Received: 27 July 2018 | Revised: 31 July 2019 | Accepted: 6 August 2019

DOI: 10.1111/1745-9125.12231

CRIMINOLOGY

ARTICLE

Testing hot-spots police patrols against no-treatment controls: Temporal and spatial deterrence effects in the London Underground experiment*

Barak Ariel[1] | **Lawrence W. Sherman**[2] | **Mark Newton**[3]

[1]Jerry Lee Centre for Experimental Criminology; Institute of Criminology, University of Cambridge, and Institute of Criminology, Faculty of Law, Hebrew University of Jerusalem

[2]Jerry Lee Centre for Experimental Criminology; Institute of Criminology, University of Cambridge, and Department of Criminology and Criminal Justice, University of Maryland—College Park

[3]Rail Delivery Group

Correspondence
Barak Ariel, Jerry Lee Centre for Experimental Criminology, Institute of Criminology, University of Cambridge, Cambridge CB3 9DA, U.K.
Email: ba285@cam.ac.uk

Funding information
Jerry Lee Centre of Experimental Criminology

*We would like to express our gratitude to the British Transport Police, with special thanks for their contribution and long-term commitment to evidence-based policing and experimental criminology. Generous support for this research was provided by The Jerry Lee Centre of Experimental Criminology. We particularly wish to thank Jerry Lee, for his commitment to experimental criminology and the global advancement of evidence-based policing. We also thank David Weisburd, Alex Sutherland, Crispian Strachan, Geoff Barnes, the *Criminology* editors, and the anonymous reviewers of this article for their useful comments. We are indebted to John MacDonald for his insightful statistical advice and critical read of earlier versions of this article.

Abstract

Our understanding of causality and effect size in randomized field experiments is challenged by variations in levels of baseline treatment dosage in control groups across experiments testing similar treatments. The clearest design is to compare treated cases with no-treatment controls in a sample that lacks any prior treatment at baseline. We applied that strategy in a randomized test of hot-spots police patrols on the previously never-patrolled, track-level platforms of the London Underground (LU). In a pretest–posttest, control-group design, we randomly assigned 57 of the LU's 115 highest crime platforms to receive foot patrol by officers in 15-minute doses, 4 times per day, during 8-hour shifts on 4 days a week for 6 months. The effect of 23,272 police arrivals at the treatment hot spots over 26 weeks was to reduce public calls for service by 21 percent on treated platforms relative to controls, primarily when police were absent (97 percent of the measured effect). This effect was six times larger than the mean standardized effect size found in the leading systematic review. This finding provides a benchmark against the baseline counterfactual of no patrol in hot spots, with strong evidence of residual deterrence and no evidence of local displacement.

KEYWORDS

baseline dosage, hot spots, no-treatment controls, randomized experiments, regional deterrence, residual deterrence

What is the logic of causality in hot-spots policing experiments? That question has become a prime case in point for debating what conclusions can and cannot be drawn from randomized field experiments (Nagin & Sampson, 2019). Although some may see the three-decade history of hot-spots experiments (Braga, Papachristos, & Hureau, 2012; Sherman & Weisburd, 1995) as providing highly certain conclusions, others may see substantial uncertainties in the same evidence. As Nagin and Sampson (2019, p. 141) concluded, "[I]t is important to be clear with policy makers and our fellow scientists about the uncertainties in the evidentiary base even where experimental evidence forms the point of departure for the translation of results to policy." The uncertainties they identified included the long-term effects of a whole-city system of hot-spots policing, which has never been compared with cities without such a system. They also cited the risk of "cognitive narrowing," by which the logic of causality is confused with research methods or designs, in which a randomized trial is seen to be not just a "gold standard" but also a magic bullet that lends certainty to any claim of causality.

In this article, we address a form of uncertainty that has not yet been debated even though the issue is central to the logic of causality in hot-spots policing. Our concern is that the effect size of redistributing police patrol presence depends just as much on the "dosage" of patrol in a *control* group as it does on the dosage in the treatment group. That logic is independent of whether a test is randomized, of whether it is short term or long term, or of exactly what police do (or do not do) when they are present in hot spots. It is also independent of whether "causality" is defined in the ambitious terms proposed by Nagin and Sampson (2019, p. 124): "[T]he difference between counterfactual worlds that emerge as a consequence of their being subjected to different universal treatment regimens over a sustained period of time, what economists would call system-wide equilibrium differences." Our purpose is not to contest that definition of causality but to accept it as one of several possible valid definitions, as well as to use that definition to reassess and enlarge the body of experimental evidence on hot-spots policing.

The logic of comparing counterfactual worlds must begin with a clear description of both worlds being compared. In our view, the major uncertainty in hot-spots policing experiments is the general failure to provide such clarity. The lack of clarity has several dimensions, including the absolute size and variance of hot-spots boundaries, the frequency and types of crime or calls for service analyzed, and even the harm level of crimes in each location (Weinborn, Ariel, Sherman, & O'Dwyer, 2017). None of those characteristics of the sample locations, however, are as theoretically important to the concept of causality as are the baseline level of the independent variable (police patrol dosage) in the control (C) group and the corollary ratio of patrol dosage in the T group to the level in the C group.

This ratio between T and C dosage has two dimensions. One is the baseline of dosage in both areas prior to beginning the experiment, which should ideally start out as equivalent at the point of random assignment. The other is the size of the gap created between T and C by implementing the experiment's random assignments. Of these two dimensions, we suggest that the absolute baseline level is the essential starting point in developing the counterfactual logic of the experiment. That, in turn, offers more help to reduce the uncertainties about the estimated effect sizes of adding patrol dosage to hot spots, at least in the short run, in any community that already provides police patrol.

The logic of causality in hot-spots policing experiments is often vaguely described as testing whether "adding more" police to high-crime locations prevents more crime compared with business-as-usual (BAU) levels of patrol in similarly high-crime locations. Systematic efforts to integrate the results of experiments so described are of great interest (Braga et al., 2012). Yet the uncertainties of any meta-analysis of the currently available research are massive. These uncertainties begin with the lack of any measurement, in most trials, of patrol dosage in even the experimental treatment (T) group. More fundamentally, the uncertainties depend on *patrol dosage in the counterfactual*: the level of police presence in the control (C) group. Absent this information, any estimates of average effect size

lack sufficient basis for inferring causality. The best demonstration of that claim is the fact that no "moderator analysis" has even been possible of whether effect size depends on the absolute value of patrol presence in either the control group or the treatment group (Braga et al., 2012).

The results of our analysis show both consensus with and departure from the Nagin and Sampson (2019, p. 123) definition of causality as "the difference between counterfactual worlds that emerge as a consequence of their being subjected to different universal treatment regimes." Our findings indicate support for every word except "universal." In the Nagin and Sampson framework, every hot spot in a city would have to have similar treatment to create a "universal treatment regime." They argued that condition is essential for the external validity of conclusions as applied to a world in which such universal treatment is created. Our claim is that such a world is unlikely to be created soon given the empirical evidence we cite in this article about the dosage of police patrol time in (aboveground) hot spots: Across experiments with precise measures, patrol dosage is both highly variable and massively underdelivered. Unless and until our nonfeasibility claim is falsified, we propose that causality tests can comprise individual hot spots, rather than entire systems of hot spots, as an appropriate unit of analysis. On that premise, we suggest that experiments within cities will remain useful for generalizing to other hot spots within cities, but only to the extent that the scholars conducting such experiments can succeed in reducing uncertainty by creating counterfactual worlds that are precisely defined, measured, and replicable.

1 | REDUCING UNCERTAINTY FOR HOT-SPOTS EFFECT SIZES: A HYPOTHESIS

Reducing uncertainty about effect sizes in conclusions from hot-spots policing experiments cannot be accomplished by better measurement alone. Reducing effect size uncertainties requires a purposive, theoretically guided research strategy for testing added patrols under a range of baseline patrol levels. That strategy can build on existing research by filling in the major gaps in the levels of baseline patrol reported, at least in the growing numbers of studies in which Control ("C") group measures are reported.

The theoretical basis for the strategy can be premised on nonrandomized evidence that the highest crime levels occur with a complete absence of policing, as in the case of police going on strike (Nagin, 1998; Sherman, 1990). It follows that if baseline levels of patrol are already high, it will be more difficult for a hot-spots experiment to approximate the radical counterfactual of a police strike versus substantial police presence. The logic of that finding, compared with a range of existing evidence about higher levels of police presence, implies our following hypothesis: *The higher the baseline levels of patrol in hot spots become, the lower the effect size of added patrols is likely to be*. Testing that hypothesis requires a portfolio of research with a wide range of baseline levels of C group patrols. To develop an evidence-based theory of effect sizes dependent on baseline dosage, that portfolio of experiments should include as many levels of control group dosage as possible.

Indispensable to such a research strategy is at least one—and better yet, many—experiment(s) in which the baseline level of C group patrol presence is zero. Without that starting point, our hypothesis remains nonfalsifiable. In our hypothesis, the largest effect size will be gained by adding patrol where it does not currently exist. Researchers testing this hypothesis could find support in a linear progression of declining effect sizes of additional T group patrol with higher baseline levels of C group patrol. These researchers may also, however, find evidence of a nonlinear progression, with empirical evidence for one or more tipping points of baseline patrols, above which increased T group patrolling could have

sharply reduced effects, or perhaps tipping points of the ratio of T to C, or even in the length of time in between patrols.

As more police officer "pracademics" design (e.g., Mitchell, 2017; Williams & Coupe, 2017) and implement their own hot-spots experiments, including GPS measures of patrol time in both C and T group hot spots, the portfolio of experiments with well-measured dosage seems likely to keep growing. A theoretical framework is needed to guide the growing research, both in designing the experiments and in drawing conclusions from them. Yet the incentives for this work will be enhanced by an experiment designed to offer a zero-patrol counterfactual world, as offered in this article.

1.1 | Effect sizes to date

The implicit concern of Nagin and Sampson (2019) in their critique of hot-spots patrol experiments is that consistently positive results may lead to overstating the case for a hot-spots policy, as well as to understating the uncertainties. Our review leads to the opposite concern: that the state of the research so far may be *understating* the case. Reported effect sizes to date are generally small. At the same time, our limited evidence is that the level of control group dosage in the existing portfolio of experiments may be substantial—thereby reducing the estimated benefits of more patrol causing less crime. On the basis of our hypothesis about effect sizes for these tests, the mere addition of more patrol time to T group hot spots may be fighting an uphill battle to deter crime—at least compared with lowering the level of patrol in the C group. This may explain the finding of Braga et al. (2012) that across 10 randomized and quasi-experiments in hot-spots patrols using nonzero dosage in the control groups the mean effect size was quite small ($d = .113$; $p \leq .001$; Braga et al., 2012, Figure 11.9).

