Reply to Sierra-Arévalo and Papachristos (2021)

Aaron Chalfin*[1] and Jacob Kaplan[2]

[1]Department of Criminology, University of Pennsylvania
[2]Princeton School of Public and International Affairs, Princeton University

April 16, 2021

Abstract

The notion that the unjustified use of force by police officers is concentrated amongst a few "bad apples" is a popular descriptor which has gained traction in scholarly research and achieved considerable influence among policymakers. Prior research documents that the top 2 percent of police officers often account for as many as 50 percent of citizen complaints. These "in-sample" statistics are cited, in part, to make the case that a large number of complaints can be abated by terminating a small number of problematic officers. But is removing the bad apples, in fact, likely to have an appreciable effect on police misconduct? In Chalfin and Kaplan (2021), forthcoming in *Criminology & Public Policy*, we motivate a simple but, we argue, informative policy simulation in which we estimate the effects of removing a small number of officers identified at the end of their probationary period based on *ex ante* risk and replacing them with alternative officers. Because citizen complaints are relatively rare and the standard probationary period of 18 months is short, future complaints are difficult to predict. We conclude that removing a small number of "bad apples" at the end of their probationary period is unlikely to have as large an impact on citizen complaints as the naive "in sample" statistics imply. In a response essay, Sierra-Arévalo and Papachristos (2021) critique our analysis and claim that our conclusions are due a selective reading of our results and inadequate incorporation of network spillovers into our analysis. In this note, we provide a brief reply to their essay.

*Please address correspondence to: Aaron Chalfin, Department of Criminology, University of Pennsylvania, E-Mail: `achalfin@sas.upenn.edu`

1 Background

The idea that a small number of "bad apples" are responsible for an outsize share of complaints against police officers has gained considerable traction over the course of the last four decades both in the scholarly literature (Berkow, 1996; Alpert and MacDonald, 2001; Walker et al., 2001; Rozema and Schanzenbach, 2019; Goncalves and Mello, 2020) and in popular media accounts (Arthur, 2018; Invisible Institute, 2018; Wu, 2019; Ba and Rivera, 2020; Kelly and Nichols, 2020; MacDonald and Klick, 2020). Empirically the claim that a small number of police officers account for an outsize share of serious misconduct rests on analyses of individually identified microdata on complaints against police officers. Prior analyses from police departments across the United States suggest that a small share of officers account for a large share of complaints against police. Indeed a common estimate is that the top 2 percent of officers account for approximately 50 percent of known misconduct by police officers (Walker et al., 2001). As the other 98 percent of officers would then be responsible for the remaining 50 percent of misconduct, the implication is that the top 2 percent of officers are, incredibly, 49 times more likely to commit misconduct than other officers.

While these "in-sample" computations point to an incredible degree of concentration of complaints among a small number of problematic officers, the key policy question is how much many complaints can be abated by incapacitating predictably problematic police officers, a question which hinges on the ability of analysts to make an *ex ante* prediction about which officers will be the subject of future complaints. In Chalfin and Kaplan (2021), we consider the likely effects of incapacitating ex ante predictable "bad apples." We motivate a simple but, we argue, informative policy simulation in which we estimate the effects of removing a small number of officers identified based on *ex ante* risk and replacing them with alternative officers. Because citizen complaints are relatively rare and the standard probationary period of 18 months is short, future complaints are difficult to predict. We conclude that removing a small number of "bad apples" at the end of their probationary period is likely to have a more modest impact on citizen complaints than in-sample estimates of the concentration of complaints (e.g., the top $X\%$ of officers account for $Y\%$ of complaints) imply.

In a response essay, Sierra-Arévalo and Papachristos (2021) (SP) critique our analysis and claim that the conclusions we draw follow from a selective reading of our results and inadequate incorporation of network spillovers into our analysis. In this note, we provide a brief reply to these critiques. Prior to providing a brief set of replies, we want to emphasize that we deeply respect SP's alternative perspective on our analyses and believe that debate about what these estimates mean is both healthy and incredibly important. While we stand by our work, we hope that our exchange in *Criminology & Public Policy* will spark further conversation among people working in this area and also further research.

2 Reply to SP's Critiques

We begin our response to SP's critiques by noting that either Sierra-Arévalo or Papachristos are among the referees who provided feedback on our original submission to *Criminology & Public Policy*. As such, our final manuscript already incorporates some of their initial feedback. While we respectfully disagree with some of how they characterize our findings in their response essay, we are deeply grateful to them for noting some important omissions in our initial draft that we have since had an opportunity to address. We offer the following responses to each of SP's remaining critiques:

1) SP point out that, regardless of the effects of doing so, terminating "bad apples" provides an important normative benefit.

