Cambridge Journal of Evidence-Based Policing (2022) 6:25–41
https://doi.org/10.1007/s41887-022-00073-y

SCIENTIFIC COMMUNICATION

"Test-As-You-Go" for Hot Spots Policing: Continuous Impact Assessment with Repeat Crossover Designs

Lawrence W. Sherman[1]

Accepted: 4 May 2022 / Published online: 13 May 2022

Abstract

Hot spots policing is rapidly changing its evidence-base. Instead of producing more results of one-off, conventional experiments that provide an evidence-base *across* police agencies (Braga et al., (*Campbell Systematic Reviews, 15*(3), 2019), hot spots policing (HSP) experiments in the UK are now providing continuous impact assessment (CIA) *within* police agencies, and within each hot spot. This new mission for experiments entails a change in research design, from the conventionally *fixed* assignment of each hot spot to a single treatment (in what is technically called a "parallel-track" design) to *alternating* treatments for each hot spot (in what is known as a "repeat crossover" design). Crossover design experiments are designed for an "on"-days-versus- "off"-days, "test-as-you-go" operating model, using test results in each hot spot for immediate operational feedback to improve performance. This feature can empower police supervisors with compelling evidence for officers about their impact on crime in recent weeks.

This approach has great promise, but it also has a great threat. The promise is for integrating evidence more tightly with daily police operations. The threat is that crossover tests may underestimate the true benefits of HSP because they ignore the lingering, "residual deterrence" effects of HSP "on" days continuing into "off" days without HSP. That "carryover" effect of HSP in prior days may take up to 4 days to "wash out" (Barnes et al., 2020). Until it does, crime during HSP "off" days may be lower than if there had been no recent HSP "on" days, thus making HSP look less effective than it truly is.

This problem is purely a matter of what analysts do, rather than what police officers do. As long as the officers deliver on their assigned tasks of which hot spots to patrol when, both research designs can have validity. But the problem of how to analyze the effects of these patrols is up to the analysts to solve. If the analysts handle the problem by deleting a certain number of days in between treatment changes for each hot spot—technically known as a "washout" period—they will provide valid impact assessments of HSP. More important, they can do that with a design that requires no long-term denial of service to large numbers of hot spots assigned to a control group, as in traditional random assignment to parallel tracks of treatment vs. control (e.g., Sherman & Weisburd, (*Justice Quarterly*, 625–648, 1995)).

Repeat crossover trials are therefore an excellent improvement over parallel-track trials, *subject to omitting crime measures from the washout period for eliminating*

Extended author information available on the last page of the article

carryover effects during crossover periods from one treatment condition to another. The following discussion shows how analysts and police leaders can use and implement crossover designs with high internal validity, without biased measures of crime on control days.

Keywords Test-as-you-go · Hot spots · Continuous impact assessment (CIA) · Carryover effects · Residual deterrence · Cumulative deterrence · Washout period

Introduction

For over three decades, hot spots policing (HSP) has been extensively tested by using "parallel track" comparisons between two (or more) groups of hot spots over long periods of time (90 to 365 days). The crime totals in hot spots receiving consistent HSP are compared to totals in similar hot spots not receiving HSP (Braga et al, 2019; Sherman & Weisburd, 1995).

In recent years, however, the parallel track trials have often been replaced by "repeat crossover" designs of HSP evaluations—especially in the UK. In this design, each hot spot serves as its own control. Using each day in each hot spot as the unit of analysis (hot spot-days), each hot spot is randomly assigned to different treatments on different days. Crime outcomes on treatment days, on average, in each hot spot are then compared to average outcomes on no-treatment days, within each hot spot.

The repeat crossover design opens the door for police to practice evidence-based policing by "testing-as-you-go" for continuous impact assessment (CIA). Unlike parallel track designs, repeat crossover designs also have a political advantage of giving every hot spot frequent attention, rather than denying hot spots policing to some locations for months (or a year) at a time.

With appropriate use of a *pause in measurement* between treatment crossovers, called a "washout" period for carryover effects of the last treatment to be "washed away," police agencies can now introduce the "test-as-you-go" strategy of using experiments as an ongoing operating strategy. Using crossover testing as the means of continuous impact assessment, every hot spot can get extra patrols—just not on every day. By using each hot spot as its own control, the repeat crossover design can gain both statistical power and complete coverage of all crime hot spots. Several such studies have already found significant reductions in crime and violence on the extra patrol days compared to no-patrol days (Barnes et al, 2020; Basford et al., 2021; Bland et al., 2021). Similar designs could also be used for testing particular tactics in hot spots, such as traffic enforcement or stop-search.

The primary purpose of test-as-you-go is not to produce published studies for the accumulation of *global* knowledge about HSP (e.g., Braga et al, 2019); it is to prevent as much crime as possible, on a continuous basis, with *local* knowledge about the cumulative and most recent outcomes of the effectiveness of HSP in each specific hot spot. Tracking outcomes this way may lead to modifications in tactics or resources that improve subsequent test results and reduce crime.