The conventional view might be that the sizes of effects of extra policing on crime or disorder outcomes depends on the amount of dosage delivered in the treatment group. Yet exact measures of treatment group dosage have so far been rare. In in all four randomized patrol experiments with measures of dosage included in the most recent systematic review of hot-spots policing (Braga et al., 2012, Figure 11.4), the effect sizes were all in the range of what Cohen (2013) called "small" ($d \leq .2$): from the largest in Jersey City (Weisburd & Green, 1995) at $d = –.147$ ($p = .59$) to the smallest in Jacksonville (Taylor, Koper, & Woods, 2011) at $d = +.055$ ($p = .57$). The only one of the four RCTs in which both T and C police presence was systematically measured was in the initial hot-spots patrol experiment, in which the ratio of mean daily T-group patrol to C-Group patrol was 1.99 to 1 (Sherman & Weisburd, 1995, p. 638), and the effect size of that treatment difference on outcomes was $d = –.061$ ($p = .00$; Braga et al., 2012).

Since then at least two other hot-spots patrol experiments have been conducted with measures of dosage in both T and C groups. Taken together, in these three "dosage-measuring" experiments, "small" effects of treatment differences on outcomes were discovered. In all of them, nontrivial amounts of dosage were measured in the control group. Arranged in order from lowest to highest baseline dosage, Sacramento (Mitchell, 2019, p. 78) had the lowest C group dosage of patrol at 3.4 percent of study hours, as measured by GPS-Automatic Vehicle Locator (AVL) devices on the patrol cars (a mean of 693 minutes per day across 21 control hot spots = 33 minutes per day per hot spot, divided by 16 hours daily, 0900 to 0100, or 16 hours × 60 minutes = 960 minutes, so that 33 / 960 = 3.4 percent of study time had a police car present in the control group). In Peterborough, England (Ariel, Weinborn, & Sherman, 2016, *p*. 296), the C group dosage of patrol was 3.8 percent of the study time of 7 hours (1500 to 2200) = 420 minutes daily as the denominator, with GPS locations measured by radios of police community service officers (PCSOs) on foot (showing 15.9 minutes per day spent inside C group hot spots divided by 420 = a mean of 3.8 percent of study time had PCSOs in the control hot spots *n* the initial hot-spots experiment). Sherman and Weisburd (1995, p. 638) reported about twice

as much patrol dosage in the control group as the other two experiments (7,500 hours of systematic observation by 16 field observers, which showed uniformed police presence at 50 control hot spots for 7.48 percent of all observed minutes).

On the basis of higher dosage of patrol in the control groups alone, we might use our theoretical perspective to predict a smaller effect size in Minneapolis than in the other two experiments. That prediction would be correct since the effect in Minneapolis (d) was –.06, compared with –.10 in Sacramento and –.21 in Peterborough. We can also note the ratio of patrol time between T and C hot spots in each of the tests, at 3 to 1 in Sacramento, 2.3 to 1 in Peterborough, and 1.99 to 1 in Minneapolis. (See table 7.) Although the ordering is not direct, we have the two newer experiments showing lower C group dosage and larger effect sizes than in Minneapolis. Even though this could be a result of differences of measurement, it is at least a clue to solving the mystery of whether control group dosage matters. [We should also note that in Peterborough, the experiment in foot patrol by PCSOs was implemented in the context of a high but constant level of automobile patrol presence by Police Constables, averaging 28 minutes daily in T group and 26 minutes in C group hot spots, a nonsignificant difference at $p = .488$].

Using the citizens calls for service data in all three experiments as a common currency for effect size on outcomes, this limited evidence does show smaller effect sizes where the baseline level of patrol (as a percent of study time) was larger. In Sacramento (Mitchell, 2019, p. 109), the mean effect size across calls for service in the 21 hot-spots pairs was –.104 ($p = .63$). In Peterborough (Ariel et al., 2016), the pooled Cohen's d across calls for service in the 34 T spots and 38 C spots was $d = –.211$ (nonsignificant CI = –.676 to +.252). In Minneapolis (Braga et al., 2012: Figure 11.4), the effect size on total calls for service was calculated at –.061 ($p = .00$). (At the same time, Minneapolis had the only statistically significant [$p = .05$] outcome effect, but that was primarily a result of its greater statistical power, with a larger sample size and a full year of testing.)

As our theory of control group dosage predicted, then, Minneapolis, as the experiment with the highest dosage in the control group (and lowest ratio of T to C dosage) had by far the smallest effect size. At $d = –.061$, the effect size in Minneapolis was 70 percent lower than the $d = –.211$ in Peterborough and 40 percent lower than the effect size in Sacramento. Although these three observations are not many, use of a Bayesian approach would not dismiss them but seek more information. That is exactly what follows, with our new experimental answer to the following key theoretical question: *Would a zero-patrol control group produce a substantially larger effect size than most hot-spots experiments to date*?

1.2 | Zero-treatment control group

A no-dosage control condition has a long history in other fields where what is called a "Knockout trial" is employed. This strategy is used in gene therapy research in which there is a complete "knockout" of one gene or genetic variant (Austin et al., 2004). The same design was used to measure the level of violence that would occur in a troupe of Macaque monkeys if the alpha males were removed from the group, with all violent acts counted carefully before, during, and after the "knockout" of policing (Flack, De Waal, & Krakauer, 2005; Flack, Girvan, de Waal, & Krakauer, 2006). The logic of a knockout comparison is that it helps to eliminate rival hypotheses, as well as to isolate a causal pathway to the hypothesized result of an intervention.

What we present in this article, in slight contrast, is a "reverse-knockout" experiment. Rather than removing police from an environment they were in already, we measure what happens when they are introduced for the first time in an environment where they had never before patrolled for more than150 years. The result is a substantial gain in our logical ability to compare counterfactual worlds, irrespective of research methods.

We accept that finding a world with baseline levels of zero police patrol does not directly address a larger question of system-wide effects (Nagin & Sampson, 2019); for that we suggest the best design to create counterfactuals is to compare a city-wide strategy of hot-spots policing (or not) across a sample of cities. What a reverse-knockout design does offer, however, is a clear counterfactual for answering a different question. That question is arguably just as valid and important as the city-wide "system" question, at least for local police commanders: *whether a given level of directed patrol at any particular location reduces crime at that location, relative to no directed patrol at all*.

Given a low level of crime, the "subway" (as New Yorkers call it) or the "Underground" rapid transit system of London had never deployed police patrols onto the platforms where passengers gather to wait for and leave trains. When they decided to do so in 2011, authorities of the London Underground agreed (after attending a Cambridge University conference on evidence-based policing) to launch the policy with a randomized controlled trial (RCT).

2 | DETERRENT EFFECTS OF POLICE PATROLS IN SPACE AND TIME

The causal logic of the London Underground (LU) experiment is aimed at offering a potential to change our understanding of deterrence at individual hot spots in several ways. In addition to testing whether a larger effect size can be found with a no-treatment control group, the design is also intended to offer new insights into the temporal and spatial dimensions of deterrent effects. It allows for a comparison of effect sizes not only between treatment and control hot spots but also between those groups in temporal-spatial boundaries defined by the following:

- Times of day when police patrol and when they do not
- Days of the week when police patrol and when they do not
- Areas in the stations that they patrol and areas that they do not

These new findings on the spatial and temporal dimensions of the effects of police presence in places are all enhanced by the LU experiment as the first test of all those effects against a control group with zero-directed police patrol and minimal total presence.

The limited evidence on temporal and spatial dimensions of hot-spots policing is central to understanding the counterfactual logic of experiments. In the years since the authors of the National Research Council report on policing (Skogan & Frydl, 2004, p. 247) concluded that experimental evidence on "hot-spots" policing reflects a "strong body of evidence [which] suggests that taking a focused geographic approach to crime problems can increase the effectiveness of policing," much more evidence has since been offered on that general question (e.g., Braga et al., 2012, 2014; Ratfcliffe, Grof, Haberman, Sorg, & Joyce, 2013; Ratcliffe, Taniguchi, Groff, & Wood, 2011; Rosenfeld, Deckard, & Blackburn, 2014; Taylor et al., 2011; Telep, Mitchell, & Weisburd, 2014). Yet on the specific issues of spatial and temporal effects of police patrols, we remain short on both theory and evidence.

Although research evidence has grown somewhat on the *spatial* issues of displacement versus diffusion of benefits (Groff et al., 2015; Ratcliffe, 2016; Taylor et al., 2011; Weisburd et al., 2006), we remain ignorant of the temporal dimensions of deterrence effects and their decay. The correlational analysis conducted by Koper (1995), in which greater 30-minute post-police presence deterrent effects from longer police patrol times were shown, has never even been replicated, let alone tested with an RCT.

A theory of deterrence from police patrols requires a framework for understanding not only *where* its effects occur but also *when* they occur: when police are present, when (and for how long) they are

absent, or both. The evidence that people commit less crime in the presence of a police officer (Ariel & Partridge, 2017; Koper, 1995) might sustain a concept of "direct" deterrence, much like the mission of a security guard. But there are too few police to provide constant "guard duty" at all hot spots at all times. The evidence so far, indeed, indicates that the effects of extra policing on crime are found mostly when police are *absent* rather than when they are *present* (Koper, 1995, p. 658). The LU experiment was designed to explore and distinguish both categories of effects.

We also propose that temporal analysis can be applied to two kinds of space in most experiments with police dosage: "local deterrence" where police are (or have been recently) present and "regional deterrence" in micro-areas surrounding and including where police are or have been present. Residual effects of police presence can also change over time: lingering, decaying, reappearing when people expect police to be present, and disappearing when police presence fades from memories of people in places.

Taken together, these moving parts indicate the following spatiotemporal local deterrence mechanisms of causation. *Spatially*, sanction threats should be highest at the immediate vicinity of the intervention site, with diminishing effect as we move outside the epicenter of the intervention location. The diffusion of the benefits should, theoretically, decay because the effect should be increasingly diluted as the geographic area expands beyond the reach of visible deterrence. *Temporally*, we expect "initial local deterrence" to carry the strongest effect at the time when the tactics are being applied. We defined "initial deterrence" (Sherman, 1990) for these purposes as a response to perceptions of sanction risks while police are still present in the hot spots, when local deterrence is theoretically at its peak. What we do not expect, but could logically occur, is a "phantom" effect: that crime would decline when people *expect* police to be present, even if they are not—based on recent patterns of police patrol.

