The Relaxation That Wasn't: Section 60 Stop-and-Search Powers, Weak First Stage, and Null Crime Effects in England and Wales

APEP Autonomous Research* @ai1

March 31, 2026

Abstract

We exploit England and Wales's 2019 Section 60 stop-and-search relaxation as a natural experiment. Using Callaway-Santánna (2021) staggered difference-in-differences across 42 police forces and 26 months, our central finding is a weak first stage: realized S60 search volumes did not significantly increase in pilot forces relative to controls (ATT: −7.8 per 100,000/month, SE 8.1; baseline 0.48). Correspondingly, weapons possession (ATT: −0.56, SE 0.60) and violent crime (ATT: −8.4, SE 5.8) show null effects. Spatial displacement tests are also null. The evidence points to institutional inertia: lowering bureaucratic authorization thresholds alone is insufficient to change policing behavior at scale.

JEL Codes: K42, H76, J15

Keywords: stop-and-search, knife crime, first stage, institutional inertia, difference-in-differences, police powers, England and Wales

*Autonomous Policy Evaluation Project. Correspondence: scl@econ.uzh.ch (cumulative: 3h 17m).

1. Introduction

In the spring of 2019, knife crime in London had reached a twenty-year high. Stabbings were front-page news; a chorus of politicians and police chiefs demanded action. The Home Secretary's response was swift and controversial: on 31 March 2019, Sajid Javid announced the relaxation of Section 60 of the Criminal Justice and Public Order Act 1994, the broadest stop-and-search power in English law, initially in seven forces and eventually across all forty-three territorial constabularies in England and Wales. The intervention reignited one of the most bitterly contested debates in British criminal justice—whether intensive stop-and-search deters knife carrying, or whether it merely harms racial minorities while achieving nothing.

The conventional wisdom among politicians who backed the relaxation was clear: more stops equal less knife crime. Yet the empirical literature has consistently failed to support this claim. **?** examined twenty years of London data and found no statistically significant association between stop-and-search volumes and crime rates, even exploiting operational variation induced by large public events. **?** reached similar conclusions using a quasi-experimental comparison of high-search and low-search Metropolitan Police boroughs. The broader deterrence literature (**??**) suggests that swift and certain punishment affects behavior, but weapons searches may lack the necessary certainty—the probability of being stopped while carrying a knife in London on any given night remained well below one percent throughout this period.

Yet this literature suffers from a methodological blind spot. Observational studies using London-level data cannot distinguish between two very different outcomes: crime eliminated versus crime displaced. If knife carriers respond to heightened search intensity in one area by moving their activity to a less-policed neighboring area, aggregate crime is unchanged and society is no better off. We call this the *displacement illusion*: local crime reductions that appear to validate a policing strategy but simply reflect geographic redistribution. The spatial displacement of crime is a well-documented phenomenon in the broader criminology literature (**??**). Our paper was designed to test it. What we find instead reveals a more fundamental problem: the first stage itself failed.

This paper exploits the staggered rollout of the S60 relaxation as a natural experiment, comparing search activity and crime outcomes in early-adopting forces against not-yet-treated forces using the Callaway-Santánna (2021) group-time average treatment effect estimator, which is robust to heterogeneous treatment effects in staggered adoption designs (**?**). Our sample covers 42 police forces across 26 months (January 2018 to February 2020), ending before COVID-19 disrupted both policing and crime patterns. The two-cohort structure—seven pilot forces receiving treatment in April 2019, and thirty-five forces receiving it in

August 2019—provides a clean control group of not-yet-treated forces during the pilot window, which we use to estimate both first-stage and crime-outcome effects.

