"CONTAGIOUS ACCOUNTABILITY"

A Global Multisite Randomized Controlled Trial on the Effect of Police Body-Worn Cameras on Citizens' Complaints Against the Police

BARAK ARIEL
University of Cambridge
The Hebrew University of Jerusalem

ALEX SUTHERLAND
RAND Europe
University of Cambridge

DARREN HENSTOCK
West Midlands Police

JOSH YOUNG
Ventura Police Department

PAUL DROVER
Wolverhampton Police

JAYNE SYKES
West Yorkshire Police

SIMON MEGICKS
Cambridgeshire Constabulary

RYAN HENDERSON
Police Service of Northern Ireland

The use of body-worn cameras (BWCs) by the police is rising. One proposed effect of BWCs is reducing complaints against police, which assumes that BWCs reduce officer non-compliance with procedures, improve suspects' demeanor, or both, leading to fewer complaints. We report results from a global, multisite randomized controlled trial on whether BWC use reduces citizens' complaints. Seven discrete tests (N = 1,847 officers), with police shifts as the unit of analysis (N = 4,264), were randomly assigned into treatment and control conditions. Using a prospective meta-analytic approach, we found a 93% before–after reduction in complaint incidence (Z = –3.234; p < .001), but no significant differences between trial arms in the studies (d = .053, SE = .11; 95% confidence interval [CI] = [–.163, .269]), and little between-site variation (Q = 4.905; p = .428). We discuss these results in terms of an "observer effect" that influences both officers' and citizens' behavior and assess what we interpret as treatment diffusion between experimental and control conditions within the framework of "contagious accountability."

Keywords: body-worn cameras; policing; complaints; multisite randomized controlled trial; accountability

Police body-worn cameras (BWCs) are becoming a common feature in the contemporary police officer's arsenal. Many, if not most, Western police departments have either already deployed these devices or are considering equipping officers with BWCs (Reaves, 2015). This enthusiasm for a technological "fix" to the perceived crisis in police legitimacy is unsurprising, as it is far cheaper to implement technology than re-train officers or solve more endemic social problems. As a result, BWCs are viewed as a panacea for enhancing police compliance with procedures, as a step toward further professionalization of the police (Reveal Media, 2015; TASER, 2015), and as "the ultimate witness" (Bodyworn.com, 2015). [AQ2] Federal judges (Santora, 2013), chief constables, and presidents of nations mention these devices as "the" technology to restore confidence in policing as a social institution (White House Office of the Press Secretary, 2014).

One particular area of interest sparked by BWCs is the potential effect on citizen complaints against the police. Formal grievances are often viewed as a mirror of police (mis)conduct, with more complaints lodged against officers interpreted as reflective of a police agency that is less compliant with police procedure, or with expected conduct, and the

proper code of practice. Thus, reducing complaints is a welcomed outcome, particularly when it is accompanied by improving public confidence and legitimacy (White, 2014) or accountability and transparency (Ready & Young, 2015; Scheindlin & Manning, 2015). It also makes sense economically, as claims resulting from grievances, particularly those that are accepted by the court, could cost police departments millions of dollars that they do not have (Mateescu, Rosenblat, & boyd, 2015).

The interest in the hypotheses supporting the use of BWCs is not matched by rigorous tests of their impact. Literature reviews on the effectiveness of BWCs have suggested that, "despite vast information sources discussing BWC technology, the operational evidence to support claims about either the pros or cons of this technology is sparse" (Cubitt, Lesic, Myers, & Corry, 2016; Lum, Koper, Merola, Scherer, & Reioux, 2015; Stratton, Clissold, & Tuson, 2014, p. 13; White, 2014). With the exception of one field experiment (Ariel, Farrar, & Sutherland, 2015), there is little evidence on whether the benefits of these devices justify their costs. Similarly, while the theory underpinning the effect of BWCs is relatively straightforward—deterrence theory juxtaposed with observer effects—estimates of the efficacy of BWC are scarce, resembling a void largely filled by conceptual research (e.g., Haggerty & Ericson, 2000; Richards, 2013).

There are two challenges for studying the effect of BWCs on policing, particularly in terms of complaints. First, complaints are relatively rare events (Ariel, 2016a; Ariel et al., 2015), which create statistical power issues (Cohen, 1988; Lipsey, 1990). Any one study on an average-sized police department would have to observe the treatment effect for a relatively long period of time to capture sufficiently large samples. The Ariel et al. (2015) study was a 12-month experiment, and this length of study cycle is not viable for most prospective field studies in criminology (Strang, 2012). The second challenge is more substantive and is about deciphering who BWCs are affecting during an encounter: the officer, the suspect, or both? From a practitioner perspective, this conundrum may seem irrelevant, but disentangling the effect is crucial to informing the wider scale rollout of BWCs, as well as for understanding how the effect of the cameras operates in practice. Questions that are, as yet, unaddressed include the following: Who is affected by the treatment? Do cameras affect the suspect first and the officer next, or the other way around? Is it a simultaneous effect?[1] These two challenges are addressed in the present study.

This article proceeds as follows: After reviewing the existing research on complaints against the police, we look at the evidence on BWCs while examining the theoretical mechanisms that underpin the effect of these devices on police–public encounters. We conducted a global multisite randomized controlled trial, with the emphasis on complaints against police officers at seven sites, encompassing nearly 2,000 officers. The results are then presented, with the practical and theoretical implications contemplated in the discussion.

COMPLAINTS AGAINST THE POLICE

For a citizen to be sufficiently aggrieved by a police officer's actions that he or she feels a complaint must be lodged against him or her should be seen as a failure of policing, let alone as a demonstration of police deviance (Barker & Carter, 1991). Negative statistics about complaints against the police can serve to undermine police–community relations (Redelet & Reed, 1973; Walker, Archbold, & Herbst, 2002), both in terms of officers' willingness to engage with the community and vice versa, particularly when coupled with factors such as ethnicity or race (Brown & Frank, 2006; Lersch, 1998; Rosenfeld, 2015; Terrill & McCluskey, 2002).[AQ3]

That formal complaints do not capture the full gamut of public grievances "against" the police is already established (Adams, 1996; Brunson, 2007; Durose, Smith, & Langan, 2007; Southgate, Ekblom, & Hough, 1984). The rarity of these reports vis-à-vis the number of both voluntary and involuntary contacts that most police departments have with citizens is minuscule, and it is very likely that many "true grievances" go underreported. Notwithstanding the rarity of the "complaints phenomenon," being able to make a complaint

provides a valuable vehicle for restoring feelings of injustice in light of what the complainant had perceived as unfair treatment (Prenzler, Allard, Curry, & Macintyre, 2010; Wagner, 1980; Waters & Brown, 2000).