3 | RESEARCH SETTING

The LU was opened in 1863 as the first underground railway in the world, with a station layout that has changed little over time. Figure 1 displays a platform in 2018. Unlike the New York MTA Subway, the Washington DC METRO, or the San Francisco BART, the platforms in the LU generally allow passengers access to only one track at a time. Therefore, at least two walled-off chambers exist in most stations: one for the trains going in one direction and one for trains going in the other (e.g., westbound or eastbound). Each platform offers only one direction of travel from one track. To get from one platform to the other, passengers may have to climb stairs. If police are present on one platform, they cannot be seen on the other. This layout is not usually found in the more suburban stations located above ground. It is characteristic, however, of the highest crime stations from which we drew our sample.

Each *station* has multiple platforms. Each station has an entry area (the concourse) and an area where tickets can be purchased, a barrier system that can be observed by the ticket-sellers and other station staff, a series of escalators and elevators, and two or more platforms below. Some stations, such as Piccadilly or Kings' Cross, have multiple city-wide train *lines*, with two platforms per line. These larger stations also have much more passenger traffic in the entry areas. Overall, LU provides more than 1.3 billion passenger rides per year (Transport For London, 2015).

With the greatest concentration of suitable victims and motivated offenders (Cohen & Felson, 1979) on the trains themselves, nearly 30 percent of reported calls for service document crimes occurring on board and in between platforms (Ariel, 2011). Within the stations, clearly identifiable areas attract substantial proportions of all crime, such as the concourse and the booking area (12 percent) and the entry/exit bars (8 percent). Crimes on the *platforms*, which historically had no police patrols, constituted just 11 percent of all reported crimes.

FIGURE 1 LU platform, 2018 [Color figure can be viewed at wileyonlinelibrary.com]

Just prior to the experiment, the distribution of crime in the LU consisted of theft (35 percent), less serious fraud (12 percent), and less serious public disorder (10 percent), with more serious offenses far less prevalent, such as robbery (.3 percent), sexual crimes (.7 percent), or violence more generally (4 percent). It is important to note, however, that many "crimes" in the LU are in fact police generated (e.g., results of stops and frisks or ticket violations accounted for by police officers). For instance, the entry/exit bars, which are the location of larcenies by turnstile jumpers, have long been the priority for police patrol, and so 65 percent of criminal events are police generated at these locations. Yet these police-generated crimes occur at the ground level of the stations, *not on the platforms* or the on the trains, because both of the latter categories of LU area had never been proactively patrolled prior to this study. In our LU experiment, in any event, we exclude police-generated crimes as outcome measures.

Crime and disorder are not distributed evenly around the LU network. The top 5 out of 270 stations (Stansted, Kings Cross, Victoria, Oxford Circus, and Leicester Square) account for more than 12 percent of all recorded crimes in the LU system. Some times are hotter than others as well, with 12 hours over 4 days (5 pm–8 pm, Wednesdays–Saturdays) accounting for a little more than 26 percent of all crime during the week. The design of the experiment allowed for us to identify the population—and not a sample—of the platforms with the highest frequencies of crimes and calls for service.

4 | DATA AND METHOD

For the purpose of the experiment, LU's detachment of the British Transport Police (BTP) granted us access to all calls for service and crime data. The experiment was in operation for 6 months, commencing in mid-September 2011 through mid-March 2012, with comparisons to the same months of the year prior (2010–2011). Calls for service data proved to be particularly challenging as key spatiotemporal

variables were recorded as text logs rather than as unique categories for every area of the station. The log for every incident on sample platforms was therefore read and coded into unique nominal groupings of space and time for purposes of the present analysis.

Data on "dosage" of police time in T and C platforms were captured manually by three dedicated sergeants assigned to record via two-way radios every entry and exit to the hot spot. GPS trackers were not available underground as GPS relies on satellite connectivity (for more recent alternative methods of tracking, see De Brito & Ariel, 2017). The extensive paper record created supervisory documentation of 20,237 visits of 15 minutes each to the treatment hot spots. No visits were recorded in the control hot spots.

4.1 | Unit of analysis: LU platforms and environs

The choice of unit of analysis in this study was driven by the theoretical question we addressed: Platforms, not stations, provided the ideal means for exploring the spatial and temporal dynamics of deterrence. Platforms experience only one tenth of all LU crime, but they are small, stable, and confined places with finite entry and exit points—characteristics that make them optimal for measuring local deterrence effects. A "platform" is therefore defined as the physical space where travelers embark on or disembark from the train, whereas a "nonplatform" is any place within the station complex that is not the platform area. Likewise, each hot spot is defined as a combination of "train line" (e.g., Piccadilly Line) times "direction" (e.g., westbound) times "station" (e.g., Russell Square London Underground Station), thus, avoiding confusion about the boundaries of the hot spots discussed in earlier studies (see Sorg, Wood, Groff, & Ratcliffe, 2014). Officers could not construct their patrols outside arbitrary digital lines on the map. They were physically constrained by the walls around them and the floor beneath them.

To target the hottest platforms, we rank-ordered all LU platforms according to the level of crime they experienced in 12 months. We excluded platforms that experienced fewer than two crimes per year. This threshold is much lower than in some previous hot-spots experiments, such as in that of Sherman and Weisburd (1995), in which they used a threshold of 20 crimes per year. We have also excluded the six major London *stations* that have been targeted and routinely patrolled by special "Hub Teams" (for example, Victoria and Kings Cross Stations) after a series of coordinated suicide bomber attacks in London on July 7, 2005. Platforms that were located too far away from other stations were excluded as well (prior to random assignment) for operational reasons: If a hot spot was 45 minutes away (distance time) from any other hot spots, then it would require too much time for a dedicated patrol unit to reach it and maintain similar dosage levels during a given shift.

Finally, we defined and located hot spots using victim-generated "hard crimes" (see Reiss, 1985; Sherman & Weisburd, 1995; Weisburd & Green, 1995), such as violence, "against-person" antisocial behavior (i.e., excluding vandalism), sex crimes, robberies, or criminal damage. This list excluded passenger theft (but only for selection purposes, not for outcome analysis; too many thefts city wide had no identifiable location). Out of thousands of platforms all over Metropolitan London, through our targeting analysis, we created a list of 115 eligible hot-spots platforms, with an mean of greater than 4.72 crimes per year (standard deviation [SD] = 4.8).

4.2 | Random assignment with partial blinding

We allocated all 115 eligible platforms by Microsoft® Excel™ using simple batch random assignment, creating 57 treatment group and 58 control group hot-spots platforms. Sergeants and patrolling officers were not given the full list of stations. Officers were informed of the location of their treatment hot

spots, but they were blinded to the location of the control hot spots. The purpose of this partial blinding was to decrease the chance of treatment contamination and spill-over violations. Nonetheless, we inspected logs of foot patrols to monitor whether any were recorded for the 58 control group stations (they were not).

4.3 | Treatment delivery

Twenty uniformed officers were selected and trained (and replaced with trained officers as normal turnover occurred) to work exclusively on foot patrols on the treatment group platforms for the prescribed time periods and rotations. They patrolled on foot usually in teams of two uniformed people. Based on the result of baseline temporal analyses, the "hot hours" and "hot days" for the sample platforms were Wednesdays through Saturdays between 3 pm and 10 pm. These were the patrol hours and days to which the officers were assigned.

Each two-person "patrol unit" was accountable for three to five hot spots depending on the traveling distance between their assigned hot spots. Officers did not have discretion about where to conduct their patrols (e.g., they were guided precisely where to go on the days and hours of the experiment); however, they were asked to attend the hot spots in a colloquially "random" or unpredictable order to avoid the predictability of visits. The 20 "ring-fenced" officers seconded to work exclusively on the experiment were debriefed daily by the sergeants, who continuously reminded them to attend the hot spots four times each treatment day for 15 minutes at a time. In these meetings, officers were also encouraged to engage as much as possible with members of the public and not to stand idle on the platforms.

At any given moment during the experimental period, there were 20 officers who conducted these patrols. Given these resource constraints and the distance between the hot spots, each hot spot was therefore assigned four patrols per day, each patrol to last 15 minutes (Koper, 1995), as in several prior hot-spots patrol experiments (Ariel et al., 2016; Telep et al., 2014). The assignments summed to 416 visits per hot spot over 6 months. Immediately after each patrol team's 15-minute visit, officers boarded a train (passengers wait less than 2 minutes on average between trains arriving on platforms in peak times [TFL 2014]) and traveled to the next platform.

Officers were tasked to remain on the platform for each 15-minute patrol visit. The officers were not tasked to problem-solve (Goldstein, 1979), conduct community policing in the classic sense (e.g. Skogan & Hartnett, 1997), or target their efforts on any particular crime category.

Several steps to limit "spillover" police presence (as violations of the stable unit treatment value assumption [SUTVA]) were taken in structuring the treatment: 1) Shifts always began at the BTP main offices, requiring no police presence on control platforms. 2) No station had both a treatment and a control platform. 3) Treatment officers moved between the treatment hot spots by traveling on the trains, although they had to walk through a nearby station to get to and from a platform at the beginning and end of each shift. 4) The officers were asked (without disclosure of the location of control platforms) to follow pathways that bypassed the control platforms in their transit between treatment platforms, to avoid spillover effects.

Notwithstanding these steps, it is possible that some spillover of police visibility onto control platforms occurred. People may have seen police on a treatment platform that they entered by mistake; after which, they went back into the station to board a different line and then exited at a control station. Offenders traveling to a control station may have seen the treatment officers on a train heading beyond the control platform to a treatment platform at the next station. Any number of such possible interactions may have occurred.

The key point about such spillover is that it would bias the estimated treatment effect *downward* because unintended police presence could have reduced crime in control as well as in experimental

platforms (see Ariel, Sutherland, & Sherman, 2018). The experiment is not perfect, but arguably the risk of *overestimation* of treatment group differences from controls has been restricted even if some risk of underestimation remains.

Finally, based on previous similar field experiments (Drover, 2015; Sherman & Weisburd, 1995), it was clear that strong leadership was essential. To generate unflagging motivation throughout the entire research cycle, a member of the research team held monthly half-day meetings with all the officers, sergeants, and senior officers, in which the importance of treatment fidelity was communicated strongly. These "motivational debriefs" provided monthly feedback to the patrol officers about particular noteworthy incidents, crime figures during the previous month, and treatment dosage data.

The total dosage in the 58 treatment hot spots was approximately 5,000 hours. No proactive patrols were assigned or recorded in any of the control hot spots.

4.4 | Measurement of the independent variable (Patrols)

As noted, three dedicated sergeants were assigned to the experiment to capture, manually, all data on patrol "dosage." Police officers called in via two-way radios every time they entered and again at every exit from each treatment hot spot. The sergeants wrote down each radio call with time, place, and date on paper forms for every officer and for every hot spot. This procedure produced 20,237 paper records that were then coded by the research team into SPSS™ for further analyses.