The test-as-you-go strategy integrates the "Triple-T" by repeat crossover testing in tracking outcomes for new targeting. It uses random assignment by days as a permanent operating model, instead of the "test-once-and-stop" pattern of parallel track designs. Its key challenge is to make ongoing testing as valid as the "test-once-and-stop" designs.

The main issue is the residual or "carryover" effects of the last treatment (Barnes, et al, 2020; Koper, 1995; Sherman, 1990). It is therefore essential that the design of any continuous impact assessment (CIA) plan identifies the number of "washout days" needed between treatment categories, so that carryover effects can be "washed out." Absent a break in measurement for a period of washout days without that HSP treatment, testing-as-you-go risks under-estimating the benefits of hot spot patrols.

That risk is growing as the spread of HSP increases. Over three decades since it was first tested in a parallel track randomized controlled trial (Sherman & Weisburd, 1995), "Hot Spots Policing" (HSP) may now be the most widely researched and adopted strategy of evidence-based policing (Sherman, 1998, 2013). With over eighty rigorous evaluations showing consistent benefits of HSP, no other policing strategy can offer more independent assessments (Braga et al, 2019; see also Barnes et al., 2020; Basford et al., 2021; Bland et al, 2021; Weisburd, et al, 2022). Hailing HSP as the crime reduction strategy with strongest evidence, one UK policing minister (Malthouse, 2021) has offered special funding for police agencies to implement it. In 2021–2022, there were 18 (of 43) police agencies using such funding, of which 13 established a repeat crossover design for continuous impact assessment (Rose, 2022).

Risks of Disappointment

While hot spots experiments have tested the effects of HSP in targeted hot spots, the precision of that aim can easily be confused with a general reduction in crime across a police force area. The two aims are not the same. One does not require the other to demonstrate effectiveness. If for no other reason, external forces (such as economics or population changes) could be driving crime up across a city, even while HSP is preventing crime from getting even worse city-wide. Yet, it is difficult to prove that HSP "works" across an entire police force when there is no comparison group to that force. This limitation makes HSP vulnerable to its critics.

Even leading criminologists (e.g., Nagin & Sampson, 2019) have argued that local crime reduction does not matter if a citywide benefit cannot be proven. While at least one quasi-experimental study of an entire city has shown a 7-year city-wide benefit (41% reduction in violence) of "system-level" HSP (Koper et al., 2021), the hot spots strategy is unlikely to have many other city-wide studies of long-term effects any time soon.

The risk of disappointing before/after results with HSP is even greater if the strategy is implemented with low levels of compliance or is abandoned after officer resistance emerged without adequate training and supervision (O'Connor, 2022). The main threat from this risk is that it does not differentiate between HSP working in some hot spots but not in others. By generalizing about HSP based on force-wide

crime trends, disappointment could cripple HSP even before a police organization can begin to develop skill at the new strategy. In sum, the risk comes from putting all your eggs in one basket—the overall, average effect of HSP on all hot spots targeted—rather than considering HSP impact for each hot spot, one at a time.

The Promises of Test-As-You-Go

The risk of disappointment can be reduced by any method that allows hot spots to be examined individually—just as doctors treat patients individually, in light of each one's individual circumstances. Much of the practice of medicine follows a "test-as-you-go" principle in which doctors first try one treatment (based in part on results of randomized trials), then switch to other treatments if the first choice did not improve the patient's condition. This trial-and-error strategy is *individualized* at the patient level, even while it is *informed* by results from studies involving thousands of patients. Some patients may even comply more with some kinds of treatments (like taking pills) than others (like increasing exercise)—just as police compliance with HSP tasks may also be a major factor in whether that treatment works. By individualizing high-crime places in which crime does not respond to HSP, police leaders do not have to fix the entire strategic system. All they have to do is look at the facts for any one hot spot, and modify the specific tactics at that location.

A further promise of test-as-you-go is to accommodate better the vast spread of crime harm and volume from the highest-ranked to lowest-ranked hot spots. Even if 100 locations out of 10,000 in a city have half of all serious violence, the top-ranked location (#1) could have 20 times as much crime as the bottom-ranked location (#100). By customizing resources and tactics for each hot spot based on its crime frequency and harm, the test-as-you-go method can minimize the over-dosing of lower-level hot spots and under-dosing of the higher-harm hot spots. While a parallel-track design demands *consistency* of patrol time across inconsistently hot spots (Sherman & Weisburd, 1995), a repeated crossover design allows *right-sizing* of patrol time relative to each hot spots crime intensity.

Each Hot Spot Is Its Own Control

Just as each patient is their own "control" for a sequential series of treatments, each hot spot can be its own control for an ongoing comparison of two different treatments. There is a slight difference: doctors could choose to keep using the first treatment that seems to work *adequately* for each patient, while a police unit can continuously compare a single hot spot alternating two or more police tactics to see which one works *optimally* over time. That difference further strengthens the promise of a test-as-you-go policy.

But how can each hot spot serve as its own comparison? As explained below, test-as-you-go uses "repeat crossover" randomized trials to deploy added patrol to each hot spot so that it can serve as its own control. Simply by randomly assigning each day to a different treatment condition, such as being patrolled for 15 minutes on

some days and not on others, the design can reveal the effects of that added patrol in that hot spot.

