Journal of Experimental Criminology
https://doi.org/10.1007/s11292-018-9344-4

Preventing treatment spillover contamination in criminological field experiments: the case of body-worn police cameras

Barak Ariel[1,2] · Alex Sutherland[3,4] · Lawrence W. Sherman[5,6]

Published online: 27 November 2018

Abstract

Objectives A central issue in experiments is protecting the integrity of causal identification from treatment spillover effects. The objective of this article is to demonstrate a bright line beyond which spillover of treatment renders experimental results misleading. We focus on a highly publicized recent test of police body cameras that violated the key assumption of a valid experiment: independence of treatment conditions for each unit of analysis.
Methods In this article, we set out arguments for and against particular units of random assignment in relation to protecting against spillover effects that violate the Stable Unit Treatment Value Assumption (SUTVA).
Results Comparisons to methodological solutions from other disciplines demonstrate several ways of dealing with interference in experiments, all of which give priority to causal identification over sample size as the best pathway to statistical power.
Conclusions Researchers contemplating which units of analysis to randomize can use the case of police body-worn cameras to argue against research designs that guarantee large spillover effects.

Keywords Body-worn cameras · Unit of randomization · Spillover effects · SUTVA · Interference · Partial interference · Experiments

✉ Barak Ariel
ba285@cam.ac.uk; barak.ariel@mail.huji.ac.il

Alex Sutherland
alex_sutherland@rand.org

Lawrence W. Sherman
ls434@cam.ac.uk

Extended author information available on the last page of the article

In any counterfactual evaluation, experimenters try to establish "what would have happened otherwise." In the case of randomized designs, those units in the treatment and control groups should be exchangeable (see Hartman et al. 2015). In quasi-experimental designs, the analysis mimics a randomized controlled trial (RCT) by conditioning on control variables, by matching, or by using an instrumental variable approach (see Morgan and Winship 2007, 2015). Whether a treatment is randomized is arguably less important than whether it is compared to similar units that do not receive the treatment. The logical meaning of "what would have happened otherwise" collapses if the control group receives the treatment (Nagin and Sampson 2018). Yet in the pursuit of other important principles, such as sample size, researchers can sometimes neglect the primacy of the counterfactual principle. Perhaps the time of greatest risk for that loss comes in the selection of units of analysis.

Several factors influence the choice of the unit of randomization in an RCT. Cost is one. Other factors include information about the intervention design, its delivery logistics (see e.g., Craig et al. 2008). In simple interventions such as a tax compliance letter, these issues are less problematic, although still salient. In more complex social interventions involving interactions between people and/or places, the unit of randomization must be directly informed by the nature of the intervention. Specific features of interventions mean that one unit of randomization is more suitable than another for plausibly answering the question, "What would have happened otherwise?"

A fundamental rule of RCTs is the axiom to "analyze as you randomize" (Boruch 1997:203; Senn 2004). Unlike non-experimental designs, in which the choice of analysis unit can be more dynamic, experiments are stuck with the units to which treatments were assigned. Whichever unit was chosen as the unit of random assignment in the experimental protocol should be the unit of analysis to estimate the causal relationship between the independent and dependent variables. Deviations from this rule are possible, but when they occur, the grading of the study is automatically reduced from a true experiment to a quasi-experimental design. The key message is that the unit of randomization matters immensely in experimental criminology.

A corollary of the analyze-as-you-randomize principle is the "independence principle": that there should be integrity in treating each unit with independence from the ways in which other units are treated (Gottfredson et al. 2015). Failure to adhere to the independence principle a well-known but often neglected issue with spillover effects. A major critique of field experiments, in fact, suggests that the principle is so difficult to follow that many randomized trials lack internal validity (Sampson 2010). While we disagree with Sampson's conclusion that randomized trials are at greater risk of this threat per se, we agree with the crucial importance of the principle.

In this article, we provide a clear demonstration of Sampson's (2010) concern for the potential violations of the independence principle, known to statisticians as "SUTVA"—the Stable Unit Treatment Value Assumption. Our case in point is a highly publicized experiment (see Ripley 2017) on the effect of police body-worn cameras (BWCs) on the rates of documented use of force and civilian complaints against police officers. Some 2000 police officers were divided randomly to two groups: a treatment group instructed to wear BWCs while on patrol and a control group who were not given the devices. The unit

of randomization was the individual officer. This study could have been powerful enough to detect small effect sizes.[1]

However, there is a catch: the design does not take into account the fact that many—if not most—police–public encounters that require use of force or could lead to complaints are dealt with by at least two officers. For example, many police patrol cars in the US are assigned to have two officers work together. Even with one-officer cars, the odds of two cars responding to the same encounter are high. Given this fact, assignment of cameras to individual officers creates a strong degree of treatment "spillover" (diffusion). Control group officers (with no cameras) who attend calls with treatment officers (who are wearing cameras) are, by definition, contaminated. By being exposed to the (manipulated) presence of the camera for the treatment officers, the control officers' treatment is no longer independent from the treatment of the experimental officers. The control officers may behave differently when working with a camera-wearing officer than when a camera is not present. The risk of spillover becomes even more pronounced when three or more officers attend the same encounters. This means that the proposed study's fidelity is at risk due to the unit of random assignment, because of experimental circumstances in which both arms are exposed to the same intervention.

Such spillover is exactly what occurred in the BWC experiment with the individual officer as the unit of analysis (Yokum et al. 2017). As one might expect, the RCT concluded that the intervention was not effective in reducing rates of either complaints or use of force, when comparing officers assigned cameras to officers who were not. It appears that the contamination is so extensive that an "intention to treat" analysis—that is, one in which all units are analyzed in the groups to which they were randomized—would result in no measurable impact. Such a study was, by the most basic principles of field experiments, not capable of fairly falsifying the null hypothesis of no differences between outcomes of two different study conditions. The conditions were virtually identical in both groups: encounters with citizens in which some officers wore cameras and others did not. If the conditions are identical, no test of causality is possible.

This example illustrates the importance of the initial choice of the appropriate unit of randomization in designs for experimental criminology. At the planning stage, experimentalists face tough choices that are simultaneously theoretical, statistical, and practical: should we randomly allocate individuals? If that does not allow independent treatment of each unit, then what about places? Or different times of day? Or clusters of any of the foregoing? The decision is critical. Ultimately, the choice of unit is a compromise between the best unit in principle and the optimal unit possible. It may also mean that scientists can make science better by refusing to conduct "randomized" trials when they know *in advance* that the treatments received cannot possibly be kept independent of each other for each unit (or most units) of analysis.

This case study begins by discussing the general problem of spillover and how critical it is when estimating the overall treatment effect. Next, we provide a strategic approach to tackling the spillover problem through careful pre-test planning. Because

[1] Even with a very conservative power calculation where alpha is 5%, desired power is 80%, 50:50 treatment:control allocation and no baseline variables, the minimum detectable effect size is $d = 0.125$ (Calculated using PowerUp!; Dong and Maynard 2013).

most field trials would suffer some type and degree of spillover effects, our recommendation is not to abandon experiments altogether (Greene 2014),[2] but rather to craft experiments that will minimize the risk of spillover effects as much as possible. That task can be accomplished by emphasizing independence over sample size in choosing the unit of analysis. Causal identification, and not sample size, has already been found to matter more in one large review of criminological field experiments (Weisburd, Petrosino and Mason 1993). Causal identification is, both empirically and theoretically, the most appropriate criterion for choosing the unit of analysis—whether individuals, places, groups, times, shifts or clusters.

We conclude this article by showing that some studies may suffer interference but result in significant results despite *modest* amounts of treatment spillover. Such experiments can be said to have arrived at a more conservative estimation of the treatment effect, but in the hypothesized direction. Moreover, some studies have positive spillovers that can be said to be desirable outcomes, thus contributing to our understanding of group dynamics, learning theories, and cost-effectiveness dilemmas. Yet these possibilities do not in any way alter the bright line between a massive and a minor violation of the SUTVA.

The spillover problem in randomized trials

Major interference

In a randomized experiment, we expect that the outcome of one unit does not depend on the outcome of any other unit. When there is interference, we can assume that the treatment effect is either inflated or deflated, meaning that the true impact of the intervention on the outcome is masked to some degree, depending on the extent of contamination. This is called the "spillover effect." There are two broad types of spillover effects: major interference and partial interference (Sobel 2006). Major interference is the contamination of the control group, whereas partial interference means spillover effects within the same treatment group. Both types are important, but partial interference is a relatively new topic of interest for experimentalists (Baird et al. 2016). We discuss major interference here and partial interference in the next section.