We agree. In fact, it's not clear that anyone who cares about the provision of high-quality police service would claim otherwise. We hope it is understood that our paper does not argue that it is appropriate to retain problematic police officers on the job. Instead, we simply point out that, at current rates of termination, firing "bad apples" is unlikely to abate a large share of citizen complaints. As we state in our paper's introduction, "our conclusion is that while incapacitating predictably problematic officers serves an important instrumental purpose, this practice is, in of itself, unlikely to lead to a large reduction in use of force complaints." This statement captures our purpose in writing this paper. While we, of course, acknowledge the importance of firing troubled officers, we believe that commonly cited in-sample statistics of the form "X% of officers account for Y% of the citizen complaints" offers policymakers and other interested observers a misleadingly optimistic view that terminating a tiny share of "bad apples" could be a panacea for reform efforts.

2) SP claim that we selectively focus on estimates in which we identify ex ante risk based on an 18-month probationary period, to the exclusion of estimates that consider a hypothetical five-year probationary period.

Given that citizen complaints against police officers are relatively rare — officers in Chicago accrue complaints at a rate of approximately 0.2 complaints per year — "bad apples" can be identified with greater accuracy when more data are brought to bear. Accordingly, we can gain more traction in identifying problematic police officers using five years of data than using 18 months of data. There is then a tradeoff in making predictions about officer risk. The longer we wait to identify “bad apples”, the greater is our ability to abate future complaints. On the other hand, by waiting longer to make a prediction, we forgo the opportunity to have abated past complaints. This tradeoff is clearly noted in the paper along with the observation that it is considerably easier to terminate a police officer at the end of his or her 18-month probationary period than later in an officer’s career.

Contrary to SP’s claim, we do not selectively present results by focusing on the 18-month analysis to the exclusion of the five-year analysis. Indeed, estimates in which we identify “bad apples” using an 18-month probationary period and estimates in which we identify “bad apples” using a hypothetical five-year probationary period are presented in the same table — Table 2. Both sets of results are reported in the main body of the paper and neither is hidden in an appendix.

In the paper’s abstract, we focus on findings from the 18-month probationary period in recognition of the tremendous difficulty in terminating officers later on. Indeed, in Chicago officers are terminated at the rate of 0.2% of officers per year. Until we see police departments terminating officers at a rate of 10% (or even 2%) per year, in our view, these estimates have the greatest relevance for public policy. An alternative view is that while it is difficult to terminate officers after their probation period has ended, early warning systems continue to have value in selecting officers for re-training or remediation. This is an important perspective. However, in this paper, the intervention we consider is termination.

3) SP claim that we selectively focus on citizen complaints rather than tactical response reports which are used by Chicago police officers to document encounters in which force was used.

In Chicago, police officers are typically required to fill out a tactical response report (TRR) when force is used during an encounter with a citizen. TRRs are more common than complaints and, as such, TRRs are easier to forecast than complaints. As such, when we focus on TRRs, we find that incapacitating the top 10% officers would abate 9% of TRRs as opposed to 5-6% of complaints. Contrary to the claim that results are reported selectively, Table 2 which presents our principal findings includes results for both citizen complaints and TRRs.

Our preferred estimates focus on citizen complaints for several reasons. First, research by Rozema and Schanzenbach (2019) has shown that complaints are a surprisingly good predictor of high-impact events such as lawsuit payouts by municipal officials. Second, by focusing on citizen complaints instead of sustained complaints or use of force incidents recorded via police department record keeping, we use data that has not been filtered through the lens of what a law enforcement agency deems problematic and which therefore may better reflect community norms. Finally, while we agree that the use of force by officers is an important outcome to consider, the application of force by police officers is often consistent with both law and policy and is sometimes a necessary part of the job. As such, it is not clear that TRRs could be straightforwardly used to terminate officers, without also considering whether those incidents led to complaints.

An underlying theme in SP's critique is that we have hidden the ball by selectively focusing on some results and ignoring uncertainty. This strikes us as an unusual claim as the figures that they have generated in their response essay which purport to show that we selectively present results are themselves derived from estimates that we include in our paper. In our discussion, we focus most intensively on discussing the results that arise from the most realistic assumptions — namely that it is difficult to terminate a large number of officers, especially after the probationary period has ended.

4) SP claim that we do not account for the importance of network spillovers in our analysis.

As Andrew Papachristos has shown in a series of important papers, social networks are important in policing. With respect to misconduct, prior research notes that an outsize share of complaints tend to cluster in a small number of peer networks (Ouellet et al., 2019; Wood et al., 2019; Zhao and Papachristos, 2020). This work is descriptively compelling and recent research suggests that peer effects may have important causal impacts as well Quispe-Torreblanca and Stewart (2019).

Quispe-Torreblanca and Stewart (2019), in particular, use data from the London Metropolitan police and find that a 10% increase in misconduct among one's peers increases one's own misconduct by 8%. While this suggests that the transmission of misconduct through peer networks is relatively inelastic, this finding nevertheless points to an important alternative channel through which misconduct could be reduced by terminating "bad apples." Taking this estimate at face value, how much would accounting for network spillovers change our estimates? Here, we note that since the relationship is multiplicative, the importance of network spillovers grows with the size of the incapacitation effect achieved. Let's say that for every 1 percent of misconduct that is incapacitated through the removal of "bad apples", we can expect an additional 0.8 percent of misconduct to be removed via network spillovers thus increasing the effect size by 80 percent. 80% of a large effect is large but 80% of a small effect is small. We estimate that removing the top 2% of bad apples abates 1% of complaints and incapacitating the top 10% of bad apples abates 5-6% of complaints. As both we and SP demonstrate, multiplying each of these estimates by 1.8 yields spillover-adjusted estimates of 1.8% and 9-11%, respectively.