In addition, the sergeants conducted at least one biweekly random "surprise" supervisory visit to the hot spots, throughout the entire experimental period, to safeguard the treatment's integrity. Officers were not notified prior to these supervisory visits but were always approached by the sergeant when contact was made. These visits were recorded and then communicated during the "motivational debriefs."

4.5 | Measuring dependent variables

We used two outcome measures to assess the treatment effect: all calls for service to the police ("999 calls") and all reported crimes within the participating hot spots, during the 6 months of the experiment (2011–2012) and for the 6 months in the (baseline) year prior to the experiment (2010–2011). As noted, police-generated crimes—that is, incidents that are *outputs,* such as drug arrests and stop-and-searches—were excluded from the outcome data (Ariel et al., 2016; Sherman & Weisburd, 1995).

Data were then broken into eight outcome variations, reflecting the "types" of deterrence to be tested in the experiment: four "incident types" (that is, both calls for service and crimes) that occurred on the platforms only—1) during the (intermittent) patrol hours on patrol days; 2) during patrol days but not during (intermittent) patrol hours; 3) on nonpatrol days but during the same hours as patrol hours on patrol days; and 4) on nonpatrol days at hours that were not normally patrol hours—and then four more times for different temporal periods in the nonplatform areas of the train station. The following eight outcomes account for varieties of deterrence effects discussed earlier:

1) Local direct deterrence on platforms during scheduled patrol hours
2) Local short-term residual deterrence outside of scheduled patrol hours on scheduled patrol days
3) Local long-term residual deterrence inside of scheduled patrol periods on nonpatrol days
4) Local long-term residual deterrence outside of scheduled patrol periods on nonpatrol days
5) Regional direct deterrence all over the station during scheduled patrol hours

6) Regional short-term residual deterrence all over the station, outside of scheduled patrol hours on scheduled patrol days
7) Regional long-term residual deterrence inside of scheduled patrol periods on nonpatrol days
8) Regional long-term residual deterrence outside of scheduled patrol periods on nonpatrol days

5 | STATISTICAL ANALYSIS

We estimated the results of the experiment using generalized linear models as a way to assess the differences between experimental and control hot spots in terms of two outcome measures: 1) citizen-generated calls for service incidents and 2) citizen-reported crime incident counts. We ran each model twice, each time with eight iterations for the eight possible outcome variations described earlier—once for calls for service and then again for crime counts. Treatment assignment was used as the predictor in each model, but we also include the preassignment value of the outcomes as a covariate to make the analysis a before–after difference of differences (therefore the model is postoutcome = intercept + preoutcome + treatment assignment). Using the baseline measure of the dependent variable as a covariate is a common method for "partialing out" the relationship between treatment assignment and the outcome, net of baseline differences between locations (see Campbell & Stanley, 1963, *p.* 23; Senn, 1989; Vickers & Altman, 2001). The process also increases the statistical power of the test (Cohen, 2013) compared with using difference scores.

We note that there are different ways to model the treatment effect in pretest–posttest control group designs (for further technical description on our design, see Shadish, Cook, & Campbell, 2002, p. 261). The two most common models are 1) using the posttest as a dependent variable and the pretest as a covariate (ANCOVA), or 2) using the difference between posttest and pretest as dependent variable (change scores). The former is focused on raw change scores, whereas the latter is focused on residual change scores. These different approaches may produce different and even conflicting results. When we applied a change score analysis, some of our comparisons resulted in nonsignificant and reversed differences between some comparisons between the study conditions. This was not the case with ANCOVA, using the pretest scores as covariates, given variation around a specific day. From a practical perspective, controls for the baseline control level were effectively estimated in the ANCOVA model in comparing experimental period differences between treatment and control. More generally, the authors of most textbooks have agreed that "analysis of covariance with pre-test scores as the covariate are usually preferable to simple gain-score comparisons" (Campbell & Stanley, 1963, p. 23). Using the pretest as a covariate is generally considered to be a statistically more powerful analysis than a difference or gain score analysis (Becker, 2000). More recently Van Breukelen (2013, p. 896) reviewed influential methodological publications over the last 60 years and concluded that they overwhelmingly recommended ANCOVA over change scores in analyzing randomized studies. We note that we did not add an interaction term as the procedure reduces the statistical power of the test.[1]

Furthermore, we have used a Poisson regression model on the basis that the dependent variables were structured as counts. We applied what are commonly referred to as Bayesian Information Criteria (BIC) to allow for comparing different distribution models. We conducted 32 tests: twice for each of our 16 comparisons. In all tests the directionality of the comparisons was identical (i.e., a reduction in crime relative to control conditions). The BIC values of the model fits indicated that the most appropriate functional form of the variance was the Poisson regression models.

[1]Ideally, we would use a Solomon Four Group Design to investigate the effects in relation to the pretesting scores; however, there were not enough units of analysis in the higher crime platform population for that design.

Based on the results of the models, we then computed estimated marginal means, which provide the mean responses for each factor, and adjusted for the covariate (i.e., baseline scores). These scores more accurately depict the adjusted treatment effect as they take into account the outcome variable at baseline values. We then converted the means into standardized mean differences (*d*) in Cohen's *d* values (Cohen, 1988) and presented the outcomes in forest plots using Comprehensive Meta-Analysis 2.0. The computations enabled us to compare the outcome variations even though the base rates of the comparison are not equal. The forest plots are used solely for visualizing the different deterrence effects, not to estimate an overall effect from comparisons with overlapping data. By the nature of the data, the point estimates (i.e., eight for each outcome variable) are not independent of each other, and therefore pooling the outcomes into a composite effect size would underestimate the degree of heterogeneity (see Cheung & Chan, 2004). Although there are statistical corrections for dependent effect sizes (see Gleser & Olkin, 2007, Hedges, Tipton, & Johnson, 2010), our interest lies primarily in comparing between the multiple end points rather than in composing an overall weighted average effect size. The interrelationship between the point estimates does not invalidate the comparisons between the multiple measures (Scammacca, Roberts, & Stuebing, 2014, p. 330).

5.1 | Statistical power

A total of 115 hot spots (57 treatment and 58 control) were used in this experiment, which created a study with sufficient statistical power, defined by Cohen (1988) as the probability of detecting a statistically significant outcome in an experiment, given the true difference between the treatment group and the control group. By using *Optimal Design* (Spybrook et al., 2013), we estimated that this sample size is large enough to detect small-to-medium effects, in which the significance level is .05. The hypotheses are assumed to be nondirectional, and the estimated power is .8.

6 | RESULTS

Tables 1 and 2 show the total number of incidents in the hot spots, broken into calls for service and crime counts, before and during the experimental period. Over the experimental period, the stations attracted 6,366 calls for service and 3,606 crimes. The 115 platforms accounted for 14 percent of the calls ($n = 898$) and 11 percent of all victim-generated crimes ($n = 402$) recorded in the 115 stations.

In table 2, the balance at baseline yielded by random assignment is tested; none of the pretreatment between-groups comparisons were significantly different for the eight comparisons. At both the levels of local and residual deterrence, the units share the same distribution of calls for service and crime prior to the allocation of treatment in the experimental group. When calculating the baseline differences of the different combinations of space and time, we found that no statistically significant differences between treatment and control conditions within the eight study comparisons were detected.

During the experimental period, a total of 3,549 calls for service were identified for control group platforms, and 2,817 calls for service in experimental hot spots, a relative difference of 21 percent fewer total calls in the experimental platforms compared with a relative difference of 12 percent during baseline (3,218 control compared with 2,834 in experimental), showing 10 percent relative reduction in calls for service. The relative difference in recorded crimes was 14 percent lower in treatment than in control in the experimental (1,668 vs. 1,938) and baseline period (2,079 vs. 2,392). The lack of total difference in crimes, however, masks important subgroup differences.

TABLE 1 Calls for service (CFS), means (*M*), standard deviations (*SD*), and counts: Baseline experimental data in control and treatment hot spots

	Control Hot Spots ($n = 58$)					Treatment Hot Spots ($n = 57$)				
	Baseline Period		Experimental Period		Post–Pre Difference	Baseline Period		Experimental Period		Post–Pre Difference
Variable	*M* (*SD*)	Count	*M* (*SD*)	Count	*M*	*M* (*SD*)	Count	*M* (*SD*)	Count	*M*
1. Platforms only, Wed. to Sat., Patrol hours	2.086 (2.187)	121	2.500 (3.230)	145	.414	2.193 (2.887)	125	1.807 (2.191)	103	−.386
2. Platforms only, Wed. to Sat., Nonpatrol hours	2.466 (2.549)	143	2.931 (3.066)	170	.465	2.509 (3.129)	143	2.035 (2.478)	116	−.474
3. Platforms only, Nonpatrol days, Patrol hours	1.224 (1.285)	71	1.603 (1.964)	93	.379	1.193 (1.737)	68	1.228 (1.414)	70	.035
4. Platforms only, Nonpatrol days, Nonpatrol hours	1.517 (1.525)	88	2.052 (2.358)	119	.535	1.333 (2.231)	76	1.439 (1.946)	82	.106
5. Nonplatforms, Patrol days, Patrol hours	14.828 (14.321)	860	16.017 (14.955)	929	1.189	13.474 (17.299)	768	12.947 (16.265)	738	−.527
6. Nonplatforms, Wed. to Sat., Nonpatrol hours	15.121 (11.997)	877	15.655 (14.166)	908	.534	12.298 (12.989)	701	12.351 (14.959)	704	.053
7. Nonplatforms, Nonpatrol days, Patrol hours	9.948 (8.676)	577	9.897 (8.725)	574	−.051	9.123 (12.169)	520	8.719 (9.985)	497	−.404
8. Nonplatforms, Nonpatrol days, Nonpatrol hours	8.293 (6.150)	481	10.534 (9.152)	611	2.241	7.596 (8.493)	433	8.895 (9.980)	507	1.299
TOTAL/AVERAGE	**6.935 (6.086)**	**3,218**	**7.649 (7.202)**	**3,549**	**.714**	**6.215 (7.617)**	**2,834**	**6.178 (7.402)**	**2,817**	**−.037**

TABLE 2 Crimes, means (*M*), standard deviations (*SD*), and counts: Baseline experimental data in control and treatment hot spots