The "repeat crossover" design has been used in several studies published in the *Cambridge Journal of Evidence-Based Policing*, starting in 2017 with the first crossover experiment in hot spots policing (Williams & Coupe, 2017) ever published (to my knowledge). Since then, Basford et al. (2021) and Bland et al. (2021) have both published successful tests of HSP by using the repeat crossover design, both of which were influenced by Barnes et al. (2020). While none of these studies has examined differential effects of HSP on different individual hot spots, all of them used a research design that has the potential to do so.

Understanding that potential requires a re-examination of the two principal designs in field experiments: parallel track vs. repeat crossover. While statisticians have shown for almost a century that these are far from the only designs possible (Cox, 1958; Fisher, 1935), they have been the most frequently deployed strategies in evidence-based policy development in medicine, education, and other operational fields. Not every research question offers a choice between the two designs, as noted below. Hot spots policing, however, can clearly be tested with either design—as long as each experiment maintains internal validity.

Two Kinds of Randomized Controlled Trials

Parallel Track Designs

A parallel track experiment is one in which random assignment occurs only once, with all units remaining in the same treatment group for the entire experiment, sometimes for many years or decades.

From the earliest experiments in criminology, parallel track designs have been the dominant research design. In the Cambridge-Somerville experiment, launched in 1939 (Powers and Witmer, 1951), two large groups of at-risk youth were randomly assigned to either a complex long-term treatment plan or no treatment, then followed up for 30 years (McCord, 1978). In the Manhattan Bail Project, hundreds of arrestees without cash to post a bond were randomly assigned to a release-on-recognizance program or not (Ares et al., 1963). In the Milwaukee domestic violence arrest experiment, 1200 arrestees were randomly assigned to be either arrested or warned for common assault (Sherman et al., 1991) and then followed up for 23 years (Sherman and Harris, 2013, 2015) .

All three studies described above were *parallel track* experiments involving people (not places). Once people were assigned to a treatment, they were intended to have no change in intervention. The logic of "staying put" in one treatment group is to reveal the long-term effect of that single decision on the rest of their lives. Adding any further treatment, in contrast, would have contaminated the test of the initial treatment. While the bail experiment had only a short-term research question, the effect of getting out of jail was to reduce further imprisonment, which also had a lifelong effect. In the case of the arrests for domestic abuse, the

arrested offenders were significantly more likely to have been murdered 23 years later, and their victims were more likely to have died from other causes.

Not all randomized experiments in criminology, of course, use people as the unit of analysis. Many of them now use places as the unit of analysis. Places appear more variable than people by having differences in populations present across different times of day or days of the week. Unlike people, places may not have much memory of what police have done with them in the recent past. Even recent police actions at a place may not necessarily affect the outcomes of what people do there in the future. It is on that basis—that effects of prior actions in a place may "wash out" or disappear in a short time period—that police have begun to use repeat crossover designs.

Repeat Crossover Designs

The crossover approach is widespread enough to have a Wikipedia page, which defines it as "a longitudinal study in which subjects receive a sequence of different treatments (or exposures)" https://en.wikipedia.org/wiki/Crossover_study. The purpose of sequencing different treatments can vary widely. The major opportunity the design offers policing is the chance to use each subject (or place) as its own control, which means that everything else about the subject place is perfectly matched. The aim of a controlled experiment is to match subjects as perfectly as possible—in order to "control" any pre-existing differences that might cause different outcomes (called "bias"). In parallel track designs, this matching can only be accomplished at the group level, so that (for example) two comparable groups of delinquents are both 75% male, or that two groups of hot spots have between 20 and 30 robberies per year. Individual hot spots would still vary, and the effect on any one hot spot would not be measureable.

In theory, measuring crime prevention would be much more precise if it could have exactly the same characteristics at the level of each case subject, not just the average for a large group. In practice, that is exactly what a crossover design can achieve.

Figure 1 depicts the logic of a single crossover design for comparing two drugs on the same people with the same chronic condition (such as asthma), in which one group starts with drug A and—after a "washout" period—changes to drug B, while the second group starts with drug B and—after a "washout" period—changes to drug A. The washout period is important because it creates a gap in measurement of patients' outcome conditions (e.g., shortness of breath). It allows each drug's effect to be measured with clarity of its effect because any traces of the last drug have "washed out" of the patient's body before a new drug is taken—and before the effects of the new drug are measured.

By measuring all of the people in both groups only after they have had prior drugs washed out, their individual scores for the benefits of each drug can be compared as follows:

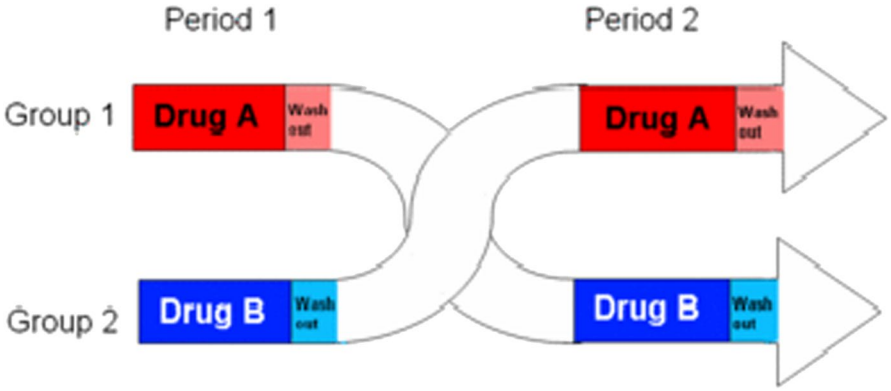

Fig. 1 A crossover design for comparing two drugs on the same people (including washout period) produced for Citizendium (2010)