Spillovers in randomized trials corrupt the core counterfactual comparison of the experimental design. The spillovers can operate at different levels, bleeding from treatment to control, between different treatment groups, within statistical blocks or clusters or within individual treatment units (Baird et al. 2016; Campbell and Stanley 1966; Shadish et al. 2002). For example, when the threat of spillover comes from major interference of the treatment group treatments into the control group, it leads to contaminated control conditions; this challenges the desired counterfactual contrast between units that were exposed to the intervention and units that were not. Rubin (1980; see also Cox 1958) and others refer to this type of contamination as a violation of the SUTVA. Put another way, when conducting an experiment—following Cox (1958) and Rubin (1980)—we are assuming that the effect of an intervention on any

[2] Recalling that the origins of experimental science were actual trials in fields conducted by Sir Ronald Fisher (see the discussion in Armitage 2003).

one individual/unit, "unit A" for example, is unrelated to the treatment assignments of other people/units in the study (units B, C, D and so on).

When spillover occurs, participants (or units) in the control group experience a direct or indirect treatment effect from the program. While not allocated to the experimental group, controls may experience a spillover from other individuals/units who were assigned to a treatment group. In the case of *spillover from treatment to control*, in which everyone gets some treatment, differences between the two groups are shrunk. This damages the primary intention to treat (ITT) analysis (Peto et al. 1976). This "analyze as you randomize" rule is the preferred method of dealing with crossover among medical scholars (e.g., Armitage 1991). Because the ITT is the only point at which differences are truly randomized, it is the only point of sorting units that has the logical power to eliminate rival hypotheses by "controlling" for baseline differences across units. Analyses of compliance with allocation subsequent to randomization, although potentially informative, suffer from the limitation that compliance is non-random. Since only the ITT analysis can hold all other factors (except for the treatment) equal, then there is limited value in analyzing any other comparisons besides groups divided by that randomly assigned intention.

While spillover effects are problematic, they are often unavoidable. Some studies have therefore dealt directly with ways of minimizing the threat of spillover to internal validity. As Gilbert et al. (2016:1) point out, the literature includes studies…

> that uncover network effects using experimental variation across treatment groups, leave some members of a group untreated, exploit plausibly exogenous variation in within-network treatments, or intersect an experiment with pre-existing networks. Further progress has been made by exploiting partial population experiments, in which clusters are assigned to treatment or control, and a subset of individuals are offered treatment within clusters assigned to treatment.

The authors' conclusion is that major interference is part and parcel of studies involving human beings, so we need to "relax the assumption around interference between units" (ibid). However, interference cannot be completely ignored; Baird et al. (2016) do not advocate this, nor do we. Empirically, the presence of spillovers may vary widely, leading to the same question that faces ITT itself (Peto et al. 1976): how much is too much? If only 10% of an intended treatment group is actually treated, compared to 5% of a no-treatment control group, many would think that ITT analysis is pointless. Yet if 85% of a treatment group received treatment, and only 15% of the controls did, there may be source of high validity for the ITT analysis.

Similarly, if 5% of controls experience spillover, we might think the ITT analysis would still have high internal validity—but not if 85% of controls experienced spillover. The issue in both cases is not whether imperfections exist, but how much tolerance the design has for such imperfections, as the history of precision engineering clearly demonstrates (Winchester 2018). The more pronounced the spillover effect, at least for a treatment that truly has an effect, the more likely the study will result in no difference (or non-significant differences) between study arms.

In these situations, we cannot actually determine whether the treatment does not have an effect, or if in fact the study's design (and SUTVA violation) made it impossible to detect an effect. This is the fundamental problem with SUTVA violations

and why we should acknowledge them. Less prosaically, this technical violation of the experimental design undermines the conclusions drawn, meaning that policy recommendations are based on flawed evidence. Nonetheless, we can see that the question of where to draw the line remains central (Sherman 1993).

While policing experiments may have underemphasized the issue of spillover and interference effects (but cf. Braga et al. 2018), a relatively developed and formalized literature in other experimental disciplines has paid closer attention to these concerns—mainly in statistics (see an early review by Page 1978 and more recently by Bowers et al. 2018, see also Hudgens and Halloran 2008 and Rosenbaum 2007).[3] Kruskal (1988), for example, discusses the causal assumption of independence and makes a key observation. "[I]ndependence seems rare in nature, and when we really want it, we go to great pains to achieve it, for example […] in randomization for allocation of treatments in experiments. […] An almost universal assumption in statistical models for repeated measurements of real-world quantities is that those measurements are independent, yet we know that such independence is fragile" (p. 935–6). At the same time, common statistical models assume independence, and when interference occurs, some fundamental assumptions of these models are not met.

Partial interference

A second component of the spillover problem is often overlooked: partial interference (Sobel 2006). For purposes of simplicity, we can define this problem (in experimental criminology) as the effects of treatment heterogeneity on the treaters, the treated, or both, which may then amplify or restrict the level of heterogeneity in the treatment actually applied to the units being treated. Beyond the assumption that the outcome of the control units will not depend on the outcome of the treatment units—and vice versa—we may also assume that, for a test of causal "efficacy," a single version of each treatment level is applied wholly to each experimental unit—"treatment homogeneity." For example, every police officer assigned to wear BWCs will use the device across all (eligible) interactions with members of the public, without exception. (Note that in "effectiveness" trials, this assumption is often relaxed, suggesting that all field experiments in criminology might be better thought of as effectiveness trials with heterogeneous treatment delivery rather than as efficacy trials with homogeneous treatment.)

Likewise, in an efficacy study (Gottfredson et al. 2015) of the effect of text messages sent to remind officers to activate their BWCs, the assumption was that every participating officer had received, read, and then acted upon the message in the same way (as implausible as that is). To emphasize, the same assumption about treatment homogeneity also applies to the other trial arms. That is, if there are more treatment conditions, then we assume that each condition was adhered to equally across units and that, crucially, the control condition (whether they receive placebos, no-treatments, business

[3] Still, these disciplines are not immune from these errors. Perhaps one reason for misunderstandings about independence is inadequate training in classrooms and lectures: "In a modest probe, I looked at two issues of the Journal of the American Statistical Association and counted 11 reviews of introductory textbooks. I inspected the six of these books in our library and graded their treatments of independence tolerantly: no As, one B, two Cs, one D, and three Fs, an unhappy record" (Kruskal 1988, footnote 135)

as usual interventions, or anything else) was maintained fully and equally across units randomly assigned to the control group.

However, in experiments in which either the treated units or their treaters interact with one another—police officers working in the same department, pupils in the same school, patients in the same care facility, offenders living in the same community—the effect from one treatment arm may often spill over to other treatment arms. Participants in the experimental group are exposed not only to the direct treatment effect from the program to which they were assigned; they are also exposed to and experience the spillover effect from the treatment of other participants in their treatment group, which may act to reinforce treatment effects. For example, if the officers are asked to video-record their interactions with members of the public, and they are often video-recorded by other officers who wear cameras and have attended the interaction, then they may be more likely to comply with the rules themselves if they know other officers are complying (and whose video-recorded evidence can get them into trouble). Even if they are not directly recorded by other officers' cameras, those other officers may behave differently because their actions *might be* recorded (assuming of course there is a true deterrent treatment effect of fearing that BWCs are recording officers' behaviors).

Consequently, there are two overlapping treatment effects: first, a direct treatment effect on treatment units (absent of any spillover effects); and second, a reinforcement spillover effect on the treated caused by other treatment units. This effect of treatment of one unit on the treatment effects of another unit is what is meant by partial interference. Thus, the ITT analysis includes the sum of these two effects of what we call treatment interference and what is called, in relation to spillover effects, partial interference.

We find the concept of partial interference to be badly labeled. For present purposes, we would prefer to describe it as "contagion effects," or "synergistic effects," similar to the concept of "herd immunity" in vaccinations (Anderson and May 1985). The basic idea is that when a critical mass of treatments of individuals is delivered, then the effect of treatment on the treated is magnified by synergy of spillover across units within the treatment group.

Understanding what is called partial interference (or contagion effects) has direct implications for policy because it addresses two interrelated issues: treatment intensity and group dynamics.