In practice, complaints filed against the police by citizens are currently tracked by most Western police departments. Official statistics inform us, for instance, that there is a "power few" phenomenon occurring in regard to complaints (Sherman, 2007), as some officers (Armacost, 2003; Brandl, Stroshine, & Frank, 2001; Dugan & Breda, 1991; Ridgeway, 2015) and some places (Lawton, Piquero, Hickman, & Greene, 2001) attract substantially more complaints than others. The data further show that the kinds of grievances on the basis of which people lodge complaints vary substantially between departments; Smith (2004) compartmentalized complaints in terms of unprofessional behavior, criminal conduct, tortious action, and unacceptable policy (see also West, 1988).[AQ4] These categories and measures are used in policing studies to illustrate how officers adhere to internal rules of conduct, as deviations from these regulations can potentially be construed as signals of non-compliance (Braga, Hureau, & Winship, 2008). Complaints patterns are conditional on force size, and the ways in which complaints are lodged, investigated, and processed vary (Russell, 1978), yet across jurisdictions, it can be said that more complaints equate to more problems for the department to deal with, as a higher prevalence of reports may be a corollary of a higher incidence of non-compliance to the rules that govern officers' behavior. Alternatively, more complaints could be interpreted as a good thing, because they suggest that citizens have not "lost hope in the system" and accept that processes will be revisited, officers reprimanded as necessary, and restitution made (Maguire & Corbett, 1991).

There is, however, a further category of complaints against the police that must be recognized, which adds yet another layer of complexity to interpretation. Some members of the public file frivolous complaints against officers (Prenzler, 2010) and abuse their right to complain by making complaints where none are warranted to make life difficult for the officer.[AQ5] Sometimes complainants are fully aware that, in fact, the outcome of their interaction with the police, the procedure adopted during the interaction, or both, were legal, proportional, and "professional," yet the citizen, nevertheless, complains.

Regardless of the motivation for lodging a complaint, and whatever the interpretation of the court might be, it can be agreed that complaints are costly; most departments would be only too happy to reduce them to a minimum (Liederbach, Boyd, Taylor, & Kawucha, 2007). These costs can be both direct and indirect (see Livingston, 1999). In the U.S. context especially, complaints are expensive and in large police departments can mean multimillion dollar settlements, not to mention the investigative, legal, and organizational expenses (see Lewis, 1999). Across the Atlantic, the U.K. Independent Police Complaints Commission (2015) reported a continuous rise in complaints across the majority of U.K. police forces, which has been a source of major concern in the context of police legitimacy issues (see College of Policing, 2015; Maguire, 1991; Seneviratne, 2004; Smith, 2004).[AQ6] Thus, any apparatus that can not only reduce the number of complaints but also simultaneously establish the reasons for complaints would be beneficial, especially as a counter-complaint behavioral narrative of "more" professional conduct, less use of force, and fairer processes emerges in concert. To that end, any means that can calm heated police–public encounters (see Fyfe, 1978; Prenzler, Porter, & Alpert, 2013; Sherman, 1983; White, 2001) should be welcomed, particularly when the evidence on the long-term impact of training programs or policy reforms is scarce and methodologically weak but also very costly.

POLICE BWCS

Police departments have increasingly begun using BWCs in daily operations all over the world (Friedman, 2015).[AQ7] BWCs are hypothesized to minimize the use of force in police–public encounters, reduce citizens' complaints, or to increase the accountability and the legitimacy of the police (House of Parliament, 2015; Scheindlin & Manning, 2015). At the same time, the massive growth in implementation of BWCs is not endorsed by research on

their cost-effectiveness or efficiency (Cubitt et al., 2016; Lum et al., 2015; Mateescu et al., 2015; White, 2014). Currently, there is a dearth of rigorous evaluation on the efficacy of BWC, with only a handful of rigorous, peer-reviewed studies on the effect of these devices in policing (see Ariel et al., 2015; Jennings, Lynch, & Fridell, 2015; Ready & Young, 2015) and some preliminary reports (e.g., Katz et al., 2014; Owens, Mann, & Mckenna, 2014a, 2014b). The published work concentrates on implementation processes (Drover & Ariel, 2015; L. Miller & Toliver, 2014; NIJ, 2012; Young, 2014), perceptions about the use of BWCs on policing and police professionalism (Ariel & Tankebe, 2016; Elliott, 2015; Gaub, Choate, Todak, Katz, & White, 2016; Jennings, Fridell, & Lynch, 2014; Wain & Ariel, 2014), administrative potential of BWCs (e.g., Dawes et al., 2015), perceptions of members of the public about deployment of BWCs (Pew Research Center, 2014; Sousa, Miethe, & Sakiyama, 2015; YouGov, 2015), legal issues associated with privacy rights in the public domain (Breitenbach, 2015; F. Harris, 2012; Kitzmueller, 2014; Schmidt, 2009; see also Dinger, 1999), and narrative reviews (Anonymous, 2015; Gates, 2016; Ramirez, 2015; White, 2014).[AQ8][AQ9][AQ10] Retrospective and non-controlled evaluations (e.g., Ellis, Jenkins, & Smith, 2015; Roy, 2014) exist as well, though their designs are somewhat unsuitable for estimating unbiased treatment effects (see Sherman, 2013). More impact evaluations are underway.[2]

One of the most influential experiments to date on the effectiveness of BWCs, conducted in the specific area of use of force and complaints, was in Rialto, California. The "Rialto experiment" found that (a) the likelihood of police use-of-force when officers did not wear BWCs was roughly twice than of when officers wore BWCs, and (b) the number of complaints lodged against officers dropped from 0.7 complaints per 1,000 contacts to 0.07 per 1,000 contacts (Ariel et al., 2015). Given the evidence from Rialto, some have argued that BWCs could be used as a technological fix that would revitalize police–public relations and prevent these incidents (President's Task Force on 21st Century Policing, 2015). Indeed, a U.S. district judge cited the Rialto study in a 2013 ruling against the New York Police Department (NYPD) over stop and search (*Floyd v. City of New York*, 2013).