	Control Hot Spots ($n = 58$)					Treatment Hot Spots ($n = 57$)				
	Baseline Period		Experimental Period		Post–Pre Differences	Baseline Period		Experimental Period		Post–Pre Difference
Variable	*M* (*SD*)	Count	*M* (*SD*)	Count	*M*	*M* (*SD*)	Count	*M* (*SD*)	Count	*M*
1. Platforms only, Wed. to Sat., Patrol hours	1.103 (1.266)	64	1.276 (1.852)	74	.173	1.544 (2.292)	88	1.105 (1.676)	63	−.439
2. Platforms only, Wed. to Sat., Nonpatrol hours	1.741 (1.681)	101	1.224 (1.312)	71	−.517	1.526 (2.399)	87	.877 (1.337)	50	−.649
3. Platforms only, Nonpatrol days, Patrol hours	.793 (1.281)	46	.638 (.931)	37	−.155	.807 (1.563)	46	.544 (.825)	31	−.263
4. Platforms only, Nonpatrol days, Nonpatrol hours	.672 (1.145)	39	.690 (.94)	40	.180	.737 (1.951)	42	.649 (1.061)	37	−.088
5. Nonplatforms, Patrol days, Patrol hours	12.103 (13.834)	702	9.931 (12.151)	576	−2.172	11.228 (15.997)	640	9.386 (13.327)	535	−1.842
6. Nonplatforms, Wed. to Sat., Nonpatrol hours	11.879 (11.23)	689	8.707 (7.199)	505	−3.172	8.842 (12.077)	504	7.333 (9.253)	418	−1.509
7. Nonplatforms, Nonpatrol days, Patrol hours	6.741 (6.083)	391	5.431 (5.225)	315	−1.310	7.053 (9.469)	402	4.702 (6.62)	268	−2.351
8. Nonplatforms, Nonpatrol days, Nonpatrol hours	6.207 (5.982)	360	5.517 (4.732)	320	−.690	4.737 (5.811)	270	4.667 (5.789)	266	−.070
TOTAL/AVERAGE	**5.155 (5.313)**	**2,392**	**4.177 (4.293)**	**1,938**	**−.957**	**4.559 (6.445)**	**2,079**	**3.658 (4.986)**	**1,668**	**−.901**

TABLE 3 Generalized linear model for calls for service—parameter estimates and standard errors (SE) (*n* Control Hot Spots = 58; *n* Treatment Hot Spots = 57)

Variable	Intercept (SE)	*B* (SE)	Covariate *β* (SE)
1. Platforms only, Patrol days, Patrol hours	.410***	−.476***	.190***
	(.099)	(.132)	(.016)
2. Platforms only, Patrol days, Nonpatrol hours	.534***	−.429***	.170***
	(.099)	(.121)	(.015)
3. Platforms only, Nonpatrol days, Patrol hours	.073	−.360**	.272***
	(.120)	(.162)	(.035)
4. Platforms only, Nonpatrol days, Nonpatrol hours	.317***	−.455***	.223***
	(.104)	(.148)	(.023)
5. Nonplatform areas, Patrol days, Patrol hours	2.127***	−.298***	.033***
	(.040)	(.050)	(.001)
6. Nonplatform areas, Patrol days, Nonpatrol hours	1.932***	−.196***	.043***
	(.044)	(.050)	(.001)
7. Nonplatform areas, Nonpatrol days, Patrol hours	1.771***	−.249***	.044***
	(.048)	(.063)	(.002)
8. Nonplatform areas, Nonpatrol days, Nonpatrol hours	.317***	−.455***	.223***
	(.104)	(.104)	(.023)

†$p < .10$; *$p < .05$; **$p < .01$; ***$p < .001$ (two-tailed).

6.1 | Fidelity of treatment delivery

We assigned 23,272 visits of 15 minutes each to the 57 treatment hot spots and 0 visits to the 58 control hot spots, with 87 percent of patrol visits delivered. Officers were able to meet the duration target in 94 percent of delivered patrols for periods within ±5 minutes. Across all assignments, approximately 13 percent were not delivered as a result of station closures, line disruptions, or engineering works. Thus, based on manual recording by the three sergeants, our dosage data indicate that the total number of visits in the 6-month period was 20,237 or 355 visits per hot spot (3.4 visits per day). Across the 7-hour (420- minute) daily period of intermittent patrols, 3.4 visits × 15 minutes each = 51 minutes of patrol per platform divided by 420 minutes = 12 percent of target time. This level of presence in the T group was slightly lower than in Minneapolis (Sherman & Weisburd, 1995) but a little higher than in Peterborough (8.9 percent) or Sacramento (10.0 percent). The level of presence in the C group (zero) of the LU experiment was, of course, far lower than the 7.5 percent of C group time in Minneapolis, 3.8 percent in Peterborough, and 3.4 percent in Sacramento (see table 7).

6.2 | Treatment effect estimates on calls for service

Tables 3 and 4 present the model estimates for calls for service in each of the eight outcomes and the estimated marginal means. Eight models, one for each outcome, reflect the eight "types" of deterrence tested in the experiment. Table 3 shows the parameter estimates and the standard errors for both the treatment predictors and the covariate values. Table 4 then lists the estimated marginal means and their associated standard errors of the means for calls for service data. We present both means and standard errors for the experimental and control groups as well as the mean pairwise comparisons.

TABLE 4 Estimated marginal means of calls for service (*n* Control Hot Spots = 58; *n* Treatment Hot Spots = 57)

Variable	Treatment (SE)	Control (SE)	Pairwise Comparisons (SE)	Fixed Baseline Value in Each Model
1. Platforms only, Patrol days, Patrol hours	1.407 (.155)	2.264 (.193)	**.857***** **(.235)**	2.139
2. Platforms only, Patrol days, Nonpatrol hours	1.696 (.169)	2.603 (.210)	**.908***** **(.254)**	2.488
3. Platforms only, Nonpatrol days, Patrol hours	1.044 (.157)	1.495 (.157)	**.452**** **(.201)**	1.209
4. Platforms only, Nonpatrol days, Nonpatrol hours	1.196 (.146)	1.885 (.176)	**.689***** **(.220)**	1.426
5. Nonplatform areas, Patrol days, Patrol hours	9.982 (.413)	13.443 (.462)	**3.461**** **(.576)**	14.157
6. Nonplatform areas, Patrol days, Nonpatrol hours	10.185 (.419)	12.392 (.441)	**2.207***** **(.563)**	13.722
7. Nonplatform areas, Nonpatrol days, Patrol hours	6.940 (.351)	8.907 (.378)	**1.967***** **(.493)**	9.539
8. Nonplatform areas, Nonpatrol days, Nonpatrol hours	7.209 (.358)	9.497 (.391)	**2.289***** **(.508)**	7.948

$^{\dagger}p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$ (two-tailed).

Nearly all outcomes (table 3) are statistically significant at least at the .05 level and all in the hypothesized direction. Looking at the estimated marginal means (table 4), the pairwise comparisons follow the same patterns.

6.3 | Treatment effect estimates on crimes

Tables 5 and 6 list the same type of results for crimes as reported for calls for service. Here, not all treatment predictors are statistically significant at the .05 level; however, all are pointing in the hypothesized direction. The estimated marginal means are equally mixed but with significant predictors on crime found for direct local deterrence during patrol hours and during nonpatrol hours.

6.4 | Residual and regional effects of patrol versus no-patrol

Next, we look at the standardized mean differences of T versus C effects (Cohen, 1988) to show the distribution of direct and residual effects across the temporal and spatial dimensions of deterrence. The forest plots presented here show these results for both call data and crime data (figures 2 and 3). Each point estimate represents the magnitude of the pooled difference between treatment and control hot spots for each of the eight definitions of the spatial and temporal nature of deterrent threat. Anything to the right of the vertical null line (.00) implies that the difference favors the treatment group (that is, fewer events) and to the left of the line indicates fewer incidents in the control group. The further away the point estimate is from the null line, the larger the magnitude of the difference between treatment and control hot spots. Horizontal lines that cross the point estimates represent the confidence intervals. When the horizontal lines cross through the vertical null line, that indicates that the effect is not statistically significant (.05).

TABLE 5 Generalized linear model for crimes—parameter estimates and standard errors (SE) (*n* Control Hot Spots = 58; *n* Treatment Hot Spots = 57)

Variable	Intercept (SE)	*B* (SE)	Covariate *β* (SE)
1. Platforms only, Patrol days, Patrol hours	−.012	−.349†	.199**
	(.125)	(.181)	(.031)
2. Platforms only, Patrol days, Nonpatrol hours	−.282*	−.437*	.230**
	(.141)	(.190)	(.030)
3. Platforms only, Nonpatrol days, Patrol hours	−.546**	−.199	.208**
	(.162)	(.236)	(.035)
4. Platforms only, Nonpatrol days, Nonpatrol hours	−.561**	−.172	.123†
	(.177)	(.244)	(.063)
5. Nonplatform areas, Patrol days, Patrol hours	1.567***	−.139*	.041***
	(.052)	(.061)	(.001)
6. Nonplatform areas, Patrol days, Nonpatrol hours	1.481**	−.104	.044***
	(.057)	(.067)	(.002)
7. Nonplatform areas, Nonpatrol days, Patrol hours	1.090***	−.413***	.071***
	(.067)	(.088)	(.003)
8. Nonplatform areas, Nonpatrol days, Nonpatrol hours	1.004***	−.062	.086***
	(.074)	(.083)	(.005)

$^{\dagger}p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$ (two-tailed).

TABLE 6 Estimated marginal means of crimes (*n* Control Hot Spots = 58; *n* Treatment Hot Spots = 57)

Variable	Treatment (SE)	Control (SE)	Pairwise Comparisons (SE)	Fixed Baseline Value in Each Model
1. Platforms only, Patrol days, Patrol hours	.907	1.286	**.379**†	1.322
	(.128)	(.150)	**(.195)**	
2. Platforms only, Patrol days, Nonpatrol hours	.710	1.099	**.389***	1.635
	(.113)	(.134)	**(.166)**	
3. Platforms only, Nonpatrol days, Patrol hours	.550	.671	**.121**	.704
	(.099)	(.106)	**(.143)**	
4. Platforms only, Nonpatrol days, Nonpatrol hours	.530	.630	**.099**	.800
	(.097)	(.103)	**(.141)**	
5. Nonplatform areas, Patrol days, Patrol hours	6.709	7.706	**.997***	11.670
	(.334)	(.347)	**(.434)**	
6. Nonplatform areas, Patrol days, Nonpatrol hours	6.240	6.921	**.681**	10.374
	(.329)	(.331)	**(.434)**	
7. Nonplatform areas, Nonpatrol days, Patrol hours	3.218	4.867	**1.648*****	6.896
	(.239)	(.281)	**(.346)**	
8. Nonplatform areas, Nonpatrol days, Nonpatrol hours	4.106	4.369	**.263**	5.478
	(.266)	(.265)	**(.351)**	

$^{\dagger}p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$ (two-tailed).

