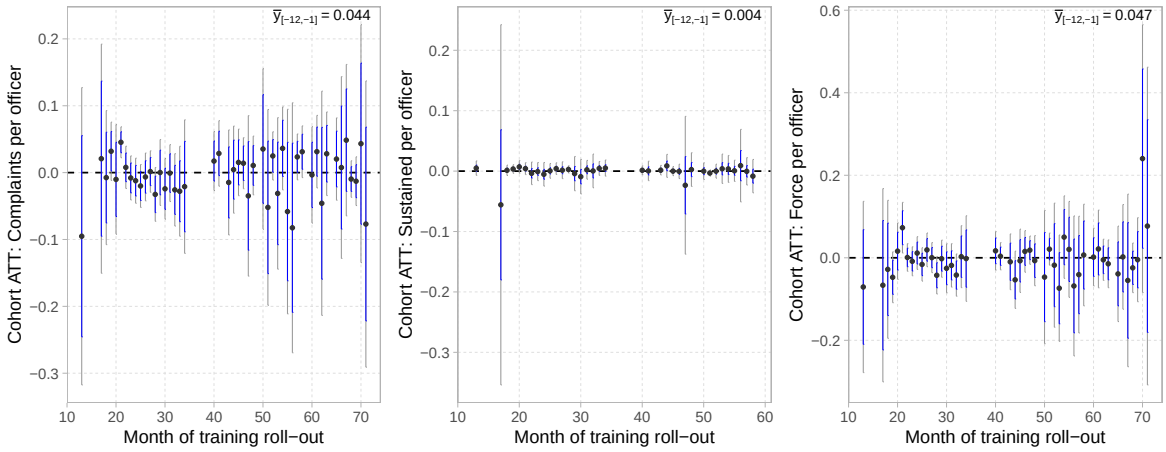

Figure 6: Estimated cohort-specific effects per officer-month. 95% pointwise confidence intervals are shown in blue and 95% simultaneous confidence intervals are shown in gray. The mean outcome per officer in the 12 months before training ($\bar{y}_{[-12,-1]}$) is reported.

Appendix

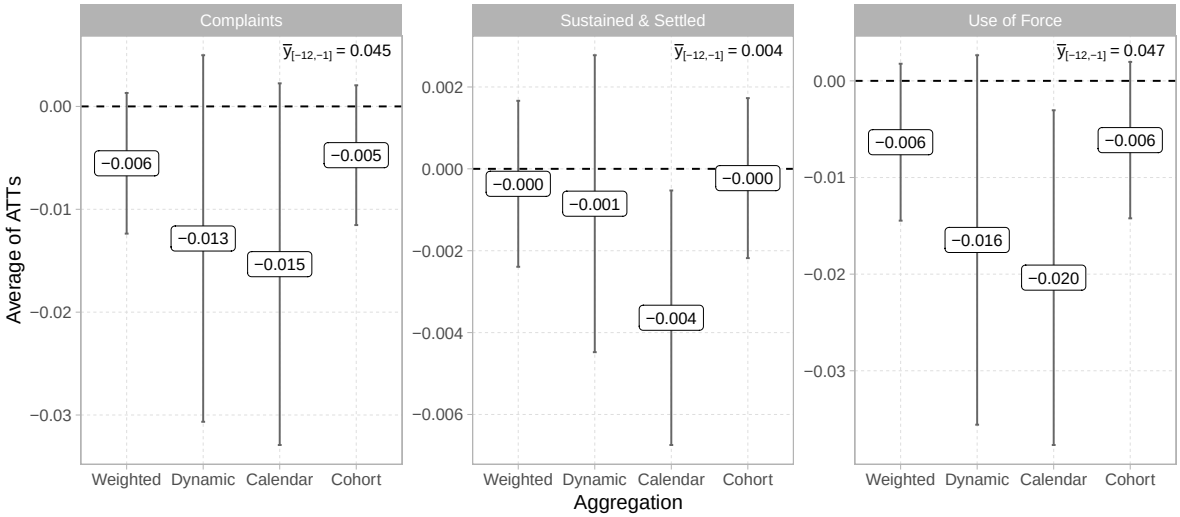

Figure 7: Estimated simple weighted average ATT, average of the dynamic effects, average of the calendar effects, and average of the cohort-specific effects on each outcome per officer by month using data from January 2011 to March 2016. The mean outcome per officer in the 12 months before training ($\bar{y}_{[-12,-1]}$) is reported. 95% confidence intervals are shown.

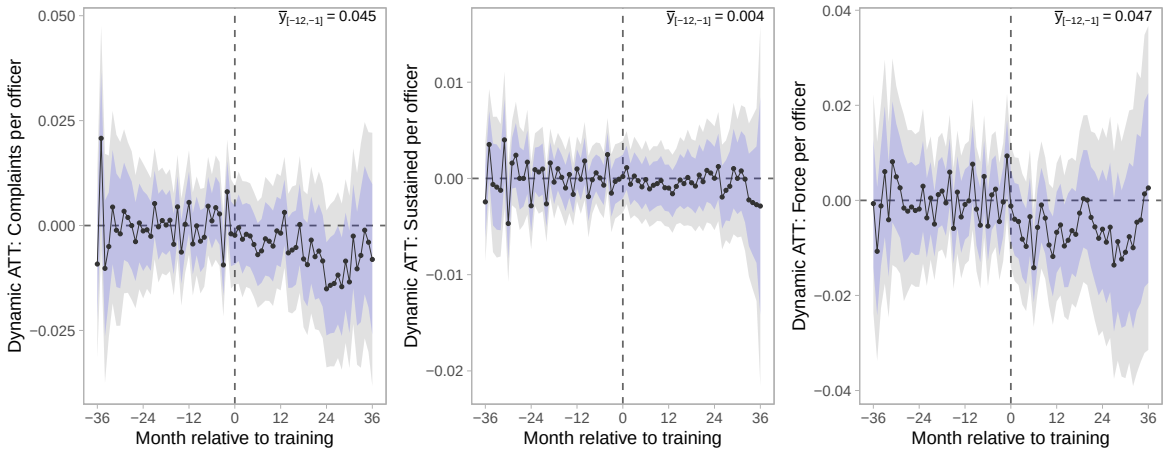

Figure 8: Estimated dynamic ATT per officer by month relative to training using data from January 2011 to March 2016. 95% pointwise confidence intervals are shown in blue and 95% simultaneous confidence intervals, which account for multiple-hypothesis testing (Callaway and Sant'Anna, 2020), are shown in gray. The mean outcome per officer in the 12 months before training ($\bar{y}_{[-12,-1]}$) is reported.