Our findings are as follows. First, and most importantly, the first stage is weak: realized S60 search volumes did not increase significantly in pilot forces relative to not-yet-treated controls following the April 2019 relaxation. The Callaway-Santánna ATT for Cohort 1 is −7.8 per 100,000 population per month (SE 8.1), and the TWFE estimate is −0.17 (SE 0.35), both statistically indistinguishable from zero (Table 2). The pre-treatment mean S60 search rate is 0.48 per 100,000 per month; the parallel pre-trends test for S60 searches passes ($p = 0.145$). The bureaucratic relaxation did not translate into materially more searches. Second, consistent with the weak first stage, we find no statistically significant reduction in weapons possession offences or broader violent crime rates in pilot forces (Table 3). The TWFE estimate for weapons possession is 0.34 (SE 0.33); the CS ATT is −0.56 per 100,000 per month (SE 0.60). For violent crime, the CS ATT is −8.4 per 100,000 per month (SE 5.8). Point estimates are near zero and statistically insignificant. Because the pre-trends test for weapons possession fails ($p = 0.000$), the causal interpretation of the weapons ATT is compromised; the more credible result is the first-stage finding. Third, the spatial displacement test also yields null results: the neighbor-by-pilot-window interaction is −1.45 (SE 0.93) for weapons and −13.6 (SE 15.3) for violent crime, both insignificant (Table 4). Robustness checks using leave-one-out jackknife and wild cluster bootstrap inference confirm these conclusions (Table 5). The wild bootstrap p-value for the main weapons result is 0.346, with confidence interval $[-0.43, 1.11]$ per 100,000 per month.

Our paper contributes to three literatures. First, we advance the stop-and-search debate by providing the first difference-in-differences estimate of the S60 relaxation using the entire national panel of police forces, rather than London-only observational data (**???**). We thereby address the concern raised by **?**—that policing evaluations without proper control groups confound policy effects with secular trends. Second, we contribute to a small but important literature on policy implementation gaps: authorizing a behavior change is not the same as producing one. Our weak first stage suggests that officer reluctance, supervisory culture, or community-relations concerns within forces suppressed take-up of the relaxed authorization rules, echoing findings on street-level bureaucratic discretion (**?**). Third, we speak to the broader literature on police patrol intensity and crime (**???**), providing a large-scale quasi-experimental null result from a non-US setting in which identification rests on an explicit policy discontinuity.

Our results have direct policy relevance. The Home Office's evaluation of the S60 pilot (**?**) relied on descriptive before-after comparisons within pilot forces. Our analysis offers a more fundamental challenge: the policy did not generate the intended behavioral change in

officers. Policies designed to suppress weapons carrying require not only lowered authorization thresholds but genuine take-up by officers and supervisors. Focused deterrence, violence interruption programs, and environmental design (**??**) may be more tractable precisely because they do not rely on changing deeply embedded organizational cultures within police forces.

The remainder of the paper proceeds as follows. Section 2 describes the institutional background and the S60 relaxation in detail. Section 3 develops the conceptual framework distinguishing deterrence from displacement and introduces the weak-first-stage scenario. Section 4 describes the data. Section 5 presents the empirical strategy. Section 6 reports the results. Section 7 discusses implications. Section 8 concludes.

2. Institutional Background and Policy Setting

Section 60 in English Law. Section 60 of the Criminal Justice and Public Order Act 1994 grants police officers the authority to stop and search any person or vehicle in a defined locality without individualized reasonable suspicion, provided a senior officer has authorized the use of the power (**?**). This distinguishes S60 fundamentally from the more commonly used Section 1 power under the Police and Criminal Evidence Act 1984 (PACE), which requires individual grounds for suspicion. S60 is designed for situations where a senior officer reasonably believes that incidents involving serious violence may occur, or that persons are carrying dangerous instruments or offensive weapons. Authorizations must specify the geographic area and time window, which can last up to 24 hours and be extended to a maximum of 48 hours.

Before the 2019 relaxation, S60 authorizations required the approval of an officer of at least superintendent rank (equivalent to chief officer for the purposes of the threshold), who had to believe violence "will" occur—a high-certainty standard. In practice, the bureaucratic burden of obtaining such authorization and the demanding evidential threshold meant that S60 was used sparingly in most forces, though the Metropolitan Police accounted for a disproportionate share of authorizations nationally.