Comparisons	Std diff in means	Standard error	Lower limit	Upper limit	p-Value
Platforms only, Patrol Days, Patrol Hours	0.680	0.192	0.304	1.056	0.000
Platforms only, Patrol Days, Non-Patrol Hours	0.667	0.192	0.291	1.042	0.001
Platforms only, Non-Patrol Days, Patrol Hours	0.419	0.189	0.050	0.789	0.026
Platforms only, Non-Patrol Days, Non-Patrol Hours	0.584	0.190	0.211	0.957	0.002
Non-Platform Areas, Patrol Days, Patrol Hours	1.121	0.201	0.727	1.514	0.000
Non-Platform Areas, Patrol Days, Non-Patrol Hours	0.731	0.193	0.354	1.109	0.000
Non-Platform Areas, Non-Patrol Days, Patrol Hours	0.744	0.193	0.366	1.122	0.000
Non-Platform Areas, Non-Patrol Days, Non-Patrol Hours	0.840	0.195	0.459	1.222	0.000

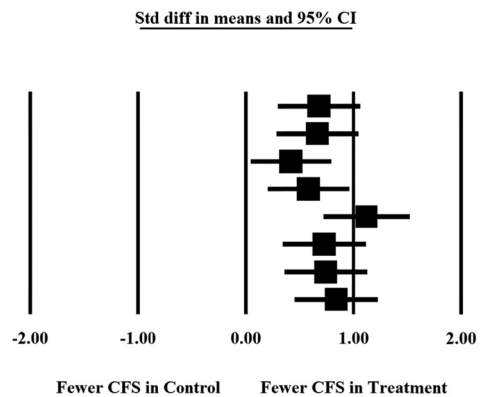

FIGURE 2 Effect of patrol on calls for service (CFS)

Comparisons	Std diff in means	Standard error	Lower limit	Upper limit	p-Value
Platforms only, Patrol Days, Patrol Hours	0.362	0.188	-0.006	0.731	0.054
Platforms only, Patrol Days, Non-Patrol Hours	0.437	0.189	0.067	0.807	0.021
Platforms only, Non-Patrol Days, Patrol Hours	0.158	0.187	-0.208	0.524	0.398
Platforms only, Non-Patrol Days, Non-Patrol Hours	0.131	0.187	-0.235	0.497	0.482
Non-Platform Areas, Patrol Days, Patrol Hours	0.428	0.189	0.059	0.798	0.023
Non-Platform Areas, Patrol Days, Non-Patrol Hours	0.293	0.188	-0.075	0.660	0.119
Non-Platform Areas, Non-Patrol Days, Patrol Hours	0.888	0.195	0.505	1.271	0.000
Non-Platform Areas, Non-Patrol Days, Non-Patrol Hours	0.140	0.187	-0.226	0.506	0.454

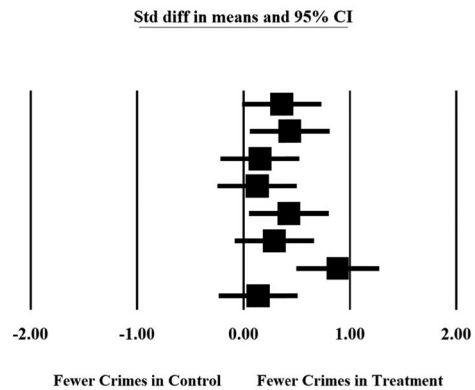

FIGURE 3 Effect of patrol on crimes

Observing the displays of effect sizes of patrols on calls for service during the "direct local deterrence" 7 hours and 4 days of intermittent patrols (comprising the targeted overall 12 percent police presence time in the T group), the estimated effect on all calls for service ($d = –.680$; $p \leq .01$) is the highest ever published for a randomized trial in hot-spots patrol policing. It is more than three times higher than the Peterborough effect on calls ($d = –.211$). Residual deterrent effects of patrols on calls for service about incidents on those platforms on patrol days outside of regular patrol hours are almost as high ($d = –.667$, $p \leq .001$). Substantial (medium) and significant effect sizes are also found on the T group platforms on the 3 days weekly when no patrols are delivered, both during usual patrol hours ($d = –.419$) and outside the usual hours ($d = –.584$).

The results for the regional effects of calls for service in station areas not including the T group platforms are even larger. Far from showing displacement of crime off of the platforms and into other areas of the station, just the opposite occurs. The presence of patrols on platforms may discourage would-be offenders from even entering the nonplatform station areas, where we have seen that most crimes are reported to have occurred.

Temporally, the biggest effect is found regionally during patrol hours and days ($d = 1.121$, 95 percent confidence interval [CI] = .727–1.514), followed by days and hours that officers were *not* conducting patrols ($d = .840$, 95 percent CI = .459–.1222). The smallest effect of hot-spots patrols in the LU was detected locally during the same patrol hours but on nonpatrol days; even that effect size is larger than in any previously reported test ($d = .419$, 95 percent CI = .050–.789).

Figure 3 displays a similar but more complex pattern. The direct deterrent effect of patrols on reported crimes during patrol hours is also larger than previously reported effects ($d = –.362$, $p = .054$). The residual benefit of that patrol in the other 17 hours a day on the same platform is even larger than the direct effect at $d = –.437$ ($p = .020$). Residual effects on those platforms on no-patrol days are not as large as on patrol days, and nonsignificant, but they are in the same order of magnitude as *direct* deterrent effects reported in other experiments. Regional effects of patrol on crime in the station but *off* the platforms on the 4 regular patrol days are almost equal to the direct deterrent effects of patrol in those hours *on* crime on the platforms. Regional, off-platform effects on the 3 no-patrol days offer the highest effect size on crime ($d = –.888$, $p \leq .001$) for the same hours as usual patrol, a "phantom" effect that is stronger than the effect when police are really there.

7 | DISCUSSION

These findings show support for our hypothesis that a no-patrol control group would yield larger effect sizes than those found in previous experiments restricted by nontrivial levels of patrol in their control groups. Table 7 provides a summary of the key features of the four experiments, including this one, with the relevant measures of direct deterrence of events causing citizen-generated calls for service during hours of intermittent patrols. The direct deterrent effect sizes in the LU experiment are three times higher on calls for service than reported in any previous randomized trial from increases in patrol time as the independent variable. It is also the only experiment to deploy a no-treatment control group, in a highly homogeneous population (not a sample) of 100 percent of the most crime-ridden Underground platforms in central London.

These findings reveal two key points. One is that all future hot-spots policing studies should be aimed at achieving both better measurement and more realistic counterfactuals to address the choices police policy makers and police officers face, whether hot spot by hot spot or across systems of hot-spots policing (see, e.g., Sherman et al., 2014).

TABLE 7 Effect size and dosage measures across four hot spots experiments

Experiment	Effect Size (*d*)	*n*	T = Percent Time Patrolled, Treatment (How Measured)	C = Percent Time Patrolled, Control (How Measured)	T:C Ratio
Minneapolis (Sherman & Weisburd, 1995)	–.06	*n* Control Hot Spots = 55; *n* Treatment Hot Spots = 55	14.9% (Observed)	7.5% (Observed)	1.99:1
Sacramento (Mitchell, 2017)	–.10	21 matched pairs	10.3% (GPS)	3.4% (GPS)	3:1
Peterborough (U.K.) (Ariel et al., 2016)	–.21	*n* Control Hot Spots = 38; *n* Treatment Hot Spots = 34	8.9% (GPS)	3.8% (GPS)	2.3:1
London Underground	–.69	*n* Control Hot Spots = 58; *n* Treatment Hot Spots = 57	12% (Admin Records)	0 (Admin Records)	N/A

The second point is that the deterrent effects of police patrol may be far greater when police are *not* present than when they are; we call this second point the "London Underground Paradox," although it was first observed by Koper (1995) in his secondary analysis of the Minneapolis experiment.

The first point is responsive in part to the concerns raised by Nagin and Sampson (2019). Although policing scholars may have failed to think as systemically as we might have done, we have at least stayed within the range of feasibility. Rather than grappling with the larger questions of how much deterrence can be obtained per police officer under different systemic counterfactuals, we have proceeded to test what we could persuade patrol officers to deliver (or not) in two different groups of hot spots. There need be no regret in accepting that conclusion. Had we tried otherwise, there would likely be no body of randomized trial evidence that we could even discuss. Every journey begins with the first step. Randomizing patrol time by location was a crucial first step for this field of science. The next steps are measuring that time precisely and theorizing about all the results combined. Those steps will increase certainty of results by clarifying the nature of counterfactual worlds in two groups of hot spots. Until that is done properly, it seems premature to design a "system" of patrol dosage across hot spots of different sizes, crime levels, physical design, and routine activities—let alone what police could do in different hot spots, such as problem-solving versus patrol. System design cannot be done well until we know much more about the effects of heterogeneity in hot-spots policing dosage and practices across experiments, especially with respect to the levels of dosage in the control group hot spots.

Three decades on from the completion of the first hot-spots policing experiment in 1989 (Sherman & Weisburd, 1995), it is now time to ask how we can do better to generate more deterrence with less intrusive policing. The Nagin and Sampson (2019) commentary usefully suggested that we should emphasize the contrasts between counterfactuals, not (just) research designs per se. Finding a counterfactual world to test causality of crime deterrence within a city is certainly difficult, as they presumed, but it is not impossible. The example of isolated platforms never patrolled before is rare but not unique. It indicates ways we might think differently about designing other experiments. For example,

many police agencies have no idea where their hot spots are (Sutherland & Mueller-Johnson, 2019) and even assign directed patrols to locations leaders *believe* to be "hot" that are not "hot" according to the results of a data analysis (MacBeth & Ariel, 2019; see also Ariel, Bland, & Sutherland, 2017; Norton, Ariel, Weinborn, & O' Dwyer, 2018; Wain, Ariel, & Tankebe, 2017; Weinborn, Ariel, Sherman, & O' Dwyer, 2017). In such agencies, experiments using GPS monitors could be designed to gather baseline measures of where patrol has been delivered and where it has not. A crossover design could be used to assign much less (or zero) patrol to some currently patrolled locations and much more to hot locations that receive very little patrol.

This point converges with Nagin and Sampson (2019) in emphasizing that a research design is not the same thing as the logic of causality. Our point is that by turning the current world upside down in some hot spots but not in others, a clear counterfactual can indeed be created for generalizing conclusions from experiments to other specific hot spots at that level of analysis–even without estimating the benefits of how each world might work at the level of a "system equilibrium" (Nagin & Sampson, 2019). Should a police agency ever fully implement "extra" policing at all hot spots in a jurisdiction—and we are unaware of any fully documented examples—we suggest that the benefits of such a system will depend heavily on what the counterfactual "business as usual" was before the system changed, and even on how long such a system could last. Recent experience in the United Kingdom, for example, indicates that targeted patrol can stop altogether when staffing is cut and demand for response policing increases. System impact questions may thus remain less important, at least in the eyes of policy makers, than questions of short-term effects at specific hot spots.