Notwithstanding the lack of evidence, the theoretical mechanism behind BWCs is predicated on a broad and largely grounded theory of deterrence. Several lines of research across many disciplines of science suggest that people alter their behavior once made aware that they are being observed (Chartrand & Bargh, 1999; Dzieweczynski, Eklund, & Rowland, 2006; Jones & Nisbett, 1971).[AQ11] A rich body of evidence on perceived social-surveillance, self-awareness (Wicklund, 1975), and socially desirable responding (Paulhus, 1988), proposes that people adhere to social norms and change their conduct because of their cognizance that someone else is watching (Munger & Shelby, 1989).[AQ12] It seems that knowing with sufficient certainty that our behavior is being observed or judged affects various social cognitive processes: We experience public self-awareness (Duval & Wicklund, 1972), become more prone to socially acceptable behavior, and sense a heightened need to cooperate with rules (Wedekind & Braithwaite, 2002).[AQ13]

Getting caught doing something morally or socially wrong is often registered as behavior that can potentially lead to negative consequences, which is an outcome rational individuals tend to avoid (Nagin, 2013a). Of course, assuming rationality in decision-making processes has increasingly been found to be a rather loose assumption (Kahneman, 2011), but the fact remains that experiments have uncovered a propensity to avoid negative outcomes, and the findings generally agree that individuals react compliantly to even the slightest cues indicating that somebody may be watching. Priming signals to people how they ought to behave and these signals can inculcate fear of reputational damage, leading to feelings of shame and aversion to the consequences of non-compliance.

It is hypothesized that the self-awareness that arises when we are aware of being watched/filmed drives us to comply with rules/norms, primarily because of the perceived certainty of punishment. In the language of deterrence theory, cameras are viewed as "credible threats" (Jervis, Lebow, & Stein, 1989, p. 3) and as Durlauf and Nagin (2011)

wrote, "for criminal decisions, what matters is the subjective probability a potential criminal assigns to apprehension" (p. 7; see also Groff et al., 2015). [AQ14] Several authors have demonstrated some of the necessary conditions in which deterrence exerts an effect on criminal decision making (e.g., Loughran, Pogarsky, Piquero, & Paternoster, 2012; Nagin, 2013a, 2013b; Nagin, Solow, & Lum, 2015). The same might be proposed for officer compliance with police regulations because the certainty of officers being sanctioned for non-compliance with laws/rules is more forceful when cameras are on.

It is worth noting, however, that the theoretical basis for BWCs, set out in the previous paragraph, rests on citizen and officer awareness of being filmed. As Ariel (2016d) discussed, the intervention in the Rialto Experiment was not simply the presence of cameras. In addition, officers were tasked to verbally warn citizens that their encounter was being recorded. Sutherland and Ariel (2014) hypothesized,

> This verbal warning could sensitize people leading them to modify their behavior. It could also serve to remind people of the rules that are in play—politeness being the bare minimum—but other rules such as laws. Similarly, the verbal prompt may jolt individuals into thinking a little more before they act, becoming more deliberative and reflecting on future consequences. In short, there could be lots of mechanisms that account for changes in behavior when camera and verbal warning are used together.

METHOD

PARTICIPANTS

The Rialto experiment generated heated debates worldwide, particularly around the transferability of the findings to other jurisdictions or to larger police departments (e.g., Reddit, 2014). For instance, Rialto is a medium-sized force, and larger law enforcement agencies operate on a different scale (e.g., Brooks & Piquero, 1998; Cordner, 1989; Regoli, Crank, & Culbertson, 1989). Cross-national comparisons are needed as well, to verify whether or not the Rialto experiment results can be generalized to other police forces and jurisdictions.

In response to these limitations, our aim was to replicate the Rialto Experiment in police departments around the world. We invited 10 police departments (e.g., Ariel et al., 2016a, 2016b), of which seven[3] agreed to adopt the same experimental protocol that was used in the Rialto experiment (Ariel & Farrar, 2012) and test the effect of BWCs on complaints against the police under controlled conditions. While a convenience sample of police departments should not be construed as representative of the entire population of police departments, results from a seven-site design, in four jurisdictions and two English-speaking countries, provide a robust framework for testing the effect of BWCs in police operations. The research encompasses 1,429,868 officer hours across 4,264 shifts. Overall, these jurisdictions cover a total population of about 2,000,000 citizens. Information on participating sites is presented in Table 1.

MEASURES

We were provided access to the number of complaints filed against police officers at each site during treatment and control shifts, as well as the number of complaints lodged against officers during the 12 months prior to our study. As not all sites completed a

TABLE 1: Seven Participating Sites—Descriptive Statistics

Site	*Population size*	*Total arrests during RCT*	n *of shifts*	n *of frontline officers*	*Officer hours during RCT*	*Follow-up period post-RA (in weeks)*
A	161,400	1,889	462	546	221,760	22
B	285,700	590	268	46	18,224	26
C	203,800	1,097	462	111	410,256	22

E	751,500	3,390	462	870	369,600	22
G	115,000	—[a]	988	54	105,996	52
H	108,817	2,591	734	115	176,160	50
K	249,470	1,175	888	105	127,872	43
Grand total	1,875,687	10,732	4,264	1,847	1,429,868	
M	267,955.3	1,788.7	609.1	263.9	204,266.9	33.9
SD	223,107.7	1,049.7	264.0	318.7	141,990.5	13.9

Note. RCT = randomized controlled trial. **[AQ15]**

[a]Data not provided by the department.

12-month trial, the complaint counts were annualized for comparability purposes (Table 1). It must be noted that our data included complaints filed with an official complaints unit within the police department, before any investigation into the allegation had been made. Complaints against police officers can be made for a range of reasons, including (perceived) excessive or unnecessary use of police force and misconduct (e.g., incivility, lack of fairness, partiality, or any other discriminatory behavior). Given data sharing issues, we were not given access to the types of complaints, but only to the total number of complaints during each shift. We discuss these limitations below.

RESEARCH DESIGN

Beyond large-scale cluster-randomized designs, randomizing shifts is the most practical approach to implementing BWC trials with the police, as even small forces can leverage a large sample size (for a more elaborate discussion on the unit of analysis, see Ariel et al., 2015; Sutherland & Ariel, 2016; however, cf. Ariel, 2016b).[4] **[AQ16]** Each study included here was a two-arm trial that randomly assigned officer shifts to either experimental (with cameras) or control (no cameras) conditions, on a weekly basis, using the Cambridge Randomizer (Ariel, Vila, & Sherman, 2012). These treatment and control shift sequences were communicated to the patrol officers, who would be deployed on patrol with or without the BWCs. This resulted in 4,264 shifts being assigned (M = 609.1; SD = 264.0 per site), with equal allocation of day and night shifts, including days of the week. No differences were observed between treatment and control conditions in terms of the distribution of shifts.[5]

PROCEDURES

Our pre-published protocol, consented to by each police department, stated that officers on "camera on" shifts had to keep the camera on during their entire shift (typically between 8 and 12 hr) and inform members of the public, during any encounter, that they were wearing a camera that was recording their interaction (e.g., see Supplementary Materials in Ariel et al., 2016a). As discussed above, the intervention consisted of both camera and notification (on the importance of the interaction between these elements, see Ariel, 2016c; Sutherland & Ariel, 2014). To be clear, the trial design meant that officers were not able to exercise any personal discretion in deciding when cameras were turned on; cameras were on throughout their shift, during every interaction with members of the public. Officers' cameras were switched off between jobs (e.g., when traveling between calls for service) and when officers were on breaks. Leaving the decision to switch on the camera during an encounter and not before officers begin engaging with a citizen may backfire (Ariel et al., 2016a). It also defeats one of the major purposes of the camera: to record the interaction from the officer's perspective, from beginning to end, therefore providing crucial evidence of the decision-making processes that have led him or her to exercise use of force (Ariel, 2016d). The only exception to this rule was when officers responded to specific types of incidents, as pre-agreed with senior staff in each force (e.g., when conversing with informants, serious sexual assaults, or major public events).