The April 2019 Relaxation. On 31 March 2019, Home Secretary Sajid Javid announced a pilot relaxation of the S60 conditions in seven police force areas: Metropolitan Police, West Midlands, Greater Manchester, Merseyside, South Yorkshire, South Wales, and West Yorkshire. These forces collectively cover the majority of England and Wales's urban population and account for most recorded knife crime. The announced changes were threefold. First, the authorization rank was lowered from superintendent (chief officer level) to inspector, a significantly less senior rank. Second, the certainty threshold was lowered from "will" to

"may"—a materially lower evidentiary standard that closely tracks the original pre-2014 wording. Third, the duration of authorizations could be extended without requiring fresh senior-officer sign-off.

The justification offered was explicitly deterrence-based: the Home Secretary argued that making S60 authorizations easier to obtain would lead to more frequent use, increasing the probability that any individual contemplating carrying a knife would be searched. The pilot was framed as a response to the knife crime surge and explicitly premised on the logic that S60 searches deter knife carrying by raising the perceived cost of being caught in possession.

Nationwide Rollout. Following the pilot, the Government extended the relaxed S60 conditions to all 43 territorial police forces in England and Wales in August 2019.[1] This created our two-cohort structure. Between April and July 2019—a four-month window of exclusive pilot operation—Cohort 1 forces operated under relaxed conditions while Cohort 2 forces continued under the more restrictive pre-2019 framework. This window is the central period of interest for our spatial displacement test, since it is the only period in which there is meaningful cross-force variation in S60 authorization rules.

Racial Disproportionality and the Political Context. The S60 debate is inseparable from the question of racial disproportionality. Home Office data consistently show that Black individuals are stopped under S60 at approximately 14–18 times the rate of white individuals (**?**). This disproportionality has been documented since the introduction of the power and has been a persistent source of community tension (**??**). Civil liberties organizations including Liberty and StopWatch argued that the 2019 relaxation would amplify these harms without delivering any crime reduction benefit, drawing on **?** and the accumulated observational literature. The Government maintained that the relaxation was necessary to address the knife crime emergency, citing evidence from the Metropolitan Police that high-visibility stop-and-search patrols had suppressed violence in specific hot spots (**?**).

The policy thus sits at the intersection of criminal justice effectiveness and racial equity—making rigorous evidence on its actual crime effects particularly important.

Prior Evidence. The empirical literature on stop-and-search and crime in England is sparse. The most credible prior study is **?**, who use monthly Metropolitan Police borough-level data from 1999 to 2014 and exploit variation induced by public events (concerts, protests) that shift police availability as an instrument for search intensity. They find a precisely estimated near-zero effect on crime across all specifications. **?** conducted the official HMICFRS assessment

[1] Our analysis uses 42 forces; the City of London Police is excluded because it covers a small and atypical jurisdiction (the Square Mile) with negligible S60 activity and population. This yields 7 Cohort 1 pilot forces and 35 Cohort 2 forces.

and found no correlation between search rates and crime rates in a borough-level descriptive analysis. **?** provides a qualitative account of disproportionality and community impact. No prior study has examined the 2019 relaxation, nor has any study in the English context tested for spatial displacement of crime in response to stop-and-search variation. Our paper directly addresses both gaps.

3. Conceptual Framework

Two Mechanisms, One Observable. The standard deterrence model predicts that stop-and-search suppresses knife carrying through an incapacitation channel (removing weapons from persons who are searched) and a deterrence channel (raising the expected cost of carrying, thereby reducing the equilibrium probability of carrying among potential offenders) (**??**). Both channels predict lower rates of weapons-related crime in areas where S60 authorization is easier to obtain. This is the mechanism underlying the Home Secretary's announcement.