Treatment intensity Treatment intensity, or dosage levels, is a measure designed to detect the level of treatment applied that is necessary to cause some level of an effect. In the study described above, assume that the protocol dictates that police officers are supposed to use cameras in every police–public encounter. However, some officers deliberately breach protocol and do not record public disorder and police-initiated contacts (e.g., stop and search and checking suspicious vehicles) because these participants feel that recording such interactions will diminish the ability of the officers to form a rapport with the subjects. While this perception may be true (Tankebe and Ariel 2016), in practice, this means there is a reduction in treatment intensity because officers are not complying with the protocol. The scope of reduction in intensity then depends on implementation—the more officers use their discretion, the more the study suffers from low fidelity and partial interference (Ariel et al. 2016).

A similar example is a study in which officers make a decision to start or stop recording an interaction at the very beginning of the encounter (Sykes 2015). Again, while there may be benefits for this type of activation policy (Ariel et al. 2017), it reduces the average assigned dosage because some other officers are likely to comply more fully with the policy. Since we are interested in the relative effect of the cameras compared to control conditions, the treatment dosage is diluted. Now, assume that these studies had detected significant treatment effects, meaning that they provide evidence against the null hypothesis of no difference. The policy implication is that BWCs are effective, but the magnitude of the effect is diminished: the intensity of the intervention is weaker than expected because the intervention was not delivered as intended.

Group dynamics The second issue, group dynamics, is more difficult to measure but creates the most difficulty in characterizing partial interference. A mature body of research offers insight into the ways in which individuals act when they are in social or group situations and the processes involved when they interact with each other in a group. Reference groups (Shibutani 1955), small-group psychology (Shaw 1911), Lucifer effects (Zimbardo 2007), social identity (Stott and Drury 1999), and a myriad of effects would either negatively or positively motivate participants in the group to act in various ways.

In the context of experiments and treatment heterogeneity, group dynamic effects can be manifested in the pull or push effects on participants to adhere to the experimental protocol. For example, imagine a study in which the unit of analysis is the police squad; some squads are assigned to treatment, and others are assigned to control conditions. If a particular officer in the treatment group is generally in favor of using body cameras in policing, but the rest of his/her squad members are against it, the group dynamics may push this officer into noncompliance with the protocol. On the other hand, if the majority of the squad members favor complying with the protocol but one officer has negative views about the usefulness of the cameras, it can create group pressure on him/her (e.g., peer pressure, informal retribution, or direct demands) to comply with the protocol. At the least, the analysis would have to be done on a squad-by-squad basis. At worst, the heterogeneity within squads would be greater than within individuals, requiring an even larger sample size for random assignment to "control" squad-level differences.

Spillover effects and units of analysis

Not all spillovers are created equal; as noted above, spillovers can vary considerably. Contamination effects are especially problematic when the sample size is large relative to the experimenter's resources to manage implementation, and thus high fidelity across units is more challenging (Weisburd, Petrosino and Mason 1993; Weisburd et al. 2001).[4]

[4] We note that an additional consideration is the size of the sample of officers. If a police department has 500 officers and 250 have cameras, then the opportunity for contamination is greater, but if a department has 500 officers and less than 100 officers have cameras and random assignment is stratified by a method described above, then the potential for contamination effects is much less. Similarly, if a department has 50 officers and 25 have cameras, this is a problem because police in smaller departments are more likely to run into one another during a shift. Department size matters too; see Braga et al. (2018) for further considerations.

Experimenter resources being equal, the larger the sample, the less control the researcher has of the application of the treatment across units or sites.

On the other hand, larger samples may make a study more externally valid by achieving more realistic implementation. For example, some participants will adhere to their allocated treatment, such as therapy or "treatment pathway" as prescribed by the treatment provider, while others will take part only partially. Similarly, police may visit some crime hot spots as assigned by the experimental protocol—e.g., 15-min visits, three times a day—but other hot spots will receive a lesser dosage.

In both these examples, the overall treatment effect may lead to statistically significant differences between the study arms, but the effect size may be attenuated as compared to more homogenous delivery or uptake of dosage. This was the case in several experiments testing the application of technological innovations in policing (see Ariel 2017).

How can the risks of spillover be minimized at the point of experimental design? The most salient direct way to do this is by selection of the unit of random assignment on the basis of how best to minimize spillover—even if the result is a smaller sample size than might result from choosing a unit with high risk of spillover. While a science writer for the *New York Times* (Ripley 2017) may conclude that a large sample size should be given more weight than independence of units of analysis, that conclusion directly contradicts a century of scholarship in statistics.

Based on the foregoing discussion, we identified two main choices of units for researchers wishing to conduct controlled trials on BWCs: individual officers or temporal units. We discuss each unit of analysis in the context of spillover effects.

Body-worn camera experiments with individual officers randomized

At the outset, we claim that individual-based experiments are the least appropriate design to study the effect of BWCs, because of the treatment interference threats. The problems of spillover effects—both intergroup and intragroup interference—are the most concerning, to the point that experiments with these designs may provide misleading evidence on the efficacy of cameras.

We take Yokum et al.'s (2017) experiment as a case in point: as a person-based randomized controlled field trial with a design in which the benefits, issues, and concerns about spillover effects can be discussed more thoroughly (herein, "the DC experiment"). Yokum et al. (2017) reported the findings from an RCT involving 2224 Metropolitan Police Department officers in Washington, DC. The experiment compared officers randomly assigned to wear BWCs to officers in the control condition who did not wear BWCs. The primary outcomes were documented uses of force and civilian complaints and judicial outcomes. The study found small average treatment effects on all measured outcomes, none of which was statistically significant. The authors conclude: "we should recalibrate our expectations of BWCs' ability to induce large-scale behavioral changes in policing, particularly in contexts similar to Washington, DC" (p. 4).

What captured our attention—aside from the strong generalization made based on this single study—was the detail that "our comparison groups were constructed from an individual level officer randomization scheme, which avoids several problems of

inference present in other methodologies used to date" (p. 22). In our view, assigning individuals rather than clusters or groups creates the greatest problems for inference because of the strong spillovers built into the study design. Choosing individuals as units of analysis risks challenges to independence by group dynamics, ecology of patrol (double rather than single-officer cars or foot patrols), and the attendance of major incidents by multiple officers. Thus, BWC experiments in which the unit of randomization/analysis is the individual officer are by definition characterized by strong spillover effects.[5] In Ariel et al.'s (2016) medical analogy, these circumstances are akin to having both experimental and control patients "take the pill." When everybody is exposed to the treatment, the experimental design is compromised, and by implication, it would not be possible to detect differences between groups.

SUTVA violations in the DC experiment

In the DC experiment, 1035 officers were assigned to the control group and 1189 officers to the treatment group, in which treatment officers were instructed to use cameras in police–public encounters. Two estimators of the average treatment effects were used: (A) difference-in-means with inverse probability weights to account for differential probabilities of assignment by block; and (B) regression of outcome on treatment assignment with controls for pre-treatment characteristics and inverse probability weights (p. 9). In theory, the overall design was powerful. In practice, however, the choice of officers as the unit of analysis in BWC experiments faces the greatest threat of spillover effects, to a point that field studies comparing any police practice assigned only to some and not others who work in the same communities are doomed to failure (Clarke and Weisburd 1994, p. 179).[6]

The issue is not statistical, but practical: there is no method for separating between treatment and control conditions. While officers in some police departments work alone in most citizen encounters, the largest departments have long deployed patrols in two-officer cars. The individual officer therefore cannot be the unit of analysis when the basic unit of patrol is delivered by two officers. Otherwise, there could be a scenario in which one of the officers was randomly assigned into treatment conditions (BWCs) while his/her partner was randomly assigned into control conditions (no-BWCs). When this patrol unit attends a call for service or conducts a stop and frisk, it is as if both officers are in the treatment conditions because a camera is present. Randomizing patrolling units would ameliorate this issue a little, but this merely relocates the problem because other units may attend.

Even if officers in Washington, DC, usually patrol in single-officer cars, the likelihood of interference between treatment and control conditions remains extremely high in the incidents that lead to use of force or complaints. Police culture, practice, safety, and situational factors require the attendance of more than one officer at the

[5] The use of automatic vehicle locators, CAD logs, or BWCs tracking data to indicate which officers respond to a call for service could provide a measure of contamination. This would enable researchers to identify which officers responded to a call and the assignment of those officers to treatment or control. While the practicality of such an analysis is an issue of time, a review of video of a subset of treatment and control officer arriving on scene together can determine which group dynamic plays out: is the officer with the camera less likely to turn on the camera on scene with a control officer, or vice-versa? At the same time, this approach may indicate ex post facto the degree of contamination, rather than reduce its likelihood ex ante.