Our second point is that in the process of thinking broadly about counterfactuals—including no-patrol control groups—we may also begin to measure things we had not previously measured. Most important would be the temporal and spatial components of deterrence. The findings of the LU experiment indicate that most crime prevention by police may occur at places and times when police are not even present—but may have been present recently. That is a major departure from the "security guard" view of police effects that some have drawn from tests like the 24-hour guards at synagogues in Buenos Aires after a terrorist attack in 1994 (DiTella & Schargrodsky, 2004). The 9-month duration of those postings yielded 75 percent reductions in crimes such as car theft but only within two blocks of the stationary presence of the police. Had the officers been able to roam within a four-block square area, however, the effects may have been substantially multiplied. Zones of patrol that embrace the now-demonstrated "regional effect" of patrols in London Underground stations could be tested in a wide variety of settings. How much presence, how often, and with how much intermittency would all be highly testable within cities. That level of granularity would not, however, be feasible for RCTs across samples of cities.

What this level of granularity has revealed is the "London Underground Paradox." By sending officers to patrol each platform for ~12 percent of a 7-hour day, 4 days a week, the result is a large reduction in crime and emergency calls for the time when police are not there: when they are not on the platforms, and when they are not in the mass transit station. The LU paradox is that *police prevented more crime when they were absent than when they were present, face to face* (figure 4). Only 6 percent of the crime reduction effect was estimated to have occurred during hours in which actually (and intermittently) patrolled; the other 94 percent of the reduction was computed by comparing crime in times in which police did no patrols. This paradox could be highly relevant for debates about cutting police funding, if nothing else. It may also have implications for police priorities for patrol versus other tasks, given limited resources, such as the benefit from investigating past crimes compared with the benefit of preventing future crimes.

The results of the LU experiment do not indicate that criminology has oversold the deterrent benefits of hot-spots policing. Rather, criminology may have underestimated the residual and regional effects

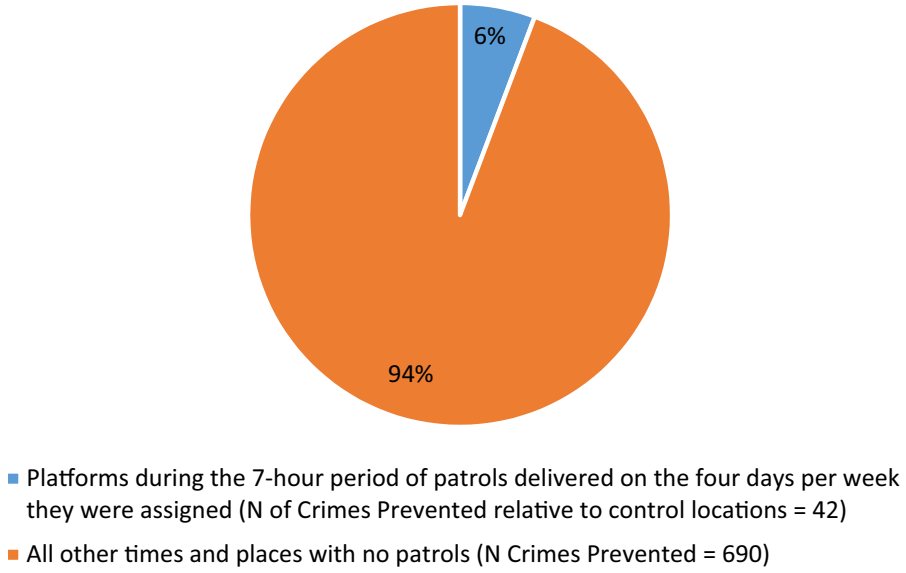

FIGURE 4 Treatment versus control postrandom assignment: Crimes prevented, relative to control sites, during hot-spots patrols and all other times and places measured in and around the hot spots [Color figure can be viewed at wileyonlinelibrary.com]

of such policing. To be sure, criminologists may have paid too little attention to both the small effect sizes of previous tests and conditions of dosage in the counterfactuals. The time may have now come when we should discourage any further research in which the counterfactuals are not fully described. The idea of "black box" experiments has had its uses. That time, however, is past. We need to know exactly what is being compared to make any sense out of the results, in either theoretical or policy terms.

This point applies far beyond hot-spots policing experiments. We cannot accept random assignment of body-worn cameras, for example, to individual officers, knowing that both T and C will be present at the same situations (Yokum, Ravishankar, & Coppock, 2019). We can only support police investing in research focused on answering research questions that are based on clearly established counterfactuals, no matter what the subject of the test may be.

8 | LIMITATIONS

The limitations of this experiment are important to consider. First, we were limited in our ability to collect accurate measures of treatment delivery in all experimental and control hot spots. The "pen-and-paper" approach we used is archaic, let alone unsustainable. More important, it may have underestimated police presence on the control group platforms. We doubt that the presence would have been substantial or consistent over 6 months. Nonetheless, much of the validity of our conclusions depends on the control platforms not having any directed patrols. Had a passive, unobtrusive measure like GPS signals been available, the experiment and its conclusions would be stronger. Checking CCTV footage on platforms may provide an alternative measure for future studies given enough funding and time (see discussion in Ariel, 2019).

A further limitation is that the volume of crime was (fortunately) so low. Although our emphasis remains on calls for service as the main outcome of the experiment (table 7), the low frequency of crime

in the London Underground limited the additional analyses that are possible with the data. Therefore, we had inadequate statistical power to examine moderator effects. Had the volume of crime been closer to the Minneapolis experiment at 20 or so per week, there may have been scope for learning whether effect sizes are bigger where crime is more frequent—or the converse. Replications of this study in higher crime cities may benefit from the added power that challenge provides. On the other hand, a zero-patrol baseline is less likely to occur naturally when crime frequency is higher.

Similarly, the residual and regional effects may also have been detectable because the baseline of patrol was so low. Although similar effects may occur with higher levels of baseline patrol, more powerful research designs may be necessary to detect them under those conditions. It would be unwise to generalize from these findings to other low-baseline patrol settings, especially if many days may pass between the presence of a proactive police patrol for even one 15-minute period.

Taking these points more generally, we acknowledge that the external validity of our results may be limited to enclosed underground public spaces on mass transit systems. Additional or unique features on such systems are not found above the ground. Even though we are testing a basic police strategy used by nearly all territorial police agencies—visible police patrols (see more generally in Wain & Ariel, 2014)—there are distinct features in the LU environment. For example, the LU does not experience the same volume nor the types of crime that areas above the ground are exposed to (Ariel, 2011). Similarly, the extent of third-party surveillance (see, e.g., Ariel et al., 2017; Ariel et al., 2019), a transient and large population flow (Currie, Jones, & Woolley, 2014), and other aspects of the LU are different from those an officer might deal with above ground. For example, a common basic patrol unit in many places in England and Wales is delivered on foot (see Ariel et al., 2016), but policing on the ground is often conducted in vehicles (Wain & Ariel 2014; see also Simpson, 2019). More reverse-knockout hot-spots experiments are needed.

Finally, we could have drawn great benefit from an analysis of the kinds of offenders who commit crimes on LU platforms and elsewhere in the stations—enhanced if possible by interviews of such people. Professional pickpockets and distraction thieves may be particularly sensitive to changes in police presence. If a large portion of the criminality was a result of them, and they were the main audience for the deterrent effects, then the results may not be generalizable to mass transit settings with less professional kinds of offenders. A better study design would have included interviewing hundreds of offenders at point of arrest to ask them what they thought, or had been told, about the different levels across stations or platforms in the risks of getting caught.

9 | CONCLUSION

In what seems to be the first reverse-knockout, no-treatment control group experiment in hot-spots policing, we tested both direct and residual local and regional deterrence effects of BTP police patrols in the London Underground. The findings indicate that platform patrols cause large reductions in both crime- and citizen-generated calls for service in these areas that had never been proactively patrolled by uniformed police. These effects were so long lasting and so spatially diffusive that the actual patrol time had geometric effects on crime. We call these effects the London Underground Paradox: The total crime prevention benefit of police patrols is greater when they are absent than when they are present.

The results of our analyses in which the effects *during* treatment delivery times from *no-treatment periods* were separated indicate a more complex nature of deterrence theory than that previously assumed. Local deterrence effects matter, but they may be only a means to a greater end: optimizing the residual deterrent effect of police patrol (Sherman, 1990) where crime is most likely to occur, which occurs by definition when police are not there. The more that uniformed police have been there,

and the more recently, the less likely the future crimes may be to occur. For every hour police spend in cars driving to answer nonemergency calls, we can now see that investment in reactive policing as a choice, not a duty. If the question is whether proactive patrols do the most good where the most harm is likely to occur, communities might finally move to reallocate preventive patrols to locations where we have documented their optimal effects, as long recommended by the National Research Council (Skogan & Frydl, 2004).

REFERENCES

Ariel, B. (2011). *Hot dots and hot lines: Analysis of crime in the London Underground*. Presented at the annual meeting of the American Society of Criminology, Washington, DC, November 6, 2011).

Ariel, B. (2019). Technology in policing. In D. L. Weisburd & A. A. Braga (Eds.), *Innovations in policing: Contrasting perspectives* (2nd ed., pp. 521–516). Cambridge, England: Cambridge University Press.

Ariel, B., Bland, M., & Sutherland, A. (2017). "Lowering the threshold of effective deterrence"—Testing the effect of private security agents in public spaces on crime: A randomized controlled trial in a mass transit system. *PloS One*, *12*(12), e0187392.

Ariel, B., Newton, M., McEwan, L., Ashbridge, G. A., Weinborn, C., & Brants, H. S. (2019). Reducing assaults against staff using body-worn cameras (BWCs) in railway stations. *Criminal Justice Review*, *44*(1), 76–93.

Ariel, B., & Partridge, H. (2017). Predictable policing: Measuring the crime control benefits of hotspots policing at bus stops. *Journal of Quantitative Criminology*, *33*(4), 809–833.

Ariel, B., Sutherland, A., & Sherman, L. W. (2018). Preventing treatment spillover contamination in criminological field experiments: The case of body-worn police cameras. *Journal of Experimental Criminology*, 1–23.

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.

Austin, C. P., Battey, J. F., Bradley, A., Bucan, M., Capecchi, M., Collins, F. S., … Grieder, F. B. (2004). The knockout mouse project. *Nature Genetics*, *36*(9), 921.