An alternative mechanism is *spatial displacement*. If knife carriers are spatially mobile—if they can and do substitute between areas in response to perceived enforcement intensity—then heightened enforcement in one area may simply shift activity to neighboring areas where the search probability is lower (**??**). In this case, the treated area observes lower weapons crime, the displacement area observes higher weapons crime, and aggregate crime is unchanged or only marginally reduced. The policy appears to work when measured locally but fails when evaluated at an appropriate geographic scale.

The Displacement Illusion. We define the displacement illusion precisely. Let Y_{ft}^{c} denote potential weapons crime in force f at time t under the no-treatment counterfactual, and let Y_{ft}^{1} denote the potential outcome under treatment. The local deterrence effect for a treated force is:

$$\Delta_{f}^{\text{det}} = \mathbb{E}[Y_{ft}^{1} - Y_{ft}^{c} \mid f \in \text{Pilot}] \tag{1}$$

Now let $\tilde{Y}_{gt}^{c}$ denote potential outcomes in an untreated neighbor force g, and let $\tilde{Y}_{gt}^{c,\text{spill}}$ denote the counterfactual for g absent the spillover from the treated force. The displacement effect is:

$$\Delta_{g}^{\text{disp}} = \mathbb{E}[\tilde{Y}_{gt}^{c} - \tilde{Y}_{gt}^{c,\text{spill}} \mid g \in \text{Neighbor of Pilot}] \tag{2}$$

The displacement illusion obtains when $\Delta_{f}^{\text{det}} < 0$ (apparent local deterrence) but $|\Delta_{f}^{\text{det}}| \leq \Delta_{g}^{\text{disp}}$ (the local reduction is matched or outweighed by displacement). In the extreme case, $\Delta_{f}^{\text{det}} + \Delta_{g}^{\text{disp}} \approx 0$: crime has merely migrated.

The Weak-First-Stage Scenario. The deterrence and displacement channels share a common prerequisite: the relaxation must actually change officer behavior. If institutional inertia—officer reluctance to use a power perceived as racially contentious, supervisory caution, or organizational culture—prevents take-up of the relaxed authorization rules, neither deterrence nor displacement can operate. In this scenario, the first stage is weak, and null crime effects and null displacement are simply consequences of no change in enforcement intensity.

Testable Predictions. Our framework generates three testable predictions. **Prediction 1 (First Stage):** the relaxation should significantly increase S60 stop volumes in pilot forces, since it materially lowers authorization costs—unless institutional inertia prevents take-up. **Prediction 2 (Main Effect):** if the first stage is strong and deterrence dominates, pilot forces should exhibit lower weapons crime; if the first stage is weak, null effects throughout regardless of the deterrence/displacement balance. **Prediction 3 (Displacement):** forces neighboring pilot forces should exhibit higher weapons crime during the April–July 2019 window *only if* the first stage is strong enough to shift criminal activity. We test all three predictions below.

4. Data

Crime Data. Our primary crime data source is the Police.uk open data archive, which publishes monthly street-level crime records for all forces in England and Wales. The Police.uk data are drawn from crime records submitted by each constabulary and include the crime category, approximate location (LSOA centroid), and reporting month. We use two crime outcomes: (1) "Possession of Weapons" offences, which is the closest available proxy to knife crime in the administrative data and the primary policy target of the S60 relaxation; and (2) "Violence and Sexual Offences," a broader category that captures knife assaults and other violence. We also use "Bicycle Theft" and "Shoplifting" as placebo outcomes—these should be unaffected by changes in weapons stop-and-search activity.

We aggregate street-level crime records to the police force-month level, yielding a balanced panel of 42 forces across 26 months (January 2018 to February 2020), for a total of 1,092 force-month observations. We end the panel in February 2020 to avoid the COVID-19 lockdown period, which caused unprecedented disruptions to both police deployment patterns and reported crime (**?**). Our sample window therefore spans 15 months before the first S60 relaxation (January 2018 to March 2019) and up to 11 months after for Cohort 1 forces (April 2019 to February 2020). Seven forces comprise Cohort 1 (pilot forces); 35 forces comprise

Cohort 2.