TREATMENT FIDELITY

Maintaining consistent treatment integrity across several sites is a challenge in experimental criminology (see MacKenzie, Umamaaheswar, & Lin, 2013; Slothower, Sherman, & Neyroud, 2015; Weisburd & Taxman, 2000). To deal with risks to fidelity, each trial was managed by a local point of contact, all graduate students on the Cambridge University Police Executive Programme who were also police officers or civilian staff (what some have referred to as "pracademics"; Morreale & McCabe, 2012). This is part of the growing role of science in policing and incorporating evidence-based policies (Sherman, 2013; Weisburd & Neyroud, 2013). All research managers had undergone extensive training in experimental designs and signed up for a rigid trial protocol, committing their respected agencies to take part in the study. Fortnightly meetings were held in which research managers were asked to provide reports on implementation fidelity, any concerns raised, or issues with treatment misassignments. This process, while costly, reduced the chance of drift in implementation, although, as we discuss below, on-site management does not fully remove the risk of diffusion of treatment conditions (i.e., some police officers used BWCs when they should not have, and some police officers did not use BWCs when they ought to have done so). This in-person monitoring is bolstered by having each force consent to the trial design in advance and coming to a consensus with senior officers on implementation (see Ariel et al., 2016b). Finally, treatment integrity was monitored with a high level of reliability using BWCs metadata, which at the very least provided a reliable measure of treatment integrity (see also Ariel et al., 2016a). Cross-tabulating between date and time of evidence uploaded from the BWCS and random assignment provides a direct measure of manipulation checks. Based on these checks, protocol breaches were immediately communicated to the supervising officers (on the importance of feedback, see Fixsen, Blase, Naoom, & Wallace, 2009).

RESULTS

All sites followed an identical protocol, so we used a prospective meta-analytic approach to analyze between-group differences. We observed the number of complaints per treatment and control shifts and computed standardized mean differences for the treatment effect, presenting these in terms of Cohen's *d* (Cohen, 1988). We used the Comprehensive Meta-Analysis v.2 software (CMA) to then synthesize the results from the

TABLE 2: Citizen Complaints Across Seven Sites: Pretreatment, Post-Treatment, and Between-Group Results

Site	*Complaints before (12 months)*	*Complaints after (12 months[a])*	N *complaints treatment shifts*	N *complaints control shifts*	*Treatment shifts*	*Control shifts*	*Rate per officer (pre-treatment)*	*Rate per officer (post-treatment)*
A	558	33	7	4	183	186	1.02	0.06
B	10	0	0	0	129	106	0.22	0.00
C	331	21	3	4	184	185	2.98	0.19
E	251	30	4	6	111	188	0.29	0.03
G	24	3	2	1	489	499	0.44	0.06
H	34	19	7	12	367	367	0.30	0.17
K	331	7	4	3	445	443	3.15	0.07

[a]Annualized.

trials and present the overall results. As each trial used the same design and outcomes, it was appropriate to combine and report them accordingly (Lipsey & Wilson, 2001). The data inputted into CMA consisted of (a) the number of treatment and control shifts and (b) the number of incidents of complaints per treatment condition.

To estimate the treatment effect on a pre-test-post-test basis, we used a non-parametric analogue to paired *t* tests, the Wilcoxon signed-rank test. This test is more appropriate for paired comparisons when looking at nominal variables, when the differences are non-normally distributed, and when the overall number of pairs is relatively small (McDonald, 2014).

Across the seven experimental sites, 1,539 complaints were lodged against police officers in the 12 months preceding the study (M = 219.86; SD = 206.9), or 1.20 complaints per officer. The number of complaints lodged against the police then dropped in the post-treatment period to 113 (M = 16.14; SD = 13.1), or 0.08 complaints per officer (Table 2). This marks an overall reduction of 93% in the incidence of complaints, mimicking findings from the Rialto experiment (Ariel et al., 2015). A Wilcoxon signed-ranks test indicated that post-test complaints were statistically significantly lower than at pre-test (Z = −3.234; p < .001).

Interestingly, post-treatment between-group differences were not statistically significant (d = .053, SE = .11; 95% confidence interval [CI] = [−.163, .269]), with minimal and non-significant heterogeneity between the sites (Q = 4.905; p = .428). These results can also be read as a non-significant 10.1% reduction in the odds of a complaint lodged against officers during a treatment shift, compared with the odds of a complaint filed during a control shift (p = .629), which is likely to be influenced by the results of one study site, Site A (d = .348, SE = .67; 95% CI = [−.931, −1.720], p = .095). Our between-groups analyses are presented within a forest plot as in Figure 1, and our before–after comparisons are depicted in Figure 2.

DISCUSSION

In this multisite randomized controlled trial, we replicated the design as well as findings of the Rialto experiment (Ariel et al., 2015) in terms of the change in magnitude and direction of complaints against the police across seven distinct police departments. Using the officer's shift as the unit of analysis, we have contributed to the evidence in three major ways, and we discuss these advances below.

Site	Point estimate	Standard error	Lower limit	Upper limit	p-Value
Site H	-0.305	0.265	-0.825	0.215	0.251
Site C	-0.165	0.250	-0.655	0.325	0.510
Site E	0.075	0.217	-0.351	0.500	0.731
Site K	0.157	0.423	-0.671	0.986	0.710
Site A	0.348	0.208	-0.061	0.756	0.095
Site G	0.395	0.676	-0.931	1.720	0.560
	0.053	0.110	-0.163	0.269	0.630

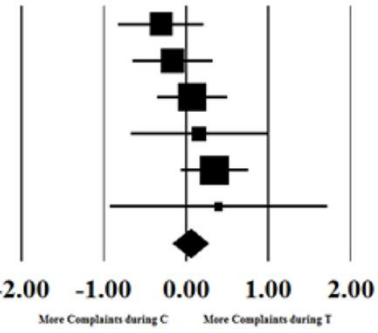

Figure 1: Complaints Against Officers per Shift. Treatment Versus Control Conditions
Note. Only six sites included in the meta-analysis, as one site—B—had nil complaints in both treatment and control arms. CI = confidence interval.