- The average difference in scores for drug A and drug B for each member of group 1 can be computed, just as it can for group 2, preferably using a standardized mean difference.
- If the average difference for each person is different in the two groups, that should raise questions about whether the conditions were different (or whether the two groups themselves differed substantially after random assignment), then it may be better to restart the experiment.
- If the average difference for each person is similar in the two groups, then the average individual difference between the two drugs can be combined across the two groups.
- The last step is important, because it increases the statistical power of the test, with each patient comprising a separate experiment. That means that the sample size can be smaller than in a parallel track design, and the study can be conducted with less expense and perhaps less time.

Examples of Each Design in Hot Spots Policing

The first experiment in hot spots policing (Sherman & Weisburd, 1995) was a parallel track design for 110 hot spots. As Fig. 2 shows, the monthly difference between the 55 hot spots receiving (by random assignment) double the patrol time of the other 55 hot spots is displayed in the pathway of two lines (solid and dotted), each representing the average crime for each group in each month (relative to crime in the same hot spots in the same month as the year before). The lines show that the difference favored the "extra patrol" group from December through July, after which the direction reversed; another chart in the study (not displayed here) showed that compliance in delivering the extra patrols collapsed in July as well. The two findings both showed the value of the extra patrols while they lasted, as well as the effect of its loss.

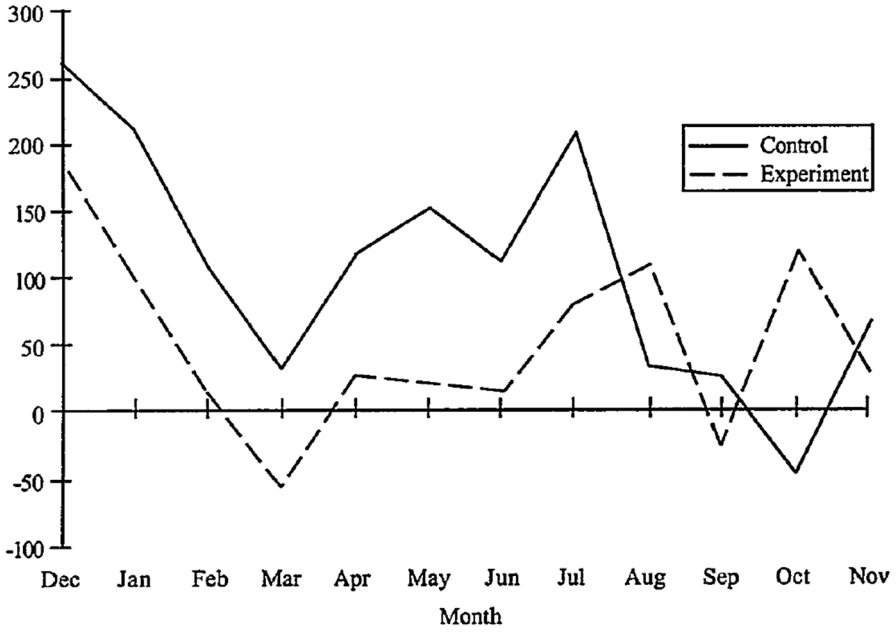

Fig. 2 Absolute differences from baseline to experimental year in total crime calls by month and treatment group

In contrast, Bland et al. (2021) applied a repeat crossover design with just 21 hot spots for just 90 days to test a similar hypothesis, but with just 15 min of patrol on scheduled patrol days and no patrol at all on other days. That design created multiple categories of days in which there was a 15-minute patrol assigned and those in which there were not, as shown in Table 1 below. The categories designate the numbers of consecutive hot spot days of patrol days that preceded each measured patrol day, and the number of consecutive hot spot days with *no patrol* that preceded each measured day of no patrol. Because the number of days in each category was not equal, the sample sizes varied in each category. But the days on which patrols were provided to each hot spot were randomly assigned to be unpredictable, which required varying numbers of consecutive prior patrols or no patrols.

Table 1 Repeat days of patrol and no patrol in Bland et al. (2021) (Calculated from Table 6, Bland et al. (2021))

Days of	Patrol	No patrol
1st day	427	434
2nd day	138	286
3rd day	49	183
4th day	15	118
5th day	1	80
Total	630	1260

Table 1 extracts the key information from Bland et al. (2021).