[6] For example, Yokum et al (2017: 20, fn. 38) report that in 70% of calls for service control officers attended with a treatment officer present, meaning that only 30% of incidents did not have contamination problems.

encounter. Therefore, in both patrol models (single- or two-officer cars), operational needs within emergency response units often require ad hoc, triple crewing, or even larger teams, particularly when responding to complicated incidents. This suggests that officers in the control group are likely to have been contaminated by responding to calls with members of the treatment group. Because the treatment is hypothesized to affect interactions with members of the public, control officers would have altered their behaviors in response to the presence of their colleagues' BWCs (again assuming the cameras are effective). At the very least, suspects and victims might behave differently when BWCs are present, even if only some officers are wearing them.

Formally, this means that participants who function together in groups usually yield scores that are correlated (Peckham et al. 1969). When there is a mishmash between the experimental group and the control group, the probability of accepting the null hypothesis of no-treatment effects when indeed there are treatment effects, that is, of making a type II error, increases dramatically as the relationship among the individuals between the group increases (Barcikowski 1981: 269).

Partial interference in the DC experiment

Furthermore, in person-based, police BWC experiments on use of force, crossover can lead to treatment heterogeneity in both experimental arms. Control officers are sometimes exposed to the intervention when treatment officers are attending the same job, and at other times, they are not. Over time, with multiple interactions between the public and control officers that are sometimes facing crossover and sometimes are not, there is no longer a control condition, only less intensive doses of treatment. A similar concern arises solely within the treatment group because treatment officers affect the dosage level of the intervention on each other (i.e., some officers attend many incidents with multiple officers wearing cameras, whereas others might only attend some such incidents).

Suppose that during the experimental period, police officers equipped with BWCs attended 100 domestic violence calls for service. Now assume that the treatment effect of the body cameras is real and that each incident is attended by two or more officers. If the experiment is specified so that the primary officer (i.e., the first officer attending) defines whether the case is experimental or control, then by definition, variations in the treatment arm will be expected. When the primary officer is a treatment officer (X) and the second attending officer is a control officer (Z), then the case is designated as experimental (X), but overall the treatment effect is (X + z); when the second attending officer is a treatment officer, then the treatment effect on the primary officer is X + x; and when a third officer is attending the scene, the exerted treatment effect on the primary officer is X + x + x or X + x + y, depending on the allocation of the third officer—and so on. Thus, multiple officers lead to a convoluted treatment heterogeneity that becomes difficult to describe (e.g., the interaction could be multiplicative rather than additive). Long causal chains with multiple responders, similar to a network of interconnected nodes, exert effect on each other. When the partial interference creates such a degree of statistical noise that the treatment efficacy cannot be quantified, it creates issues for assessing the magnitude of the treatment effect.

Selection bias and chance in rare events

One related issue is the Pareto curve concentration of rare events in certain situations or with certain officers. The number of contacts per 10,000 encounters that lead to complaints against the officer, for example, or that result in the use of force, is remarkably small (see Terrill and McCluskey 2002). These infrequencies mean that in experiments in most departments, the majority of complaint-conducive or force-response contacts can fall into one of the treatment arms by purposeful selection bias—or because of chance in how the random allocation has worked. Because officers may be able to anticipate problematic calls (e.g., to specific locations, during specific hours of the day, or when dealing with particular types of known offenders), a subset of officers may simply avoid contact in such high-risk situations. Ariel et al. (2017) construe this type of camera-induced inaction as a form of "de-policing." (However, Headley et al. 2017 find no supportive evidence for abstaining from community contacts in the Hallandale Beach Police Department in Florida.)

Analytical considerations for individuals as units of random assignment

A growing body of literature attempts to deal directly with the analysis of SUTVA-violating trials. However, these solutions are partial and often deal with groups or clusters as the units of analysis, rather than individual participants. One reason using the individual officer as the unit of analysis is problematic is that it ignores group dynamics and organizational factors that are very difficult to control for in any statistical model. Underlying forces and cultural codes of behavior can characterize entire forces or shifts, and most of these factors are not recorded and therefore cannot be included in the statistical model. These may include the character of the sergeant managing the shift, the degree of officers' cynicism, comradery, and codes of silence. A host of institutional undercurrents that are recognized in the literature (Sherman 1980), but cannot be factored into a statistical protocol without detailed information about the officers themselves, may affect the "independence" of individuals from factors affecting the deployment of officers with cameras. Furthermore, adding statistical controls may exacerbate problems if they are uncorrelated with outcomes or open back-door pathways that corrupt treatment allocation (Morgan and Winship 2007).

Body-worn camera experiments with temporal units randomized

As an alternative to the individual-based RCT on BWCs, experimentalists can choose to randomize temporal units, as with the original Rialto experiment (Ariel et al. 2015) and as designated in its first experimental protocol (Ariel and Farrar 2012) and replicated more recently by Headley et al. (2017). In all these tests, officers were exposed to both treatment and control conditions—this is similar to a crossover trial with more than one switch between conditions, for each officer. It is a repeated measurement design, such that each experimental unit (e.g., an officer) receives different treatments during the different time periods, that is, the officers "crossover" from one treatment condition to another condition, during the course of the trial.

A major consideration in favor of a crossover design is that it could yield a more efficient comparison of treatments than a parallel design. For example, fewer units are

required to attain the same level of statistical power or precision. In practice, this means that every officer is serving as their own matched control, which leads to a fundamental benefit: a crossover design is strongly balanced with respect to the carryover effects, when each treatment precedes every other treatment, including itself, the same number of times.

By making police shifts (e.g., a 08:00–17:00 shift) the unit of analysis, the sample size available can be increased significantly, allowing much smaller effect sizes to be detected but with relatively few front-line officers. When there are more shifts or other temporal units (e.g., days of the week) than police officers, especially in midsize departments, substitutes ought to be considered to satisfy the sample size problem (unless a Bayesian approach is possible; see Lenth 2001). One thousand shifts is sufficient to detect small effects ($d = 0.178$) with an alpha of 0.05, power of 80% for a two-tailed statistical test (with no covariates and thus no variance explained by covariates), but those 1000 shifts could be generated by as few as 60 officers, as in Rialto. In contrast, with a study of approximately 128 officers and no covariates to increase the statistical power of the test, a study is unlikely to detect effects below $d = 0.499$, and the practice in some studies had been to relax some of these statistical assumptions of the power test (e.g., Jennings et al. 2015, p. 482).

Randomly assigning shifts as the unit of analysis is not a perfect solution, given the potential spillover effect (Ariel et al. 2015). The same officers are randomly assigned to use the cameras and also randomly assigned not to use the cameras. However, it represents a least worst option (what is sometimes called the maximin rule; see Rawls 2009, p. 72). The issue with contamination when using shifts is that the same officers experience both treatment and control shifts, so there is the likelihood that behavioral modifications due to treatment conditions can be carried over into control conditions. If BWCs affect behavior, then a learning mechanism may be at play in which officers adapt their overall behavior (and possibly attitudes), and this broader change affects control conditions as well (Ariel 2016a, b; Ariel et al. 2015, p. 528). However, we believe the story is more nuanced than to discount this unit of randomization.

SUTVA in the context of shift-based experiments

Ariel et al. (2015, p. 623) were the first to note that the fact that officers participated on multiple occasions in both treatment and control conditions creates "interference," as it does in many other crossover designs in which each unit serves sequentially as treatment and control (Brown Jr 1980). However, as the authors note, the unit of analysis is the shift, not the officer. The set of conditions encountered in each shift cannot be repeated because time moves in only one direction. The manipulation was whether the shift involves all police with cameras or no police with cameras.[7] Outcomes (use of force, complaints, etc.) are essentially driven by how officers act and

[7] In fact, the intervention should be said to consist of both the camera and an activation notification, such as a verbal warning, a visual cue, or any sort of announcement that notifies the suspect that s/he is being videotaped. The verbal warning is also in place to initiate the cognitive process of public social-awareness within the officer using the body-worn camera. Police departments should be mindful of this aspect of the intervention because most suspects, witnesses, and victims are not in a position to identify the body-worn camera among the wide range of gadgets modern police officers wear on a daily basis.

how citizens perceive those actions during each shift. Likewise, because the whole 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.