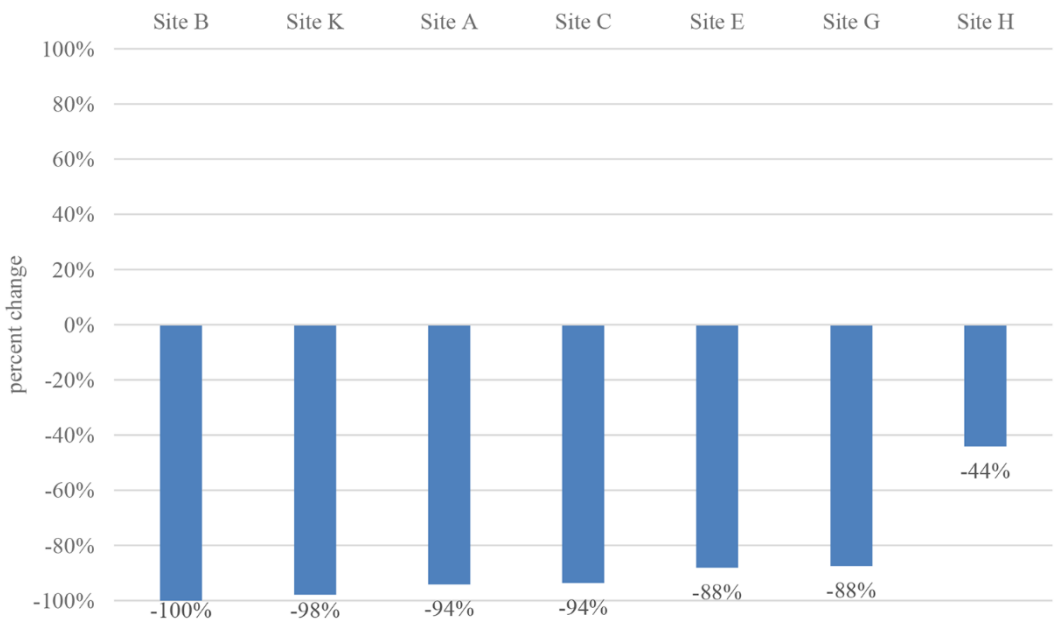

Figure 2: Complaints Filed Against Police Officers (Seven Experimental Sites): Before–After Percent Changes

A TECHNOLOGICAL SOLUTION FOR POLICE ACCOUNTABILITY, LEGITIMACY, AND POLICE–COMMUNITY RELATIONS?

Citizens initiate complaints against the police. This might be because an officer has behaved so badly that a citizen feels he must seek redress or acknowledgment, or because the citizen seeks to "make trouble" for the officer ("vexatious complaints"). We demonstrated that the use of BWCs in police operations dramatically reduces the incidence of complaints lodged against police officers, thus illustrating that the treatment effect, first detected in a relatively small force in Rialto, carried a strong external validity.

If complaints are a proxy to police (mis)conduct, and if cases of police misconduct predict perceptions of illegitimacy, such a highly significant, large drop can potentially be interpreted as a technological solution for the legitimacy problem (Hasisi & Weitzer, 2007; Myhill & Bradford, 2012; Pew Research Center, 2014a; Tillyer & Kennedy, 2008). Cooling-down potentially volatile police–public encounters to the point where official grievances against the police have virtually vanished may well lead to the conclusion that the use of BWCs indeed signals a profound sea change in modern policing—or what Sherman and Strang (2015) accurately described as a turning point in policing.

Indeed, one view of police legitimacy (e.g., Tyler, 1990) lends weight to the claim that the impressive reduction in complaints, evidenced by the research, should be interpreted as "increased procedural compliance."[AQ17] In other words, BWCs lead to increased procedural justice and, consequently, to greater legitimacy. This view is, to an extent, supported by correlational studies. Complaints are often used to measure the extent and scope of police legitimacy. For instance, Braga et al. (2008) used complaint databases in Boston as a proxy to various types of legitimacy and justice-related outcomes. If a causal link exists between procedures and legitimacy (see Jonathan-Zamir & Harpaz, 2014; Langley, 2014; Mazerolle, Antrobus, Bennett, & Tyler, 2013; Tyler, 1990; see also Stevens, Willis, & Oxby, 1981; Tankebe, 2009, 2013), then the lower rates of complaints detected across our seven tests can be viewed as a marker of enhanced perceptions of police legitimacy and satisfaction with police performance.

At the same time, BWCs are probably not the panacea for a deeply rooted issue such as police legitimacy. As the literature suggests, the correlation between complaints and legitimacy is not strong, and it is presently not at all clear to what extent the prevalence of citizen complaints correlates with general police legitimacy, beyond our assumption that they

are a proxy of the latter. Even if BWCs can lead to perfectly executed police procedures—an ideal Tylerist world of 100% implementation of procedural fairness (Tyler & Bies, 1990)—what happens before or after the encounter might still be perceived as unfair, racist, unprofessional, or malicious. Officers' decisions to arrest (Brown & Frank, 2006; Kochel, Wilson, & Mastrofski, 2011), willingness to initiate stop and frisk encounters, distributive justice, historical accounts (Braga et al., 2008), and even the co-location of crime and race (Eck & Weisburd, 1995) all affect perceptions of police and policing—and this is likely to be but a partial list of factors. That procedural justice is not the whole story of legitimacy has already been established (Bottoms & Tankebe, 2013). BWCs directly mediate the physical encounter between the police and the public but have no impact on other factors associated with legitimacy and are therefore only likely to have a limited effect on legitimacy.

Furthermore, any assumption that a change in officers' behavior in police–public encounters engenders a more "general effect on legitimacy"—that is, on the broad group of members of the community who are not directly involved in the videotaped encounter, is tenuous. Ariel (2016b) recently illustrated that wearing BWCs in one police district (Denver, Colorado) was linked to an increase in cooperation with the police compared with other districts, but only in residential and low-crime street segments (see also Murphy & Barkworth, 2014). However, to the extent that strong evidence is concerned, there is currently no research available on the effect of BWCs on police legitimacy in places where BWCs are already in use (White, 2014). Nor are there experimental studies comparing legitimacy scores in wide policing environments with and without BWCs (which are likely to be shown only through large cluster-randomized controlled trials that are difficult to implement). More broadly speaking, it would be a mistake to assume that police initiatives—even major policy changes such as BWCs—are even noticed by the public (see Brain, 2014; Home Office, 2010). Even the infamous stop-and-frisk practice in New York City carried a significant, but only modest, deterrent effect on the general offender population (Weisburd, Telep, & Lawton, 2014) and we would speculate that the recent 95% reduction in the use of stop-and-account policy by the NYPD has gone largely unnoticed by the local residents. Erroneous assumptions about how the public understands police work can be referred to as an "awareness illusion," whereby police interventions may or may not work, but the public is unaware of policy changes or their consequences. In this respect, if BWCs were to have an effect on public support for the police and on legitimacy, it would prove a difficult phenomenon to measure unbiasedly, again, without large cluster-randomized trials. We raise this matter given the weight assigned to legitimacy in the discourse on BWCs, even though our study does not offer evidence of this sort, and call for more research on the link between complaints and general legitimacy within, or without the specific context of BWCs.