Becker, L. A. (2000). Analysis of pretest and posttest scores with gain scores and repeated measures. *FrontPage Workshop*. Retrieved from http://www.uccs.edu/lbecker/gainscore.html

Braga, A., Papachristos, A., & Hureau, D. (2012). Hot spots policing effects on crime. *Campbell Systematic Reviews*, *8*(8), 1–96.

Braga, A. A., Papachristos, A. V., & Hureau, D. M. (2014). The effects of hot spots policing on crime: An updated systematic review and meta-analysis. *Justice Quarterly*, *31*(4), 633–663.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Boston: Houghton, Mifflin.

Cheung, S. F., & Chan, D. K. S. (2004). Dependent effect sizes in meta-analysis: Incorporating the degree of interdependence. *Journal of Applied Psychology*, *89*(5), 780.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences* (3rd ed.). London, England: Routledge.

Cohen, L. E., & Felson, M. (1979). On estimating the social costs of national economic policy: A critical examination of the Brenner study. *Social Indicators Research*, *6*(2), 251–259.

Currie, G., Jones, A., & Woolley, J. (2014). Travel demand management and the big scare: Impacts and lessons on travel in London during the 2012 summer Olympic Games. *Transportation Research Record*, *2469*(1), 11–22.

De Brito, C., & Ariel, B. (2017). Does tracking and feedback boost patrol time in hot spots? Two tests. *Cambridge Journal of Evidence Based Policing*, *1*(4), 244–226.

Di Tella, R., & Schargrodsky, E. (2004). Do police reduce crime? Estimates using the allocation of police forces after a terrorist attack. *American Economic Review*, *94*(1), 115–133.

Drover, P. (2015). Leading an experiment in police body-worn video cameras. *International Criminal Justice Review*, *25*(1), 80–97.

Flack, J. C., de Waal, F. B., & Krakauer, D. C. (2005). Social structure, robustness, and policing cost in a cognitively sophisticated species. *The American Naturalist*, *165*(5), E126–E139.

Flack, J. C., Girvan, M., de Waal, F. B., & Krakauer, D. C. (2006). Policing stabilizes construction of social niches in primates. *Nature*, *439*(7075), 426.

Gleser, L. J., & Olkin, I. (2007). *Stochastically dependent effect sizes* (technical report 2007-2).

Goldstein, H. (1979). Improving policing: A problem-oriented approach. *Crime & Delinquency*, *25*(2), 236–258.

Groff, E. R., Ratcliffe, J. H., Haberman, C. P., Sorg, E. T., Joyce, N. M., & Taylor, R. B. (2015). Does what police do at hot spots matter? The Philadelphia policing tactics experiment. *Criminology*, *53*(1), 23–53.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research synthesis methods*, *1*(1), 39–65.

Koper, C. S. (1995). Just enough police presence: Reducing crime and disorderly behavior by optimizing patrol time in crime hot spots. *Justice Quarterly*, *12*(4), 649–672.

Macbeth, E., & Ariel, B. (2019). Place-based statistical versus clinical predictions of crime hot spots and harm locations in Northern Ireland. *Justice Quarterly*, *36*(1), 93–126.

Mitchell, Renée, J. (2017). *The Sacramento hot spots policing experiment: An extension and sensitivity analysis*. (Ph.D. a dissertation, University of Cambridge).

Mitchell, R. J. (2019). The usefulness of a crime harm index: analyzing the Sacramento Hot Spot Experiment using the California Crime Harm Index (CA-CHI). *Journal of Experimental Criminology*, *15*(1), 103–113.

Nagin, D. S. (1998). Criminal deterrence research at the outset of the twenty-first century. *Crime and justice*, *23*, 1–42.

Nagin, D. S., & Sampson, R. J. (2019). The real gold standard: Measuring counterfactual worlds that matter most to social science and policy. *Annual Review of Criminology*, *2*, 123–145.

Norton, S., Ariel, B., Weinborn, C., & O'Dwyer, E. (2018). Spatiotemporal patterns and distributions of harm within street segments: The story of the "harmspot". *Policing: An International Journal*, *41*(3), 352–371.

Ratcliffe, J. H. (2016). *Intelligence-led policing*. New York: Routledge.

Ratcliffe, J. H., Grof, E. R., Haberman, C. P., Sorg, E. T., & Joyce, N. (2013). *Philadelphia, Pennsylvania Smart Policing Initiative: Testing the impacts of differential police strategies on violent crime hotspots* (Smart Policing Initiative: Site Spotlight). Washington: Bureau of Justice Assistance.

Ratcliffe, J. H., Taniguchi, T., Groff, E. R., & Wood, J. D. (2011). The Philadelphia foot patrol experiment: A randomized controlled trial of police patrol effectiveness in violent crime hotspots. *Criminology*, *49*(3), 795–831.

Reiss, A. J. (1985). *Policing a city's central district: The Oakland story*. US Department of Justice, National Institute of Justice.

Rosenfeld, R., Deckard, M. J., & Blackburn, E. (2014). The effects of directed patrol and self-initiated enforcement on firearm violence: A randomized controlled study of hot spot policing. *Criminology*, *52*(3), 428–449.

Scammacca, N., Roberts, G., & Stuebing, K. K. (2014). Meta-analysis with complex research designs: Dealing with dependence from multiple measures and multiple group comparisons. *Review of Educational Research*, *84*(3), 328–364.

Senn, S. J. (1989). Covariate imbalance and random allocation in clinical trials. *Statistics in Medicine*, *8*(4), 467–475.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Belmont: Wadsworth Cengage Learning.

Sherman, L. W. (1990). Police crackdowns: Initial and residual deterrence. *Crime and Justice*, *12*, 1–48.

Sherman, L. W., & Weisburd, D. (1995). General deterrent effects of police patrol in crime 'hot spots': A randomized, controlled trial. *Justice Quarterly*, *12*(4), 625–648.

Sherman, L. W., Williams, S., Ariel, B., Strang, L. R., Wain, N., Slothower, M., & Norton, A. (2014). An integrated theory of hot spots patrol strategy: Implementing prevention by scaling up and feeding back. *Journal of Contemporary Criminal Justice*, *30*(2), 95–122.

Simpson, R. (2019). Police vehicles as symbols of legitimacy. *Journal of Experimental Criminology*, *15*(1), 87–101.

Skogan, W. G., & Frydl, K. (2004). *Fairness and Effectiveness in Policing: The Evidence*. Washington: National Academies Press.

Skogan, W. G., & Hartnett, S. M. (1997). *Community Policing, Chicago Style* (p. 13). Oxford University Press.

Sorg, E. T., Wood, J. D., Groff, E. R., & Ratcliffe, J. H. (2014). Boundary adherence during place-based policing evaluations: A research note. *Journal of Research in Crime and Delinquency*, *51*(3), 377–393.

Spybrook, J., Bloom, H., Congdon, R., Hill, C., Liu, X., Martinez, A., & Raudenbush, S. (2013). *Optimal design plus empirical evidence* (Version 3.01). [Computer software]. Retrieved from http://hlmsoft.net/od

Sutherland, J., & Mueller-Johnson, K. (2019). Evidence vs. professional judgment in ranking "power few" crime targets: A comparative analysis. *Cambridge Journal of Evidence-Based Policing*, *3*(1–2), 54–72.

Taylor, B., Koper, C. S., & Woods, D. J. (2011). A randomized controlled trial of different policing strategies at hot spots of violent crime. *Journal of Experimental Criminology*, *7*(2), 149–181.

Telep, C. W., Mitchell, R. J., & Weisburd, D. (2014). How much time should the police spend at crime hot spots? Answers from a police agency directed randomized field trial in Sacramento, California. *Justice Quarterly*, *31*(5), 905–933.

Transport for London. (2015). Annual Report and Statement of Accounts. Transport For London: Mayor of London. Retrieved from http://content.tfl.gov.uk/tfl-annual-report-2015-16.pdf

van Breukelen, G. J. (2013). ANCOVA versus CHANGE from baseline in nonrandomized studies: The difference. *Multivariate Behavioral Research*, *48*(6), 895–922.

Vickers, A. J., & Altman, D. G. (2001). Analysing controlled trials with baseline and follow up measurements. *British Medical Journal*, *323*(7321), 1123–1124.

Wain, N., & Ariel, B. (2014). Tracking of police patrol. *Policing: A Journal of Policy and Practice*, *8*(3), 274–283.

Wain, N., Ariel, B., & Tankebe, J. (2017). The collateral consequences of GPS-LED supervision in hot spots policing. *Police Practice and Research*, *18*(4), 376–390.

Weinborn, C., Ariel, B., Sherman, L. W., & O' Dwyer, E. (2017). Hotspots vs. harmspots: Shifting the focus from counts to harm in the criminology of place. *Applied geography*, *86*, 226–244.

Weisburd, D., Wyckoff, L. A., Ready, J., Eck, J. E., Hinkle, J. C., & Gajewski, F. (2006). Does crime just move around the corner? A controlled study of spatial displacement and diffusion of crime control benefits. *Criminology*, *44*(3), 549–592.

Weisburd, D. L., & Green, L. (1995). Policing drug hot spots: The Jersey City drug market analysis experiment. *Justice Quarterly*, *12*(4), 711–735.

Williams, S., & Coupe, T. (2017). Frequency vs. length of hot spots patrols: A randomised controlled trial. *Cambridge Journal of Evidence-Based Policing*, *1*(1), 5–21.

Yokum, D., Ravishankar, A., & Coppock, A. (2019). A randomized control trial evaluating the effects of police body-worn cameras. *Proceedings of the National Academy of Sciences*, *116*(21), 10329–10332.

AUTHOR BIOGRAPHIES

Barak Ariel is professor of criminology at the Hebrew University of Jerusalem and a lecturer in experimental criminology at the University of Cambridge, where he is the chief analyst of the Jerry Lee Centre of Experimental Criminology in the Institute of Criminology. His research interests include deterrence theory and social control, compliance, evidence-based policy, experimental criminology, and policing.

Lawrence W. Sherman is Wolfson Professor of Criminology Emeritus at the University of Cambridge and director of the Cambridge Centre for Evidence-Based Policing. He is also a distinguished professor in the University of Maryland—College Park's Department of Criminology & Criminal Justice. His research interests include residual deterrence, hot spots of predatory crime, evidence-based policing, experimental criminology, defiance theory, and experimental designs.

Mark Newton is a former assistant chief constable in the British Transport Police and the head of railway policing and security with the Rail Delivery Group.

How to cite this article: Ariel B, Sherman LW, Newton M. Testing hot-spots police patrols against no-treatment controls: Temporal and spatial deterrence effects in the London Underground experiment. *Criminology*. 2020;58:101–128. https://doi.org/10.1111/1745-9125.12231