Taking Table 1 into account, Fig. 3 reveals the important discovery of what Bland et al. call "cumulative deterrence." The more prior consecutive days of patrol, the greater the preventive effect on crime. Even more important, the first day of patrol in this experiment had no reduction in crime. It was only after the second day of patrol that an effect was found. Had the experiment been designed as an every-other-day test, it would not have shown any difference in violence between days with and without patrols. We can therefore decide to classify the first day in Fig. 3 to be a "washout" day after a period of no patrol effects, a carryover of the no-patrol period prior to initial deterrence being realized (Sherman, 1990). The deterrent effect of patrol "on" days was only made discernible by random assignment of multiple days of patrol, creating what the authors call "cumulative deterrence."

Advantages and Disadvantages of the Two Designs

Seen solely from the standpoint of an externally valid research design, we must remember that both parallel-track and repeat crossover designs have advantages and disadvantages. Some of these vary by whether we experiment with places or individual people. What follows below is intended to apply only to experimenting with places, not individuals.

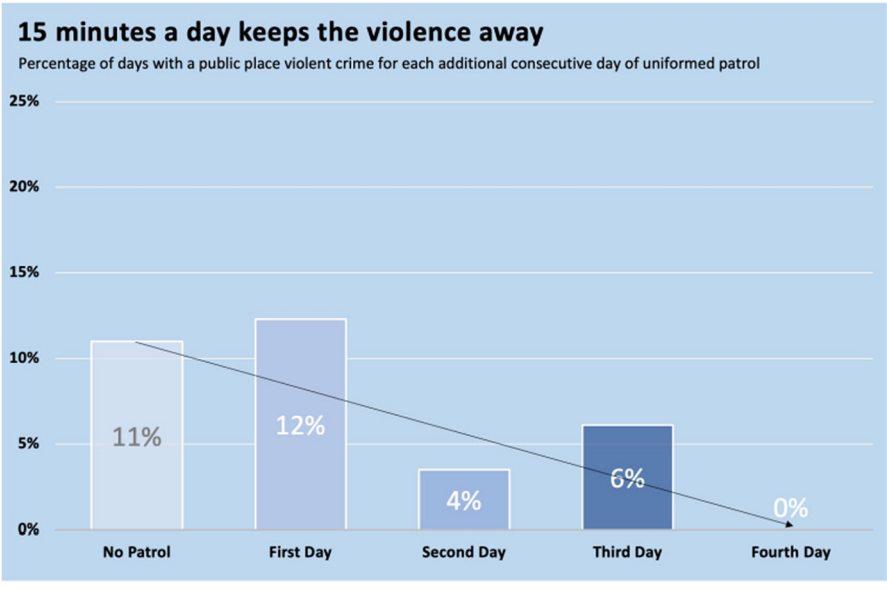

Fig. 3 Cumulative deterrence from repeated days of patrol (Bland & Leggetter, 2021)

Advantages of Parallel Track Designs for Places

When a large sample of high-crime places is split in half, what a test can show is the average effect of about the same level of patrol in all of the treatment group, compared to the average crime measures in all of the control group. The creation of an average effect makes the *results more generalizable*: it increases their "external validity." What you learn from a test in Minneapolis in 1995 may also apply in Minneapolis, or even Manchester, in 2005. That is why parallel track designs are more attractive as the basis for global knowledge, since average effects embrace a wide range of variation from one hot spot to another. This advantage is especially powerful when the differences are clear, large and certain (i.e., statistically significant). Yet that is not always the case. It is especially problematic when the sample size is small.

Disadvantages of Parallel Designs: the Kansas City Patrol Experiment

When the Kansas City (Missouri) Police conducted the first modern experiment in patrol, they did not design it around hot spots—for the very good reason that hot spot crime concentrations had not yet been discovered in 1971, when the experiment was planned. Instead, the experimenters divided 15 patrol beats into three groups of frequency of patrol visits (high, medium, and low). With only five units in each condition for a full year, the analysis was limited to a total number of 15 cases. Statistically speaking, 15 is a very small number with which to draw conclusions about different anything, let alone the number of *minutes* a patrol car is driving by.

The disadvantages of the Kansas City design (Kelling et al., 1974) were as follows.

Power

The design had very low statistical power, with the test unable to detect as statistically significant even large differences in robbery rates (which were higher where patrol frequency was reduced).

Speed

The experiment took an entire year, but that only increased the number of crimes in each of the 15 beats—and not the total number of "cases" in the analysis.

In an important critique, Feinberg et al. (1976) suggested that the experiment would have given patrol a better chance to show its effects if the Kansas City design had been a repeat crossover design, rather than a parallel-track experiment. The advantages for that 15-beat experiment are similar to those that can be found in micro-place hot spot designs of similar size. Had each of the 15 beats been randomly assigned to different patrol levels on each of 365 days—at the same cost

for the experiment—the number of cases for analysis would have jumped from 15 to 5475 (15 beats × 365 days = 5475 beat-days)!

Correctability

Because there are so many units in a large parallel-track randomized trial, the resources needed to correct non-compliance (i.e., patrol dosage as assigned) are far greater than with a smaller sample. Both designs require constant checking. The crossover design requires that fewer places be checked.

Advantages of Repeat Crossover Designs

The disadvantages of the Kansas City design are the flip side of the advantages of the Bedford and Essex and Western Australian designs, each of which randomly assigned a different treatment condition every day.