CONTAGIOUS ACCOUNTABILITY

However, in terms of police accountability, BWCs can very well be construed as a "fix" (for a wider discussion regarding technology usage in law enforcement, see King, 2000; Weiss, 1997). We hold this view because complaints reflect most directly on procedural compliance (irrespective of legitimacy), and procedural compliance is an essential part of the definition of accountability (United Nations Office on Drugs and Crime, 2011):

> [A] system of internal and external checks and balances aimed at ensuring that police carry out their duties properly and are held responsible if they fail to do so. Such a system is meant to uphold police integrity and deter misconduct and to restore or enhance public confidence in policing. Police integrity refers to normative and other safeguards that keep police from misusing their powers and abusing their rights and privileges. (p. 9)

There can be no doubt that BWCs increase the transparency of frontline policing (Ready & Young, 2015; Scheindlin & Manning, 2015). Anything that has been recorded can be subsequently reviewed or scrutinized. Individual officers become more accountable as BWCs

accentuate the need for oversight and reflection on their own actions (Lumina, 2006; Reiner, 1993; Walsh & Conway, 2011). BWCs thus sit squarely within what D. A. Harris (2010) referred to as a "holistic" approach to police oversight, which "combines the traditional 'reactive' functions (i.e., tracking cases of individual misconduct) with 'proactive' functions designed to promote organizational changes that might reduce individual misconduct" (p. 240). In turn, greater transparency not only primes power-holders to adhere to protocols, guidelines, and "best practice" (for the same reasons we alluded to earlier in the context of deterrence theory and self-awareness) but also creates an equilibrium between the account of the officer and the account of the suspect about the same event (see Frank, Smith, & Novak, 2005). Without corroborating evidence (e.g., bystanders testimonies, forensic evidence), a complainant would find it difficult to prove, in the necessary forums, that police misconduct had, indeed, occurred (see J. Miller & Merrick, 2002). With the evidence from BWCs, however, an officer's transgression can be revealed and legally proven, much as a suspect's transgression can. For this reason, there will be little value of recording incidents as they become more aggressive; they need to stay on for the entire police–public contact, from start to finish, and as a method of corroborating the rationale for any action that the officer has taken to deal with the situation. In deterrence theory terms, the "credible threat" of apprehension is elevated through complete transparency, to the extent that complainants now have the necessary evidence to support their claim that they have been wronged by the officer (and, again, vice versa). The awareness of being videotaped leads to transparency, and transparency leads to accountability; ergo, greater awareness leads to increased police accountability. Our evidence supports this model.

Our findings, however, go beyond this model. The pre/post results show that the introduction of cameras reduced complaints overall, but why were reductions seen in both the treatment and control arms of the trial? There are several ways to interpret these results. We offer two approaches in the context of accountability although they are not mutually exclusive. Both interpretations are born out of our research design, where the unit of analysis was the officer's shift.

Experiments test treatment contrasts rather than single treatments (Holland, 1986). Cook and Campbell (1979) drew attention to a novel threat that affects this treatment contrast, without necessarily influencing the major treatment purportedly under test: diffusion of one treatment to other treatments. Reflecting on this diffusion, Cook and Shadish (1994) wrote,

> Treatment providers or recipients learn what other treatment groups are doing and, impressed by the new practices, copy them. . . . It threatens to bias treatment effect estimates when compared to situations where the experimental units cannot communicate about the different treatments. Statisticians now subsume them under the general rubric of SUTVA—the stable-unit-treatment-value assumption (Holland & Rubin, 1988)—in order to highlight how much the interpretation of experimental results depends on the unique components of one treatment group not diffusing to other groups and hence contributing to the misidentification of the causal agent operative within a treatment contrast. Is it the planned treatment that influences the outcome, or is it serving in a control group? (p. 555)[AQ18]

Our study somewhat reflects this scenario. While we found a reduction in complaints in pre-post comparisons, we also found no significant differences across the seven tests when looking at post-treatment comparisons, and no discernible heterogeneity between the sites. Taken at face value, it may be unclear as to whether deterrence theory was wrong, the implementation weak, or the statistical power to detect the effect of BWCs on police behavior inadequate. As the same officers wore the cameras and did not wear the cameras, we cannot rule out a violation of stable unit treatment value assumption (SUTVA) and treatment diffusion (Ariel et al., 2015; Bloom, Bos, & Lee, 1999; Sampson, 2010).[AQ19] The same officers, all of whom were participating in the same program, experienced both treatment and control conditions. We return to these points when discussing the additional limitations of our

study below. However, the evidence must be read holistically, along with the before–after comparisons. This multisite experiment provides direct evidence that repeated and systematic exposure to a stimulus that elicits deterrence can increase accountability, even when the stimulus has vanished. Our officers learned, by their repeated exposure to the surveillance effect of the cameras, what normative or appropriate reactions are, even when they were no longer under surveillance. This may be true for officers who once wore BWCs and no longer do (through the process of random assignment), or officers in the department who did not take part in the experiment (e.g., neighborhood police teams, special victim support units, etc.). We argue that that BWCs affect entire police departments through a process we label *contagious accountability*. Perhaps naively, we find it difficult to consider alternatives to the treatment effect beyond the panopticonic observer effect when the reduction in complaints is by nearly 100%. Whatever the precise mechanism of the deterrence effect of being watched and, by implication, accountability, all officers in the departments were acutely aware of being observed more closely, with an enhanced transparency apparatus that has never been seen before in day-to-day policing operations. Everyone was affected by it, even when the cameras were not in use, and collectively everyone in the department(s) attracted fewer complaints.

There is, however, a caveat associated with this conclusion, which is important for future experiments on BWCs. It is not the camera device alone that caused the contagious accountability, but rather a two-stage process. First, the treatment effect incorporated the camera as well as a warning at the beginning of every interaction that the encounter was being videotaped. We urge practitioners to acknowledge that the verbal warning, which our protocol dictated should be announced as soon as possible when engaging with members of the public, is a quintessential component of the treatment effect. It primed both parties that a civilized manner was required and served as a nudge to enhance the participants' awareness of being observed. Without the warning, the effect might easily have been reduced or failed to materialize.