Despite the potential SUTVA violation in any crossover design, it is still the case that when a treatment is present the individuals are in a different context from when it is absent—regardless of their prior experience in both conditions. Officers in a shift-based experiment, while serving during control shifts, do not wear BWCs. Officers are certainly aware that their actions and conduct are under surveillance in both treatment and control shifts. However, awareness of potential surveillance is not equal in credible deterrent threat. In control shifts, detection of rule breaking (by either citizen or officer) is less likely because the cameras are not present. In the treatment condition, every recorded interaction can be viewed and audited. This may lead to an officer's reprimand or a citizen's complaint being challenged. An unrecorded interaction does not necessarily lead to similar costs, since a recorded incident of excessive use of force can very likely lead to criminal prosecution of the officer. An unrecorded incident of excessive use of force, in contrast, can more easily be left to subjective interpretations. In deterrence theory terms, the perceived likelihood of apprehension is more substantially elevated in treatment conditions than control conditions. While under both experimental arms, the behavior may have been modified as a result of the spillover, the extent of the behavioral modification under control conditions cannot be assumed to be the same as that which has taken place under treatment conditions—otherwise we would not observe significant differences between treatment and control conditions across multiple outcomes using this research design (e.g., Ariel et al. 2015; Ariel et al. 2016a; Ariel et al. 2016b).[8]

To summarize, a shift-based design can create, in theory, both negative and positive spillover effects. The negative effects would be to contaminate the control group with treatment. The positive effect would be to reinforce the treated officers with the effects of treatment on each other's behaviors.

Being able to define units, treatments, and outcomes in this way means, we can be more specific about when SUTVA violations might be occurring. More importantly, spillover effects often result from experiments, which indeed may be the intention (Angelucci and Maro 2016). The spillover means that officers in control conditions were affected by their counterpart treatment conditions and altered their behavior enough, regardless of treatment condition.

Therefore, spillovers that take place within the experimental arm are not necessarily undesirable. Glennerster and Takavarasha (2013: 348) described the role of positive spillovers in experiments: when there are positive spillovers, those in the treatment group benefit from the fact that they are "surrounded by those who take up the program and the benefits they received are assumed […] to be experienced by those who take up the program." Braga et al. (2013) and Braga and Weisburd (2014, pp. 583–586) show how positive spillover effects contributed to meaningful reductions in gang violence. A systematic review of positive spillover effects in medical impact evaluations suggested

[8] In experimental sites in which the cameras had no effect, Ariel et al. (2016b) have found strong evidence of noncompliance, which they described as implementation errors because the cameras were not used as intended.

that such effects are not only desirable, but also carry cost-effective externalities for public health programs (Benjamin-Chung et al. 2015).

Individual vs. temporal units: statistical considerations

Let us return to the notion of "analyze as you randomize" (see Senn 2004; Demir et al. 2018 ; Ariel et al. 2018; Maskaly et al. 2017). Analyzing at the officer-level following shift-level randomization (i.e., ignoring the fact that officers are clustered by shift) would undermine the experimental design, becoming the exercise in self-deception against which Cornfield (1976) warns. Analyzing officers after randomizing shifts may also require the scholar to measure different variables at the outset as baseline covariates and, plausibly, to control for them, including interactions. With all this in mind, we discuss the analytical considerations of inferring causation between the shift—and the shift only—on the outcomes of interest.

The critical issue in terms of spillover effects is that a shift-based design explicitly creates risks to type II error rather than type I error. In practice, using the shift as the random assignment unit, with the potential of cross-unit contamination, means that it becomes more difficult to reject the null hypothesis of differences between treatment and control conditions. Because both arms of the trial are exposed to some level of the manipulation (at least as it is applied to the officers), it becomes more challenging to detect statistically significant differences. Hence, if anything, a statistically significant difference between the experimental and control arms under these conditions implies that the true treatment effect is more pronounced. Put another way, the exposure of officers to both treatment and control conditions is likely to affect the estimation of treatment effects asymmetrically. Officers in control shifts are likely to change their behavior because of exposure to cameras during their own treatment shifts.

The "shift effect spillover hypothesis" is that during *control* shifts, officers would change their behavior to be more like that during treatment shifts. The spillover would therefore act to shrink the gap between treatment and control conditions by making control shifts more like treatment shifts. If true, this means that the estimated effect sizes for high compliance experiments would represent lower-bound estimates of effect sizes—or underestimation of the treatment effect. In other words, this so-called flaw makes the job of this test in showing a significant outcome harder, not easier, resulting in a more conservative test rather than a less stringent type I error rate.[9] As more robustly concluded by De La and Rubenson (2010, p. 195), in such circumstances, "the intervention's indirect negative effect on non-recipients would produce a diluted effect of the program," but if the findings are nevertheless in favor of the hypothesized direction, the issue is not of reliability, but of magnitude. In other words, in a shift-based RCT, there is a threat of spillover comparable to any crossover design, but it is not as large a threat as giving patients in the control group the active pills rather than placebos.

Moreover, the degree of contamination with shift randomization is more limited than when using officers as the unit of randomization/analysis. An implicit assumption of using officers as the unit of analysis in a simple statistical model is that the effect of the

[9] A statistical type I error indicates that the null hypothesis is rejected when it ought not to be not (false positive), while type II error implies that the null hypothesis is not rejected when it ought to be rejected (false negative).

suspect variation and the effect of officer-suspect interaction are negligible (Whiting-O'Keefe et al. 1984). Nevertheless, the error rates are not and should not be assumed to be distributed equally between units or across study groups. From a theoretical perspective that would then affect the computation of the predictors, BWCs may have at least as much of an effect on citizens as they do on officers, particularly if citizens are verbally warned that cameras are being used (Ariel et al. 2016c). Because the officer here is not the unit of randomization, the analytical procedure ought to be centered on the unit of randomization and generalized to the universe of police shifts, not officers. This argument obviously does not apply if officers are the unit of randomization.

Conclusions and recommendations

Research designs that fail to account for spillovers produce biased estimates of treatment effects, and studies that produce biased treatment effects can lead to misconstrued policy recommendations. This issue is present in all experiments, regardless of sample size. Having a large study that suffers from spillover is less powerful than a small study that adequately handles spillover effects. Consequently, choice of units in experimental criminology is critical. Unlike observational studies, the trial cannot go back and change the unit once it has been assigned. Deviation from these basic rules means the trial is no longer an RCT, but rather a controlled quasi-experimental design.

Contamination is not only plausible when the units are directly exposed to the manipulation, but also indirectly or vicariously. If the treatment-providers—police, probation officers, judges, or therapists—are aware of which participant they are treating, they may behave differently, set different expectations, or lead to self-fulfilling prophecies that may indivertibly bleed from one study arm to the next. For example, when police officers in the Minneapolis Domestic Violence Experiment were able to accurately predict the next random assignment sequence, they treated the case differently (Gartin 1995). This can happen in other research designs, depending on officer preferences about the study outcomes (e.g., disparities in arrival time, the application of procedural justice, expectations from the party with which he or she is engaged). Patients interacting with one another in the waiting room before entering singly into a clinical trial can contaminate each other; police officers participating in an experiment on hot spots policing can purposely patrol control sites even though they were instructed otherwise (as they did in the first systematic patrol experiment; see Kelling et al. 1974), and prisoners randomly assigned to a particular rehabilitation program can engage with control prisoners, all in a way in which the treatment spills over to other individuals. The effect can also take place within subject, when the participant affects him- or herself over time. It can also occur within the treatment group only, when some participants are exposed to different levels of the treatment, or when they affect each other given varying attitudes, expectations, or degrees of implementation success. Hence, researchers should expect some degree of spillover when conducting real-world tests.

However, experimenters should equally try to minimize these contaminations as much as possible, both between and within the study groups (partial interference). We

recommend that future scholars avoid using officer-level randomization because it creates spillover effects that lead to design failures unless the scholars are confident that officers are not interacting with one another, and not just overall, but during encounters that are force-conducive or prone to generate complaints. Because officers in most large departments patrol in pairs or larger formations (not least due to officers' safety), by definition, the unit of analysis is not the individual officer, but the patrolling unit. SUTVA violations cannot be characterized at all in these individual-based experiments. Ultimately, it is no surprise that a study such as the DC experiment failed to reject the null hypothesis: its design was not suitable to the question it was trying to answer.