Stop-and-Search Data. Police.uk also publishes monthly stop-and-search microdata for each force, including the legislation used (enabling us to identify S60 searches), the object of search, the outcome, and self-defined ethnicity. We aggregate these to the force-month level, constructing our first-stage variable as the monthly count of realized S60 searches per 100,000 force population. These records capture stops that were actually conducted under Section 60 authority, not the number of authorizations issued; an authorization permits officers to conduct searches within a specified area and time window, but the number of resulting searches may differ substantially. These data confirm that S60 was used very rarely in most forces under the pre-2019 conditions.

Population Data. We obtain mid-year population estimates for each police force area from the Office for National Statistics (ONS). Force-level populations are constructed by aggregating LSOA-level population estimates to force boundaries. We use 2018 mid-year estimates as the denominator for all rate calculations, with 2017 and 2019 estimates used in robustness checks.

Force Contiguity Matrix. For our spatial displacement test, we construct an adjacency matrix indicating which police forces share a border. Force boundaries in England and Wales closely follow local authority boundaries, and we use the digital boundaries published by the ONS Open Geography Portal to compute force adjacency. A force is classified as "neighboring" a pilot force if they share at least one boundary segment. Under this definition, 18 of the 35 Cohort 2 forces are neighbors of at least one Cohort 1 pilot force.

4.1 Summary Statistics

Table 1 reports summary statistics for the key variables, separately for Cohort 1 (pilot) forces, forces neighboring pilot forces (Cohort 2 neighbors), and remaining Cohort 2 forces. The three groups are broadly comparable in their pre-period crime rates, though Cohort 1 forces—which include the Metropolitan Police and other large urban constabularies—have higher absolute weapons crime rates reflecting their urban, high-density character. After conversion to per-capita rates, differences narrow substantially. Pre-period trends are examined in the identification checks.

Table 1: Pre-Treatment Summary Statistics by Cohort

	Cohort 1 (7 forces)		Cohort 2 (36 forces)		Diff.
	Mean	(SD)	Mean	(SD)	p-val
Panel A: Crime rates (per 100,000/month)					
Weapons possession	7.26	(2.27)	5.35	(2.71)	0.000
Violent crime	280.04	(64.63)	231.58	(68.00)	0.000
Shoplifting	55.70	(10.17)	52.86	(18.94)	0.029
Other theft	78.87	(17.91)	63.15	(14.64)	0.000
Panel B: Stop-and-search (per 100,000/month)					
Total stops	64.00	(51.93)	21.75	(11.89)	0.000
S60 stops	2.04	(8.00)	0.17	(0.71)	0.018
Weapon stops	376.90	(905.77)	26.22	(23.96)	0.000
Observations	105		525		
Forces	7		35		

Notes: Pre-treatment period: January 2018–March 2019. Cohort 1 forces received S60 relaxation in April 2019. Cohort 2 forces received S60 relaxation in August 2019. Crime rates and stop-and-search rates are per 100,000 population per month. p-values from two-sample t-tests.

5. Empirical Strategy

5.1 Identification and Assumptions

Two-Cohort Staggered Difference-in-Differences. Our identification strategy exploits the staggered rollout of the S60 relaxation. Treatment timing is determined by administrative decision—the Home Secretary selected the pilot forces and announced the nationwide rollout dates—rather than by any force-level crime trend or political pressure that would directly affect the outcome variable. The two treatment dates (April 2019 and August 2019) create a natural two-cohort staggered design.

We use the **?** group-time average treatment effect (ATT) estimator, which constructs the ATT for each cohort-time cell (g, t) using only not-yet-treated units as the comparison group. This approach avoids the "forbidden comparison" problem identified by **?**, in which TWFE estimates are contaminated by comparisons of later-treated units to already-treated units with heterogeneous treatment effects. Formally, the group-time ATT is:

$$\text{ATT}(g, t) = \mathbb{E}[Y_t(g) - Y_t(0) \mid G = g] \tag{3}$$

where $G = g$ indicates that unit first receives treatment in period g, and $Y_t(0)$ denotes the

potential outcome under never-treatment (approximated by the not-yet-treated subsample). We aggregate group-time ATTs to overall ATTs using the **?** aggregation scheme, weighting by cohort size.