Power

By using each day in each place (hot spots or beats or even districts) for experimental analysis, the power of the test is greatly increased. The power level benefits by having (nearly) identical place characteristics across all 365 days of a year. But while all those variables are held constant as the backdrop, the slight differences in crime or police presence are able to stand out more clearly.

Speed

The increase in power then allows the police force to shrink the cost and time needed for the experiment. Instead of 12 months, 3 months may do—and did do, in several experiments with significant results.

Correctability

Given the smaller number of hot spots targeted in repeat crossover designs, it is cheaper and easier to check up daily on patrol dosage delivery compliance. Other important issues, including procedural justice, local police legitimacy, and tactics such as stop and search can also be monitored more closely. If correction is required, it is not spread so widely across a much larger number of places.

Key Disadvantage of Repeat Crossover Design: the "Carryover" Problem

The main disadvantage of a crossover design, as noted above, is the period of transition between one treatment condition and the next. This period poses a problem of "carryover" of the persisting effect of the previous treatment even after it has stopped. This includes the concept of residual deterrence (Koper, 1995; Sherman, 1990) in which deterrence persists even after police are no longer present. It also

includes the concept of deterrence decay, which is the longer term effect of increasing crime in the absence of previous levels of visible police presence (see Fig. 3 above, in which crime does not drop on the first day of patrol after no patrol).

The carryover problem is a general issue in medical and other experiments, in which cessation of a treatment cannot guarantee cessation of its effects. Even in physics, pedaling a bicycle creates motion effects that persist after you stop pedaling. Those effects can be stopped by putting on the brakes on the bicycle. But there is no obvious parallel for putting the "brakes" on the effects of policing.

Across different kinds of experiments, the key question for carryover effects is how long they last. The answer appears to be that the length of any carryover effect depends upon the nature of the treatment that has just ceased. If the average period of a carryover of the last treatment can be estimated from empirical observations, then the problem is solved. But that is not so easy.

Crossover Designs: the "Washout Days" Solution

In the experimental literature, there is a wide range of time periods designated for carryover effects to end. In a test of eating walnuts to control diabetes, the design called for 1 week of eating walnuts as the treatment before crossover. This test period was followed by a pause in daily measurement of diabetes effects, in order to allow 1 month of eating no walnuts before the effects were declared washed out and a different diet could be assigned (Farr et al, 2017). In a pre-COVID test of telling employees to stay home unless they had to come into the office for a meeting the washout period between that policy and coming to work daily was only 2 days of the weekend between crossover treatments (E.L. Sherman, 2020). Absent prior research, neither study enjoyed an empirical basis for deciding the length of the wash-out. Yet hot spots policing has substantial research on which to estimate the length of a wash-out period.

Choosing the Washout Period

The solution to estimating a washout period for hot spots patrol days is to look first to local data; if that does not work, look second to global data, from studies in other communities.

Estimating Residual Deterrence

In police forces like Bedfordshire that have conducted several hot spots experiments, it may be possible to estimate the average period for a full washout of residual deterrence (Bland et al., 2021). That would be a study in itself. Yet given the distribution of consecutive days with and without patrol, the raw material for the estimate is there—if sufficient power can be found.

In Perth, Australia, Barnes and colleagues (2020) found that four days of residual deterrence ensued after patrols stopped, even as deterrence decay began to build—yet

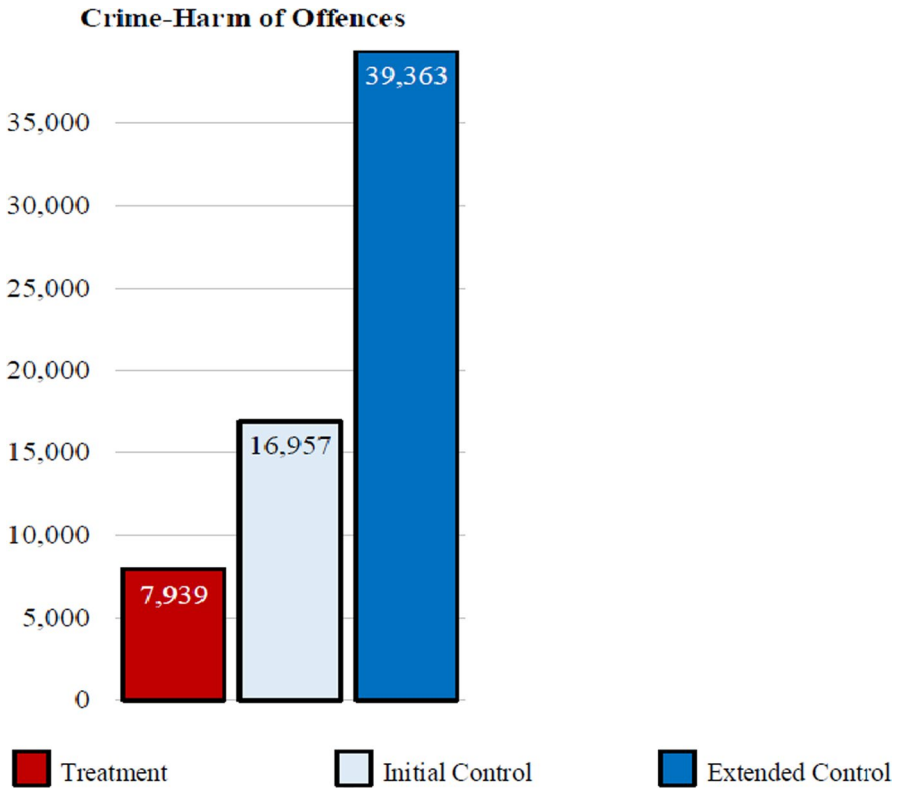

Fig. 4 Treatment vs. initial control period (4 days) vs. extended control period (5-10 days) Crime Harm Index totals, from Barnes et al. (2020)