As Morgan and Winship point out, for many applications, SUTVA is a restrictive assumption (2007, pp. 37–39). Therefore, studies ultimately leading to statistically significant differences between no-camera and camera conditions in the test can be interpreted as producing favorable outcomes. Yet when the study produces non-significant results, the outcomes are more challenging to interpret. Are the findings a result of a true no-effect, or was the design incapable of producing reliable estimates? Contrary to our global experience with BWCs, the findings are not mixed; they are, as far as we can tell, consistent with the hypothesized civilizing effect of BWCs on police–public contacts. Thus, the DC study is the exception rather than the norm, which leads us to conclude that methodological challenges and in particular the contaminated spillover effects of using a person-based randomization sequence reduced the ability of the experiment to detect true effects.

Possible design solutions for future experiments at risk of interference

More broadly, we note that there are recent and helpful solutions to the interference concern. One solution to the partial interference scenario is to take advantage of the treatment propagation by assigning less than half of the pool to treatment from the perspective of statistical efficiency (Bowers et al. 2018). This seems logical because when treatment spreads rapidly across a network, then "comparisons of outcomes between treated and control units will become very small or even vanish as the control units to which the treatment spread will act just like treated units" (p. 197).[10]

Network analysis techniques also provide a useful solution to handling treatment propagation in clusters, and these are becoming more common in observational data (Lyons 2011; Shalizi and Thomas 2011) and randomized experimental designs (Aral and Walker 2011; Aronow and Samii 2013; Bapna and Umyarov 2015; Bond et al. 2012; Eckles et al. 2017; Ichino and Schündeln 2012; Ostrovsky and Schwarz 2011; Rosenbaum 2007; Toulis and Kao 2013). In fact, treatment propagation is now considered in research as both a target of inference and as a nuisance. Network analysis can show graphical scenarios where the potential outcomes of a unit are a function of the treatment assigned to a unit and of the treatment assigned to other units that are related to a unit through the network (Basse and Airoldi 2015a, b). This interest had led

[10] As Bowers et al. (2018) explain, this process entails the following procedure: "A node could be treated directly by an experimenter, isolated from treatment (i.e., several hops away from any treated nodes) or exposed to the treatment at one degree of separation by virtue of the network relationship—without control by the experimenter." For a more elaborate discussion, see Aronow and Samii (2017), Bowers et al. (2013), and Toulis and Kao (2013).

to recent methodological work on statistical inferences about peer effects or total average effects, when the topology of the network can be explained (Aronow and Samii 2017; Bowers et al. 2013; Eckles et al. 2017; Toulis and Kao 2013).

In terms of random assignment, statisticians offer a partial although convincing solution to the interference issue: model-assisted restricted randomization strategies that take into account these interference effects (see Yates 1948, but more recently see Ariel and Farrington 2010). The premise of these techniques is that some assignments are considered problematic (e.g., when interference happens or when covariates are potentially unbalanced between the treatment arms) and can be excluded. In networks, the challenge is to identify which features must be balanced, which makes it challenging to know how to restrict the randomization. Basse and Airoldi (2017) and Basse and Feller (2017) suggest a novice approach called "two-stage experiments" to identify subsets of units that can indeed be construed as independent (free of spillover), which is a subset of constrained randomization techniques. This approach utilizes a subset of units that might be assumed or known to be independent of one another and allocates treatment conditions to these units. The statistical literature should be consulted (Basse and Airoldi 2017; Basse and Feller 2017; and others).

Finally, there is recent work on dyadic relationships that should be considered in future experiments when interdependence is unavoidable. This so-called "actor-partner interdependence model" (APIM; Kashy and Kenny 2000) can be used to analyze dyadic data. It integrates a conceptual view of interdependence with the relevant statistical techniques for measuring and testing it (Cook and Kenny 2005). This approach enables experimentalists to simultaneously examine the effect of the treatment effect on the actor and then on the partner; interestingly, this "partner effect" illustrates the interdependent nature of relationships. APIM can be used for dyads only or for groups, but the latter can become mathematically complex. For further reading on this approach, see Ledermann and Kenny (2015), Cook and Kenny (2005), and Garcia et al. (2015).

A final word about the link between interference and compliance with the experimental protocol

Throughout this note, we suggested that the commitment (or lack thereof) to using BWCs appropriately is vital to understanding whether spillover occurs. Compliance with the protocol is therefore a key feature. Incidents that involve officers with (treatment) and without (control) BWCs result in contamination to the control group, assuming prima facie that treatment officers indeed turn their camera on. Following the release of Yokum et al.'s (2017) study, the Washington DC Office of Police Complaints (OPC) found that officers failed to comply with department guidelines for BWC use in 34% of cases the OPC investigated. To be sure, these are just situations that the OPC investigated and not an overall assessment of the cameras. Similarly, in Phoenix, AZ, evaluation of BWCs implementation found that 1 month after deployment, 42.2% of all incidents that should have been recorded with a BWC were not, and compliance declined over time, to 13.2% (Katz et al. 2015). Thus, if treatment officers are not turning their cameras on, this reduces not only the intervention effect size, but also the concern for spillover. We believe this concern is exacerbated with person-based experiments.

A second issue with officers not turning their camera on and spillover effects is group dynamics, which could facilitate a change in behavior of non-compliant officers through partial interference effect. However, this again assumes that officers are turning their cameras on. If compliance to the intervention is not occurring, then group dynamics may lead to negative behavior (on adverse group dynamics, see Xia et al. 2009) Hedberg et al. (2017) support this contention: their evaluation of BWCs showed that compliance worsened due to a lack of oversight and thus there was no deterrent effect against noncompliance, which may explain their results.

Summary

We conclude by reiterating that shift randomization allows researchers to maximize sample sizes, be in a better position to characterize SUTVA violations (see Sampson 2010) and minimize problems arising from spillover effects. After all, the experience with the most shift-based trials on BWCs has led to significant results, and those that did not produce discernible effects were characterized by poor implementation (Ariel et al. 2016). One must also consider alternative designs. Practitioners and policy-makers should be encouraged by the consistency of most of the results from the range of studies that appear to support the implementation of BWCs. A series of properly designed cluster-randomized trials will assist in providing an overall conclusion about the utility of the cameras for policing. Finally, we encourage researchers, practitioners, and policy-makers to look beyond the results from single studies, regardless of size, and think about the whole evidence puzzle.

References

Anderson, R. M., & May, R. M. (1985). Vaccination and herd immunity to infectious diseases. *Nature, 318*(6044), 323.

Angelucci, M., and Di Maro, V. (2016). Programme evaluation and spillover effects. Journal of Development Effectiveness, 8(1) 22–43.

Aral, S., & Walker, D. (2011). Creating social contagion through viral product design: a randomized trial of peer influence in networks. *Management Science, 57*(9), 1623–1639.

Ariel, B. (2016a). The effect of police body-worn videos on use of force, complaints and arrests in large police departments. *Journal of Criminal Law and Criminology, 106*(1), 101.

Ariel, B. (2016b). Increasing cooperation with the police using body worn cameras. *Police Quarterly, 1098611116653723*.

Ariel, B. (2017). Technology in policing. In D. Weisburd & A. A. Braga (Eds.), *Innovations in policing: contrasting perspectives* (2nd ed.). Cambridge: Cambridge University Press.

Ariel, B., & Farrington, D. (2010). Randomised block designs. In D. Weisburd & A. Piquero (Eds.), *Handbook of quantitative criminology* (pp. 437–457). New York: Springer Verlag.

Ariel, B., & Farrar, W. A. (2012). The Rialto police department wearable cameras experiment: Crimport (Experimental protocol available at https://www.crim.cam.ac.uk/global/docs/rialto.pdf.

Ariel, B., Farrar, W. A., & Sutherland, A. (2015). The effect of police body-worn cameras on use of force and citizens' complaints against the police: a randomized controlled trial. *Journal of Quantitative Criminology, 31*(3), 509–535.

Ariel, B., Weinborn, C., & Sherman, L. W. (2016). "Soft" policing at hot spots—do police community support officers work? A randomized controlled trial. *Journal of Experimental Criminology, 12*(3), 277–317.