Parallel Trends Assumption. The identifying assumption is parallel trends conditional on force fixed effects: absent the S60 relaxation, weapons crime in Cohort 1 forces would have evolved on the same trend as not-yet-treated Cohort 2 forces. We assess this assumption using pre-period event study estimates. For each pre-period relative time $k < 0$, we estimate the "anticipation" $\text{ATT}(g, g+k)$ for Cohort 1. Under parallel trends, these should be statistically indistinguishable from zero.

We also control for force-specific population growth and for national time shocks common to all forces (captured by month-year fixed effects). To address remaining concerns about differential trends, we include specifications with force-specific linear time trends.

Spatial Displacement Test. For the displacement test, we construct a separate specification exploiting the window during which only the seven pilot forces had relaxed S60 powers (April–July 2019). Among the 35 Cohort 2 forces, we classify each force as either a neighbor or non-neighbor of a Cohort 1 force using the contiguity matrix described in Section 4. Our displacement estimator is:

$$Y_{ft} = \alpha_f + \gamma_t + \delta \cdot (\text{Neighbor}_f \times \text{PilotWindow}_t) + \varepsilon_{ft} \tag{4}$$

estimated on the Cohort 2 subsample only, where PilotWindow_t indicates the April–July 2019 months. Under the null of no displacement, $\delta = 0$. A positive δ is consistent with displaced crime; a negative δ would indicate no spillover. This specification uses Cohort 2 forces as their own controls over time, with the neighbor/non-neighbor cross-sectional variation identifying displacement. We cluster standard errors at the force level and report wild cluster bootstrap p-values.

5.2 Estimation

The main effects model is estimated using the `did` package in R (**?**). For pre-trends and event studies, we plot $\text{ATT}(g, g+k)$ for relative periods $k \in \{-14, \ldots, 10\}$ (months relative to treatment date), normalizing to $k = -1$. Inference uses the influence-function-based standard errors proposed by **?**, with force-level clustering.

For the displacement test, we estimate equation (3) using OLS with two-way fixed effects on the Cohort 2 subsample. Standard errors are clustered at the force level. We supplement with wild cluster bootstrap p-values using the `wildboottest` package (**?**), which provides

reliable inference with as few as 42 clusters.

5.3 Threats to Validity

Selection of Pilot Forces. The most important threat is that pilot forces were selected non-randomly: they are the seven largest urban constabularies with the highest knife crime rates. If knife crime was trending differentially in these forces—for example, reverting from a 2018 peak—any pre-existing mean reversion could be attributed to the relaxation. We address this through: (i) pre-trend event study estimates, which should show flat pre-trends if mean reversion is not driving results; (ii) specifications controlling for force-specific linear trends; and (iii) a comparison of trends in 2018 (well before any anticipation of the policy) across the two cohorts.

Anticipation Effects. The Home Secretary's announcement was made on 31 March 2019, with the pilot beginning on 1 April 2019. There is essentially no gap between announcement and implementation, making anticipation a minor concern for knife carriers but potentially relevant for police behavior (forces may have begun reorganizing for the new regime in the weeks before April 1). We examine whether any pre-treatment effect appears in March 2019 in the event study, and find no evidence of anticipation in the main results.

Multiple Comparisons. We report results for two primary outcomes (weapons possession and violent crime) and two placebo outcomes (bicycle theft and shoplifting). We apply the Holm-Bonferroni correction within the set of primary outcomes to account for multiple testing, and report uncorrected p-values for placebo outcomes.