The second element to the process is the need for affirmation that the videotaped footage can be used. People may be aware of CCTV or bystanders filming the encounter but still conduct themselves inappropriately, believing the camera to either not be recording or not monitoring their demeanor. Without the actualization of the warning, transgressors may be quick to assume that the threat of apprehension and risk of sanctioning are not real. Therefore, the fact that the officially collated, recorded footage can be used against the participants moves this intervention from being a "toothless policy" (Ariel, 2012, p. 57) into an effective technological solution.

UNTANGLING OBSERVER EFFECTS

Since BWCs came out, a question has been raised about which person in the encounter is most affected. Does the self-awareness effect or the "announcement effect" of surveillance (see Mann, 2002; Surette, 2005), concentrate on the person the camera is videotaping or the person holding the camera?**[AQ20]** Does the causal chain start with the police officer, who is holding the BWC and is deterred from reacting with excessive or unnecessary demeanor, or does it first cool down the aggressive demeanor of the suspect?

On one hand, most empirical research suggests that the suspect's actions and resistance during police–public encounters precipitate force reactions from police officers (Alpert & Dunham, 1997; Alpert et al., 2004; Crawford & Burns, 2002; Engel, Sobol, & Worden, 2000; Terrill, 2001; Terrill & Mastrofski, 2002; Worden & Shepard, 1996).**[AQ21]** This demeanor hypothesis (Croft & Austin, 1987; Engel et al., 2000; Garner et al., 2002) shows that the relationship between police–public encounter characteristics and police use-of-force is significantly dependent on resisting arrest and therefore concentrates on the suspect.**[AQ22]** If the effect of BWCs is wholly on the suspect, then BWCs should deter members of the public from resisting arrest, "talking trash" to the officer, or exhibiting behavior that may result in force being used by the officer. This in turn might lead to officers refraining from

reacting in ways that will more likely lead to a complaint. If this is the case, then the causal chain starts with the suspect:

> BWC + verbal warning → suspect's demeanor "cools down" → officers do not "react" aggressively → fewer complaints than without BWCs.

Alternatively, the amount of response officers use is dependent on the cognitive and emotional capacities of the officer, as well as his or her training and experiences (see Paoline & Terrill, 2011). Common to all organizations, the police force includes "thin skinned" individuals who are primed to act aggressively when encountering certain members of the public (Holmes, 2000; Reiss, 1968; Worden, 2015; see also Fielding & Fielding, 1991). Recent studies have shown that identifiable officer characteristics can predict police behavior; for instance, officers with prior problematic performance are 3 times more likely to discharge a firearm (Ridgeway, 2015). Likewise, the ability to de-escalate a situation is critical in public–police encounters (Sherman, 1983), and while the officer ought to respond with a proportionate "response dosage," he or she also needs to have the skill to reduce the need to engage in what can, in hindsight, be construed by the citizen as misconduct. Under this model, the causal chain is as follows:

> BWC + verbal warning → officer's reaction to suspect's demeanor is "cooler" → fewer complaints than without BWCs.

It is also worth considering the possibility that both the officer and suspect are being affected simultaneously. The mechanism that might bring this about is as described above; the verbal reminder, combined with the camera, "cools down" both participants at the same time, owing to the jolt of the verbal reminder of being watched, nudging them to think about their actions more consciously. This might mean that officers are beginning encounters with more awareness of rules of conduct (internal behavior control), and suspects' demeanor is less likely to elicit an aggressive response because they are aware they are being watched (external behavior control). In this scenario the causal chain would be

> BWC + verbal warning → officer's starting point for the interaction is cooler + suspect's demeanor "cooler" → officers less likely to react aggressively → fewer complaints than without BWCs.

Leaving aside the officer–suspect interaction, we have sufficient evidence to conclude that the BWC had a more pervasive effect on officers. This conclusion is based on the contagious accountability effect of the intervention across all sites. Because we detected a reduction in complaints in our before–after analysis, we conclude that officers changed their behavior in encounters during control conditions as well as treatment conditions. To use an analogy from the medical world, suspects were not given the medication during control conditions, but officers were. The treatment effect carried over to no-treatment shifts as well, and officers' behavior was affected by it. The alternative model—that the chain of causality begins with the suspect—requires us to assume that the suspect anticipated that his interaction with the officer would be videotaped and therefore significantly amended his behavior so as not to provoke the officer. This seems unlikely in the many thousands of interactions we have reported, and we are not aware of any large-scale advertising campaigns having taken place at research sites that might have alerted suspects to the use of BWCs beforehand, the evidence about the ineffectiveness of such campaigns notwithstanding. It is therefore logical and more parsimonious to conclude that, insofar as the demeanor hypothesis is concerned, BWCs offer a novel nudge for de-escalation by affecting police officers' approaches to encounters.

REVISITING THE SUTVA CONCERN

The fact that officers participated on multiple occasions in both treatment and control conditions, potentially violates the SUTVA (Rubin, 1986) and/or the requirement that observations are independent. However, the unit of analysis is the shift, not the officer. The set of conditions encountered in each shift cannot be repeated, because time moves on in one direction. The treatment manipulation was whether the shift involves cameras and a verbal warning or no cameras and no verbal warning. Outcomes (complaints) filed by citizens against officers are essentially driven by how officers act and citizens perceive those actions. Likewise, because the shift was randomized and officers experienced multiple shifts with and without cameras, we know that on average, all else was equal—including which officer was involved. Being able to define units, treatments, and outcomes in this way makes SUTVA less plausible (Rubin, 1986). However, spillover effects often result from experiments, and indeed may be the intention (Angelucci & Di Maro, 2016). In all of our tests reported above, officers were exposed to both treatment and control conditions. This is akin to a cross-over trial with more than one switch between conditions for each officer. The spillover, we propose, is that officers in control conditions were influenced by their counterpart treatment conditions and altered their behavior enough, overall, to reduce the likelihood of complaints being filed against them, regardless of treatment condition. In short, we believe that our results point to the success of this intervention in terms of complaints: Complaints fell so far, as much as to zero, that between-group differences in treatment conditions could not be detected.