Ariel, B., Sutherland, A., Henstock, D., Young, J., Drover, P., Sykes, J., . . . Henderson, R. (2016b). Report: increases in police use of force in the presence of body-worn cameras are driven by officer discretion: a protocol-based subgroup analysis of ten randomized experiments. *Journal of Experimental Criminology*, 1–11.

Ariel, B., Sutherland, A., Henstock, D., Young, J., Drover, P., Sykes, J., et al. (2016c). Wearing body cameras increases assaults against officers and does not reduce police use of force: results from a global multi-site experiment. *European Journal of Criminology, 1477370816643734*.

Ariel, B., Sutherland, A., Henstock, D., Young, J., Drover, P., Sykes, J., et al. (2017). "Contagious accountability" a global multisite randomized controlled trial on the effect of police body-worn cameras on citizens' complaints against the police. *Criminal Justice and Behavior, 44*(2), 293–316.

Ariel, B., Sutherland, A., Henstock, D., Young, J., Drover, P., Sykes, J., et al. (2018). Paradoxical effects of self-awareness of being observed: testing the effect of police body-worn cameras on assaults and aggression against officers. *Journal of Experimental Criminology, 14*(1), 19–47.

Armitage, P. (1991). Should we cross off the crossover? *British Journal of Clinical Pharmacology, 32*(1), 1–2.

Armitage, P. (2003). Fisher, Bradford Hill, and randomization. *International Journal of Epidemiology, 32*(6), 925–928.

Aronow, P. M., & Samii, C. (2013). Estimating Average Causal Effects Under General Interference, with Application to a Social Network Experiment. arXiv preprint arXiv:1305.6156.

Aronow, P. M., & Samii, C. (2017). Estimating average causal effects under general interference, with application to a social network experiment. *Ann. Appl. Stat., 11*(4), 1912–1947.

Baird, S., Bohren, J. A., McIntosh, C. and Ozler, B. (2016). *Optimal design of experiments in the presence of interference* (December 1 2016). PIER Working Paper No. 16-025. Available at SSRN: https://ssrn.com/abstract=2900967.

Bapna, R., & Umyarov, A. (2015). Do your online friends make you pay? A randomized field experiment on peer influence in online social networks. *Management Science, 61*(8), 1902–1920.

Barcikowski, R. S. (1981). Statistical power with group mean as the unit of analysis. *Journal of Educational Statistics, 6*(3), 267–285.

Basse, G. W. and Airoldi, E. M. (2015a). *Optimal design of experiments in the presence of network-correlated outcomes*. arXiv preprint arXiv:1507.00803.

Basse, G. W., and Airoldi, E. M. (2015b). *Model-assisted design of experiments in the presence of network correlated outcomes*. *arXiv preprint arXiv:1507.00803*.

Basse, G., & Airoldi, E. (2017). Limitations of design-based causal inference and A/B testing under arbitrary and network interference. arXiv preprint arXiv:1705.05752.

Basse, G., & Feller, A. (2017). Analysing two-stage experiments in the presence of interference. *Journal of the American Statistical Association, 113*(531). https://doi.org/10.1080/01621459.2017.1323641.

Benjamin-Chung, J., Abedin, J., Berger, D., Clark, A., Falcao, L., Jimenez, V., ... and Luby, S. P. (2015). The identification and measurement of health-related spillovers in impact evaluations: a systematic review. *Grantee Final Review. New Delhi: 3ie*.

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature, 489*(7415), 295–298.

Boruch, R. F. (1997). Randomized experiments for planning and evaluation: A practical guide (Vol. 44). Sage Publications.

Bowers, J., Fredrickson, M. M., & Panagopoulos, C. (2013). Reasoning about interference between units: a general framework. *Political Analysis, 21*(1), 97–124.

Bowers, J., Desmarais, B. A., Frederickson, M., Ichino, N., Lee, H. W., & Wang, S. (2018). Models, methods and network topology: experimental design for the study of interference. *arXiv preprint arXiv: 1601.00992*. *Social Networks, 54*, 196–208.

Braga, A. A., & Weisburd, D. L. (2014). Must we settle for less rigorous evaluations in large area-based crime prevention programs? Lessons from a Campbell review of focused deterrence. *Journal of Experimental Criminology, 10*(4), 573–597.

Braga, A. A., Apel, R., & Welsh, B. C. (2013). The spillover effects of focused deterrence on gang violence. *Evaluation Review, 37*(3–4), 314–342.

Braga, A. A., Sousa, W. H., Coldren Jr., J. R., & Rodriguez, D. (2018). The effects of body-worn cameras on police activity and police-citizen encounters: a randomized controlled trial. *Journal of Criminal Law and Criminology, 108*(3), 511–538.

Brown Jr., B. W. (1980). The crossover experiment for clinical trials. *Biometrics*, 69–79.

Campbell, D. T., & Stanley, J. C. (1966). Experimental and quasiexperimental designs for research. Chicago: Rand McNally & Company.

Clarke, R. V., & Weisburd, D. (1994). Diffusion of crime control benefits: observations on the reverse of displacement. *Crime Prevention Studies, 2*, 165–184.

Cook, W. L., & Kenny, D. A. (2005). The actor–partner interdependence model: a model of bidirectional effects in developmental studies. *International Journal of Behavioral Development, 29*(2), 101–109.

Cornfield, J. (1976). Recent methodological contributions to clinical trials. *American Journal of Epidemiology, 104*(4), 408–421.

Cornfield, J. (1978). Symposium on CHD prevention trials: design issues in testing life style intervention: randomization by group: a formal analysis. *American Journal of Epidemiology, 108*(2), 100–102.

Cox, D. R. (1958). *The planning of experiments*. New York: Wiley.

Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I., & Petticrew, M. (2008). Developing and evaluating complex interventions: the new Medical Research Council guidance. *BMJ, 337*, a1655.

De La, O. A., & Rubenson, D. (2010). Strategies for dealing with the problem of non-overlapping units of assignment and outcome measurement in field experiments. *The Annals of the American Academy of Political Science, 628*(1), 189–199.

Demir, M., Apel, R., Braga, A., Brunson, R., & Ariel, B. (2018). Body worn cameras, procedural justice, and police legitimacy: a controlled experimental evaluation of traffic stops. *Justice Quarterly*.

Dong, N., & Maynard, R. (2013). PowerUP!: a tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness, 6*(1), 24–67.

Eckles, D., Karrer, B., & Ugander, J. (2017). Design and analysis of experiments in networks: reducing bias from interference. *Journal of Causal Inference, 5*(1). https://doi.org/10.1515/jci-2015-0021.

Garcia, R. L., Kenny, D. A., & Ledermann, T. (2015). Moderation in the actor–partner interdependence model. *Personal Relationships, 22*(1), 8–29.

Gartin, P. R. (1995). Dealing with design failures in randomized field experiments: analytic issues regarding the evaluation of treatment effects. *Journal of Research in Crime and Delinquency, 32*(4), 425–445.

Gilbert, D., King, G., Pettigrew, S., Wilson, T. (2016). *More on "estimating the reproducibility of psychological science"* Available at projects.iq.harvard.edu/files/psychology-replications/files/gkpw_post_publication_response.pdf.

Glennerster, R., & Takavarasha, K. (2013). *Running randomized evaluations: a practical guide.* Princeton: University Press.

Gottfredson, D. C., Cook, T. D., Gardner, F. E., Gorman-Smith, D., Howe, G. W., Sandler, I. N., & Zafft, K. M. (2015). Standards of evidence for efficacy, effectiveness, and scale-up research in prevention science: next generation. *Prevention Science, 16*(7), 893–926.

Greene, J. R. (2014). The upside and downside of the "police science" epistemic community. *Policing: A Journal of Policy and Practice, 8*(4), 379–392.

Hartman, E., Grieve, R., Ramsahai, R., & Sekhon, J. S. (2015). From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 178*(3), 757–778.

Headley, A. M., Guerette, R. T., & Shariati, A. (2017). A field experiment of the impact of body-worn cameras (BWCs) on police officer behaviour and perceptions. *Journal of Criminal Justice, 53*, 102–109.

Hedberg, E. C., Katz, C. M., & Choate, D. E. (2017). Body-worn cameras and citizen interactions with police officers: Estimating plausible effects given varying compliance levels. *Justice Quarterly, 34*(4), 627–651.

Hudgens, M. G., & Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association, 103*(482), 832–842.

Ichino, N., & Schündeln, M. (2012). Deterring or displacing electoral irregularities? Spillover effects of observers in a randomized field experiment in Ghana. *The Journal of Politics, 74*(1), 292–307.