SP suggest that the difference between 5-6% and 9-11% is more meaningful than we allow in our paper. We accept that such a difference may well be meaningful in a variety of ways. However, when compared to in-sample estimates that note that the top 2% of officers account for as much as 50% of the misconduct, these estimates are considerably more modest regardless of how they are sliced. Likewise, to the extent that a police department terminates a fewer than 10% of officers for cause at the end of the probationary period, the quantitative importance of spillovers becomes proportionately smaller.

5) SP note that the magnitude of our estimates are similar to many prevailing estimates of the effect of alternative interventions that are intended to reduce police misconduct.

This is an excellent point. As SP point out, many evaluations of training programs that are intended to reduce police misconduct point to modest impacts. While some evidence points to potentially larger effects — for example, procedural justice-inspired training in Seattle (Owens et al., 2018) or body-worn cameras nationally (Kim, 2019), we agree with SP's characterization of our estimated incapacitation effects in comparison to other interventions.

Whether to think of our estimates as "large" or "small" requires making a normative judgement and depends on the counterfactual that is under consideration. In our paper, we characterize our estimates as "modest." In doing so, we are thinking about the size of our estimates in comparison to in-sample statistics which indicate a large degree of concentration of police misconduct. We believe that our estimates are valuable in dispelling the misunderstanding that when 2% of officers account for 50% misconduct, this does not mean that such an outcome can be achieved through terminating 2% of officers.

With respect to what our estimates mean for policy decisions in this area, we remain fairly agnostic. On the one hand, our results suggest that terminating "bad apples" can play a role in reducing complaints against police officers. Likewise, terminating problematic officers is procedurally just and an achieves an important normative goal. On the other hand, our research cautions that efforts to expand the ability to terminate a slightly larger share of officers may involve a great deal of political capital and so policymakers should be realistic about what they believe can be achieved with respect to reducing misconduct. With regard to the politics of firing more officers, since this paper was written, the policy landscape has shifted. If probationary periods are extended or police departments invest in greater data gathering, it remains entirely possible that the incapacitation effects documented in our paper could become larger.

References

Alpert, G. P. and J. M. MacDonald (2001). Police use of force: An analysis of organizational characteristics. *Justice Quarterly 18*(2), 393–409.

Arthur, R. (2018). 130 chicago officers account for 29 percent of police shootings. *The Intercept*.

Ba, B. and R. Rivera (2020). Police think they can get away with anything. that's because they usually do.

Berkow, M. (1996). Weeding out problem officers. *Police Chief 63*, 21–29.

Chalfin, A. and J. Kaplan (2021). How many complaints against police officers can be abated by incapacitating a few 'bad apples?'. *Criminology Public Policy*.

Goncalves, F. and S. Mello (2020). A few bad apples?: Racial bias in policing. *The American Economic Review*.

Invisible Institute, T. (2018). The citizens police data project.

Kelly, J. and M. Nichols (2020). We found 85,000 cops who've been investigated for misconduct. now you can read their records. *USA Today*.

Kim, T. (2019). The impact of body worn cameras on police use of force and performance. *Available at SSRN 3474634*.

MacDonald, J. and J. Klick (2020). Hire more cops to reduce crime. *City Journal*.

Ouellet, M., S. Hashimi, J. Gravel, and A. V. Papachristos (2019). Network exposure and excessive use of force: Investigating the social transmission of police misconduct. *Criminology & Public Policy 18*(3), 675–704.

Owens, E., D. Weisburd, K. L. Amendola, and G. P. Alpert (2018). Can you build a better cop? experimental evidence on supervision, training, and policing in the community. *Criminology & Public Policy 17*(1), 41–87.

Quispe-Torreblanca, E. G. and N. Stewart (2019). Causal peer effects in police misconduct. *Nature Human Behaviour 3*(8), 797–807.

Rozema, K. and M. Schanzenbach (2019). Good cop, bad cop: Using civilian allegations to predict police misconduct. *American Economic Journal: Economic Policy 11*(2), 225–68.

Sierra-Arévalo, M. and A. Papachristos (2021). Bad apples and incredible certitude. *Criminology Public Policy*.

Walker, S., G. P. Alpert, and D. J. Kenney (2001). *Early warning systems: Responding to the problem police officer*. US Department of Justice, Office of Justice Programs, National Institute of Justice.

Wood, G., D. Roithmayr, and A. V. Papachristos (2019). The network structure of police misconduct. *Socius 5*, 2378023119879798.

Wu, K. J. (2019). Study finds misconduct spreads among police officers like contagion. *PBS*.

Zhao, L. and A. V. Papachristos (2020). Network position and police who shoot. *The ANNALS of the American Academy of Political and Social Science 687*(1), 89–112.