As we noted earlier, despite these reservations about our choice of unit of analysis, it is still the best of all possible worst options of units of analysis. With every officer on the shift either wearing the device or not, depending on the research protocol, and with every participating force maintaining the same random allocation of shift patterns, the shift is the best option for non-cluster experiments, given the practical and statistical implications briefly mentioned herein. In addition, there is also a good reason for using a shift-based approach, particularly in light of our interest in the impact of the observer effect. Suppose that indeed there is a 100% diffusion effect, namely, that policing during control conditions is completely affected by policing during treatment conditions. This means that officers modified their behavior entirely and followed protocol when cameras were used and when the cameras were not used. One way to interpret this is clinically, as a contamination effect. Following this line of reasoning, then one may simply ignore the between-group analyses of our findings and view this as a pre-experimental multisite study, with seven independent before–after analyses. The results still seem informative to us. Nonetheless, from a theoretical as well as practical perspective, the entire premise of BWCs in police operations is to cause a change in behavior through the deterrent effect of being observed. The diffusion of the treatment effect, if it exists, is the treatment effect, which therefore makes the lessons learned from our estimates even more informative. The purpose of BWCs is to modify behavior, and ultimately this was achieved, with direct and unmediated measures of change. Therefore, the implications for learning theorists, psychology scholars interested in priming or nudges, are consequently substantive.

With these in mind, there are also implications for social policy researchers, who would benefit from understanding that a social control apparatus can be implemented partially (e.g., in half of the temporal shifts rather than all the time), and still provide a desirable effect; the cost and resources consequences are therefore noted as well, particularly in an age of public domain austerity. One can achieve the same goal of reducing complaints against the police through the systematic application of BWCs at half of the costs (for a more elaborate discussion on these cost implications, see Sutherland & Ariel, 2014).

ADDITIONAL LIMITATIONS

We cannot rule out the possibility that forces' policies on how complaints were handled and processed changed over the period of the experiment, for example, by forces making complaints more difficult for citizens to register, or by ignoring them. The literature on how written policies and their enforcement or non-enforcement influence officer behavior is developed. This line of research observed that policies could influence street officers'

behavior. However, the ways in which policies modify behavior is heavily dependent on how the policies are enforced in the context of use of force guidelines (see Fyfe, 1978; Sherman, 1980; Sherman & Langworthy, 1979; White, 2001), pursuit policies (see Hansen, Rojek, Wolfe, & Alpert, 2015), and in the area of domestic violence mandatory arrest policies (see Eitle, 2005).[AQ23] If we assume that was the case, then it would be disingenuous in the extreme if forces were already under pressure to reform and improve officer behavior (and be seen to do so), while making it more difficult for citizens to complain. However, the level of Machiavellian pre-planning that would have been required for this to happen, in all participating forces, just in time for the trials, seems highly implausible.

Future studies will need to look into the type of complaints that are reduced as a result of the intervention and against whom these complaints are filed. As in the Rialto experiment, we did not observe the categories of complaints filed against officers. It may be the case that certain kinds of complaints are reduced because of wearing BWCs, whereas others are not. Recently, Ariel (2016a) has shown a reduction in complaints against use of force but an increase in complaints against misconduct because of using BWCs. The situations from which complaints arise most often, whether during police-initiated contacts, emergency calls for service, multiplayer events, or solo contacts, as well as offense types, also bear future scrutiny as these may condition the effect of BWCs and are presently unknown. It is also important for future research to more closely observe how the cameras affect the power few (Sherman, 2007) of complaints, as we already know that certain officers are more likely to attract complaints than others. These are important avenues for future research, as they may carry cost-benefit implications for police departments around the globe.

NOTES

1. One possible solution could be systematic social observations of police patrols with or without body-worn cameras (BWCs; Mastrofski, Parks, & McCluskey, 2010; Sampson & Raudenbush, 1999); however, when considering the first challenge mentioned above (rarity of events and wide spatiotemporal dispersion), one would have to assign costly observers to every police vehicle, during all hours of the day, for many months. This was beyond the scope of the present study, nor would qualitative ride-a-longs methodologies go without critique (Mastrofski & Parks, 1990).
2. http://www.justice.gov/opa/pr/justice-department-awards-over-23-million-funding-body-worn-camera-pilot-program-support-law
3. Names of departments omitted for the blind review process.[AQ24]
4. While randomly assigning shifts as the unit of analysis is not ideal, given the potential spillover effect, it represents the best of all worst options (what is sometimes referred to as the maximin rule—that is, to compare the alternatives by the worst possible outcome under each alternative and choose the option which maximizes the utility of the worst outcome). As alluded to by Ariel et al. (2015)—but shown more strongly in light of recent evidence—the risk for a contamination effect is substantially more concerning, almost detrimental, when the unit of analysis is the single officer or the squad (Jennings et al., 2015), the borough (Owens, Mann, & Mckenna, 2014), or studies where the allocation was confounded by geography (Ready & Young, 2015)[AQ25]. First, if the unit of analysis were the individual police officer (or squad), then the contamination would be even greater: The overwhelming majority of police responses that result in complaints have at least one or two more police officers arriving at the scene as backup (Ariel, 2016).[AQ26] Suppose the primary/first responding officer is a "control officer," but the second officer arriving at the scene as backup is a "treatment officer," then would this case be a "treatment case" or a "control case?" Suppose that both the primaries as well as the secondary response units are "control officers," but the third unit is a "treatment officer?" Furthermore, another reason why using the individual officer as the unit of analysis seems wrong is that it dismisses group dynamics and organizational factors that cannot be controlled for. There are underlying forces and cultural codes of behavior that characterize entire shifts, such as the character of the "officer in charge" (OIC) or the sergeant managing the shift, the degree of officers' cynicism, comradery, codes of silence, and a host of institutional undercurrents that are recognized in the literature—but cannot be factored into a statistical model.

The best methodological solution for this spillover effect problem is to randomly assign entire forces, or a population of geographic divisions and sectors, in what are called cluster random assignment designs (see Bloom et al., 1999; Cook, 2005). Yet from a practical perspective, this *beau idéal* methodology is close to impossible to implement in policing because it requires many police forces to achieve sufficient statistical power. This is why, to

the best of our knowledge, cluster-randomized trials are not common in criminology. Obtaining the necessary buy-in from these forces, securing thousands of BWCs for treatment conditions, having full-time field management to this scale, and safeguarding treatment integrity, are only but a few of the operational challenges that such a design would entail, which would go beyond the scope of our project—which had zero funding of any kind. For example, obtaining permission to conduct a multisite experiment with a selective sample of up to 10 small- to midrange forces and local divisions was lucky; but systematically assigning a "sufficiently large" sample of forces for a clustered design, who would all commence their trial simultaneously, is a dream not imagined. Simply, from a statistical power perspective (see Cook, 2005; Cohen, 1988), we would require such a large number of willing and able forces to randomly assign into treatment and control arms, which we cautiously conclude that this design is practically impossible, at least in BWCs research at present.

5. Following consolidated standards of reporting trials (CONSORT) guidelines, we did not test statistically for differences between trial conditions (http://www.consort-statement.org/checklists/view/32-consort/510-baseline-data). If the reviewer(s) see fit, we can report these distributions as well. **[AQ27]**