Jennings, W. G., Lynch, M. D., & Fridell, L. A. (2015). Evaluating the impact of police officer body-worn cameras (BWCs) on response-to-resistance and serious external complaints: evidence from the Orlando police department (OPD) experience utilizing a randomized controlled experiment. *Journal of Criminal Justice, 43*(6), 480–486.

Kashy, D. A., & Kenny, D. A. (2000). The analysis of data from dyads and groups. *Handbook of research methods in social and personality psychology, 38*, 451–477.
Katz, C. M., Kurtenbach, M., Choate, D. E., & White, M. D. (2015). *Phoenix, Arizona, smart policing initiative: evaluating the impact of police officer body-worn cameras*. Washington, DC: Bureau of Justice Assistance, US Department of Justice.
Kelling , G. L., Pate, T., Dieckman, D., and Brown, C. E. (1974). *The Kansas City preventive patrol experiment*. Washington, DC: Police Foundation.
Kruskal, W. (1988). Miracles and statistics: the causal assumption of independence. *Journal of the American Statistical Association, 83*(404), 929–940.
Ledermann, T., & Kenny, D. A. (2015). A toolbox with programs to restructure and describe dyadic data. *Journal of Social and Personal Relationships, 32*(8), 997–1011.
Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician, 55*(3), 187–193.
Lyons, R. (2011). The spread of evidence-poor medicine via flawed social-network analysis. *Statistics, Politics, and Policy, 2*(1). https://doi.org/10.2202/2151-7509.1024.
Maskaly, J., Donner, C., Jennings, W. G., Ariel, B., & Sutherland, A. (2017). The effects of body-worn cameras (BWCs) on police and citizen outcomes: a state-of-the-art review. *Policing: An International Journal of Police Strategies & Management, 40*(4), 672–688.
Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal analysis: methods and principles for social research*. Cambridge: Harvard University Press.
Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference*. Cambridge University Press.
Nagin, D. S., & Sampson, R. J. (2018). The real gold standard: measuring counterfactual worlds that matter most to social science and policy. https://doi.org/10.1146/annurev-criminol-011518-024838.
Ostrovsky, M., and Schwarz, M. (2011, June). Reserve prices in internet advertising auctions: a field experiment. In *Proceedings of the 12th ACM conference on electronic commerce* (pp. 59-60). ACM.
Page, E. S. (1978). Process control. In W. H. Kruskal & J. M. Tanur (Eds.), *International encyclopedia of statistics* (pp. 809–812). New York: Free Press.
Peckham, P. D., Glass, G. V., & Hopkins, K. D. (1969). The experimental unit in statistical analysis. *The Journal of Special Education, 3*(4), 337–349.
Peto, R., Pike, M., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., et al. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *British Journal of Cancer, 34*(6), 585.
Rawls, J. (2009). *A theory of justice* (Revised ed.). Cambridge, MA: Harvard University Press.
Ripley, A. (2017). The upshot: a big test of police body cameras defies expectations. New York Times, October 20 [last accessed 08 Oct 2018] from https://www.nytimes.com/2017/10/20/upshot/a-big-test-of-police-body-camerasdefies-expectations.html.
Rosenbaum, P. R. (2007). Interference between units in randomized experiments. *Journal of the American Statistical Association, 102*(477), 191–200.
Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher Randomization Test Comment. *Journal of the American Statistical Association, 75*(371), 591–563.
Sampson, R. J. (2010). Gold standard myths: observations on the experimental turn in quantitative criminology. *Journal of Quantitative Criminology, 26*(4), 489–500.
Senn, S. (2004). Controversies concerning randomization and additivity in clinical trials. *Statistics in Medicine, 23*(24), 3729–3753.
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton, Mifflin and Company.
Shalizi, C. R., & Thomas, A. C. (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods and Research, 40*(2), 211–239.
Shaw, M. E. (1911). *Group dynamics: the psychology of small group behaviour*. New York: McGraw-Hill.
Sherman, L. W. (1980). Causes of police behavior: the current state of quantitative research. *Journal of Research in Crime and Delinquency, 17*(1), 69–100.
Sherman, L. W. (1993). Defiance, deterrence, and irrelevance: a theory of the criminal sanction. *Journal of Research in Crime and Delinquency, 30*(4), 445–473.
Shibutani, T. (1955). Reference groups as perspectives. *American Journal of Sociology, 60*(6), 562–569.
Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference. *Journal of the American Statistical Association, 101*(476), 1398–1407.
Stott, C., & Drury, J. (1999). The inter-group dynamics of empowerment: a social identity model. In J. Hearn & P. Bagguley (Eds.), *Transforming politics* (pp. 32–45). UK: Palgrave Macmillan.
Sykes, J. (2015). Overcoming the threat to fidelity in RCTs. [Masters Dissertation, University of Cambridge].

Tankebe, J., Ariel, B. (2016). *Cynicism towards change: The case of body-worn cameras among police officers*. Hebrew University of Jerusalem Legal Research Paper No. 16–42. Available at SSRN: https://ssrn.com/abstract=2850743.

Terrill, W., & McCluskey, J. (2002). Citizen complaints and problem officers: examining officer behaviour. *Journal of Criminal Justice, 30*(2), 143–155.

Toulis, P., and Kao, E. (2013, February). Estimation of causal peer influence effects. In *International conference on machine learning* (pp. 1489–1497).

Weisburd, D., Petrosino, A., & Mason, G. (1993). Design sensitivity in criminal justice experiments. *Crime and justice, 17*, 337–379.

Weisburd, D., Lum, C. M., & Petrosino, A. (2001). Does research design affect study outcomes in criminal justice? *The Annals of the American Academy of Political and Social Science, 578*(1), 50–70.

Whiting-O'Keefe, Q. E., Henke, C., & Simborg, D. W. (1984). Choosing the correct unit of analysis in medical care experiments. *Medical care*, 1101–1114.

Winchester, S. (2018). *The perfectionists: how precision engineers created the modern world*. NY: HarperCollins.

Xia, L., Yuan, Y. C., & Gay, G. (2009). Exploring negative group dynamics: adversarial network, personality, and performance in project groups. *Management Communication Quarterly, 23*(1), 32–62.

Yates, F. (1948). Systematic sampling. *Philosophical Transactions of the Royal Society of London Series A, 241*(834), 345–377.

Yokum, D., Ravishankar, A., and Coppock, A. (2017). *Evaluating the effects of police body-worn cameras: a randomized controlled trial Working Paper (October 20 2017)*. Available at http://bwc.thelab.dc.gov/TheLabDC_MPD_BWC_Working_Paper_10.20.17.pdf. Last accessed 04 November 2017.

Zimbardo, P. G. (2007). *Lucifer effect: understanding how good people turn evil*. Hoboken: Blackwell Publishing Ltd..

Dr Barak Ariel is a Lecturer in Experimental Criminology and a Jerry Lee Fellow of Experimental Criminology at the Institute of Criminology of the University of Cambridge. He is also an Associate Professor in Criminology at the Hebrew University of Jerusalem, Israel.

Dr Alex Sutherland is a Senior Research Leader at RAND Europe, where he leads RAND's experimental research. His primary research areas at RAND are criminal justice and education.

Lawrence W. Sherman is the founding president of the Academy of Experimental Criminology and Chair of the Division of Experimental Criminology of the American Society of Criminology. Director of the Jerry Lee Centre for Experimental Criminology at Cambridge University, he is also a Distinguished University Professor at the University of Maryland.

Affiliations

Barak Ariel [1,2] · Alex Sutherland [3,4] · Lawrence W. Sherman [5,6]

1 Institute of Criminology, University of Cambridge, Sidgwick Avenue, Cambridge CB3 9DA, UK

2 Institute of Criminology, Faculty of Law, Hebrew University, Mount Scopus, 91905 Jerusalem, Israel

3 Communities, Safety & Justice RAND Europe, Westbrook Centre, Milton Road, Cambridge CB4 1YG, UK

4 Research Associate & Member of Violence Research Centre Institute of Criminology, University of Cambridge, Cambridge, UK

5 Wolfson Professor of Criminology Emeritus, University of Cambridge, Sidgwick Avenue, Cambridge CB3 9DA, UK

6 Distinguished University Professor, University of Maryland, 2220 Samuel J. LeFrak Hall, College Park, MD 20742, USA