crime harm exploded on the 5th day. Figure 4 shows that in the first four consecutive days of cessation of patrols, crime harm per day (only) doubled. But in the period from 5 to 10 days with patrol "off," daily crime harm was five times higher than it was during the patrol "on" days. Figure 4 could be read as showing the best test of patrol effects was in the comparison of "treatment" days and "extended control" period days, with the "initial control" period days eliminated from the calculation. Put another way, patrol cut crime harm by 80%, if the comparison is to no patrol at all (after at least 4 days as a washout period). This finding is a strong argument for including a washout period in every HSP experiment—for global knowledge as well as local.

But what about local knowledge? Does test-as-you-go also require the use of a washout period? The answer is clearly "yes."

Custom-Tailored ("Bespoke") Policing for Each Hot Spot

Regardless of how long a washout period may need to be, it seems essential that research designs include such a period for both local and global knowledge—for both test-and-stop as well as test-as-you-go. That is especially true for testing

each hot spot separately, as distinct from seeking average effects across all hot spots. If each hot spot has different lengths of carryover effects—about which we still lack good evidence—then each one may need a different washout period to estimate the amount of crime or harm prevented. On a test-as-you-go basis, each hot spot can be given a different sequence of patrols and no patrols, just as patients can receive different medical interventions in their best interest. Once a washout period is identified with several replications in any particular hot spot, the test pattern for that hot spot can be established for estimating the amount of crime police are preventing (or not).

This logic applies not only to how many minutes police spend in each hot spot. It also applies to other issues, such as changing tactics, or even switching from visible patrols to problem-oriented policing. Finally, there may even be ways to calibrate more patrol for more crime or other adjustments of policing to the particulars of the hot spot.

Designing an Operating Model

In summary, the best strategy for police to measure the impact of hot spots policing for operational purposes is "test-as-you-go." This strategy offers a window for repeat crossover designs to provide individual estimates for each hot spot. Those individual estimates provide an opportunity to rightsize the policing of each hot spot, with no need to make patrol dosage consistent across hot spots. Test-as-you-go is testing for a strategy of right-sizing by trial-and-error and not the effects of a constant dosage level. In that regard alone, it is a means of revolutionizing the way hot spots experiments are undertaken, as well as why.

Choosing Your Measures: Outputs and Outcomes

Other features of HSP testing would remain the same under test-as-you-go. Outputs such as patrol minutes, officer minutes, or arrests would still need to be chosen. Outcomes such as crime counts, crime harm index scores, and calls for service about disorder (antisocial behavior) also must be selected. Holding these metrics constant, even over years of repeat-crossover innovations, would be useful for building local knowledge.

Targeting Successive Days of Treatments

For both global and local knowledge, the value of tracking successive days in each treatment group now seems more important than ever. Alternating patrols in a continuously predictable, "every-otherday-is- on- or- off" design is clearly inadvisable. Such a strategy can be endlessly contaminated by residual deterrence effects. Random assignment of up to 10 days of consecutive patrols or no patrols—as Barnes et al (2020) did—may reveal much more about how to optimize patrol tasking in hot spots. Washout periods seem essential to sort out the carryover issue, one that has not arisen in the context of parallel track designs.

Conclusion

While the evidence from crossover designs is increasing its value for operational policing, the evidence from parallel track designs continues to provide invaluable global knowledge. A recently published 9-month comparison of two parallel tracks of hot spot patrols (in 3 US cities at once), for example, found that a 5-day course of procedural justice training for hot spots patrol officers reduced both arrests and crime, as well as perceptions by local residents that police used excessive force or harassed local citizens (Weisburd, et al, 2022). Only a long-term study using that design could have discovered such important facts.

Yet in order to make the most of the global body of knowledge from long-term parallel track experiments, policing craves local knowledge. The best test of external validity from global research is to conduct similar tests locally, not once, but continuously. Context is always important, but context can change. From an operating perspective, continuous testing is needed for maximally evidence-based policing.

Repeat crossover designs therefore offer great promise, as well as great risks. Unless analysts can solve the problem of estimating the length of washout periods, they run the risk of underestimating the effects of hot spots policing—and stopping the strategy dead in its tracks. But if they can identify washout periods for use with test-as-you-go, they may bring a new meaning to evidence-based practice as a process of continuous impact assessment.

Acknowledgements The author wishes to acknowledge the many years of conversation on these matters with Dr. Geoffrey Barnes, Simon Williams, Professor Matt Bland, Professor Christopher Koper, Dr. Heather Strang, and Professor Eliot L. Sherman.

References

Ares, C. E., Rankin, A., & Sturz, H. (1963). The Manhattan Bail Project: An interim report on the use of pretrial parole. *New York University Law Review, 38*, 67.

Barnes, G. C., Williams, S., Sherman, L. W., Parmar, J., House, P., & Brown, S. A. (2020). Sweet spots of residual deterrence: A randomized crossover experiment in minimalist police patrol. https://scholar.google.co.uk/scholar?cluster=15405365609504279630&hl=en&as_sdt=0,5. Accessed 19 Mar 2022

Basford, L., Sims, C., Agar, I., Harinam, V., & Strang, H. (2021). Effects of one-a-day foot patrols on hot spots of serious violence and crime harm: A randomised crossover trial. *Cambridge Journal of Evidence-Based Policing, 5*(3), 119–133.

Bland, M., Leggetter, M., Cestaro, D., & Sebire, J. (2021). Fifteen minutes per day keeps the violence away: A crossover randomised controlled trial on the impact of foot patrols on serious violence in large hot spot areas. *Cambridge Journal of Evidence-Based Policing, 5*(3), 93–118.

Bland and Leggetter. (2021). Presentation to the Evidence-Based Policing 2021 Conference, University of Cambridge, July.

Braga, A., Turchan, B., Papachristos, A. and Hureau, D. (2019). Hot spots policing of small geographic areas effects on crime. *Campbell Systematic Reviews*, 15(3). https://doi.org/10.1002/cl2.1046

Cox, D. R. (1958). *Planning of experiments*. Wiley.

Citizendium (2010). Cross-over study. (2010, September 23)., Retrieved 17:23, April 17, 2022 from https://citizendium.org/wiki/index.php?title=Cross-over_study&oldid=725049

Farr, O. M., Tuccinardi, D., Upadhyay, J., Oussaada, S. M., & Mantzoros, C. S. (2017). Walnut consumption increases activation of the insula to highly desirable food cues: A randomized, double-blind, placebo-controlled, cross-over fMRI study. *Diabetes, Obesity and Metabolism, 20*(1), 173–177.

Feinberg, S., Larntz, K., & Reiss, A. J., Jr. (1976). Redesigning the Kansas City preventive patrol experiment. *Evaluation, 3*, 124–131.

Fisher, R. A. (1935). *The design of experiments*. Oliver & Boyd.

Kelling, G. L., Pate, T., Dieckman, D., & Brown, C. (1974). *The Kansas City preventive patrol experiment: A technical report*. Police Foundation.

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.

Koper, C., Lum, C., Wu, X., & Hegarty, T. (2021). The long-term and system-level impacts of institutionalizing hot spot policing in a small city. *Policing, 15*(2), 1110–1128.

McCord, J. (1978). A thirty-year follow-up of treatment effects. *American Psychologist, 33*(3), 284.

Malthouse, C. L. (2021). Address to the Cambridge International Conference on Evidence-Based Policing 2021, July.

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*(1), 123–145.

O'Connor, D. (2022). Lecture to the Cambridge Police Executive Programme, March, 2022.

Powers, E., & Witmer, H. (1951). *An experiment in the prevention of delinquency*. Columbia University Press.

Rose, S. (2022). Personal communication based on UK field research

Sherman, E. L. (2020). Discretionary remote working helps mothers without harming non-mothers: Evidence from a field experiment. *Management Science, 66*(3), 1351–1374.

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

Sherman, L. W. (1998). *Evidence-based policing*. Police Foundation.

Sherman, L. W. (2013). The rise of evidence-based policing: Targeting, testing, and tracking. *Crime and Justice, 42*(1), 377–451.

Sherman, L. W., & Harris, H. M. (2013). Increased homicide victimization of suspects arrested for domestic assault: A 23-year follow-up of the Milwaukee Domestic Violence Experiment (MilDVE). *Journal of Experimental Criminology, 9*(4), 491–514.

Sherman, L. W., & Harris, H. M. (2015). Increased death rates of domestic violence victims from arresting vs. warning suspects in the Milwaukee domestic violence experiment (MilDVE). *Journal of Experimental Criminology, 11*(1), 1–20.

Sherman, L. W., Schmidt, J. D., Rogan, D. P., Gartin, P. R., Cohn, E. G., Collins, D. J., & Bacich, A. R. (1991). From initial deterrence to longterm escalation: Short-custody arrest for poverty ghetto domestic violence. *Criminology, 29*(4), 821–850.

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

Weisburd, D., Telep, C. W., Vovak, H., Zastrow, T., Braga, A. A., & Turchan, B. (2022). Reforming the police through procedural justice training: A multicity randomized trial at crime hot spots. *Proceedings of the National Academy of Sciences, 119*(14), e2118780119.

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

Lawrence W. Sherman is Editor in Chief of the Cambridge Journal of Evidence-Based Policing, and Wolfson Professor of Criminology Emeritus in the University of Cambridge.

Authors and Affiliations

Lawrence W. Sherman[1]

✉ Lawrence W. Sherman
LS434@cam.ac.uk

[1] University of Cambridge, Cambridge, UK