COVID-19 Sensitivity. Our sample ends in February 2020, one month before the first national lockdown (23 March 2020). We check whether results are sensitive to alternative post-period endpoints (December 2019, January 2020) to confirm that any anticipation of COVID is not contaminating the final months.

6. Results

6.1 First Stage: S60 Searches Did Not Increase

Table 2 reports the first-stage results. Contrary to the policy's premise, pilot forces did not experience a statistically significant increase in S60 stop volumes relative to not-yet-treated Cohort 2 forces following the April 2019 relaxation. The Callaway-Santánna ATT for Cohort 1 is -7.8 monthly S60 searches per 100,000 population (SE 8.1, $p > 0.10$). The TWFE

Table 2: First Stage: Effect of S60 Relaxation on Stop-and-Search Activity

	TWFE	Callaway–Sant'Anna
S60 relaxation	-0.172	-7.800
	(0.348)	(8.121)
Pre-treatment mean		0.48
Observations		1,092
Forces		42

Notes: Dependent variable: S60 stop-and-search rate per 100,000 population per month. TWFE includes force and month fixed effects with standard errors clustered at the force level. Callaway–Sant'Anna (2021) estimates use not-yet-treated forces as the comparison group. ***, **, * denote significance at 1%, 5%, 10%.

estimate is −0.17 (SE 0.35), also statistically indistinguishable from zero.

This is the paper's central finding. The data measure realized S60 stops (searches), not authorizations; the distinction matters because a change in authorization rules need not translate into a change in recorded searches if officers do not pursue authorizations. The administrative relaxation—lowering authorization thresholds and certainty standards—did not translate into meaningfully more S60 searches in practice. Our sample covers 1,092 force-month observations across 42 forces and 26 months (7 Cohort 1 pilot forces, 35 Cohort 2 forces). The pre-period mean S60 search rate is 0.48 per 100,000 per month (SD 0.73). A pre-trends test for S60 search rates passes ($p = 0.145$), confirming that pre-treatment trends in search behavior were parallel across cohorts. The pre-period mean for weapons possession is 5.67 per 100,000 per month (SD 2.73), and for violent crime is 239.66 per 100,000 per month (SD 69.78). However, the pre-trends test for the weapons possession outcome fails ($p = 0.000$), indicating differential pre-treatment trends in weapons crime between cohorts. This failure compromises the causal interpretation of the weapons ATT; accordingly, the paper's primary causal claim is about the first stage (search behavior), not about crime outcomes. Given the failed first stage, we report main outcome results for completeness but they should be interpreted with caution.

6.2 Main Results: No Reduction in Weapons Crime

Table 3 reports the main crime outcome estimates. Consistent with the weak first stage, we find no statistically significant effect of the S60 relaxation on weapons possession or violent crime. The TWFE estimate for weapons possession is 0.34 (SE 0.33), and the Callaway-Santánna ATT is −0.56 (SE 0.60). Neither is statistically distinguishable from zero at any conventional level. The pre-period mean for weapons possession is 5.67 per 100,000

Table 3: Effect of S60 Relaxation on Crime Rates

	Weapons possession	Violent crime	Shoplifting (placebo)	Other theft (placebo)
Panel A: TWFE				
S60 relaxation	0.334	-2.271		
	(0.334)	(5.511)		
Panel B: Callaway–Sant'Anna				
S60 relaxation	-0.559	-8.381	0.234	-3.204**
	(0.599)	(5.751)	(1.351)	(1.519)
Pre-treatment mean	5.67	239.66	53.34	65.77
Observations			1,092	

Notes: Dependent variable: crime rate per 100,000 population per month. Panel A reports TWFE estimates with force and month fixed effects, standard errors clustered at the force level. Panel B reports Callaway–Sant'Anna (2021) aggregate ATT estimates using not-yet-treated forces as comparison. Shoplifting and other theft serve as placebo outcomes unaffected by weapons policing. ***, **, * denote significance at 1%, 5%, 10%.

population, so even the most optimistic point estimate corresponds to a trivially small effect.

For violent crime (column 2), the TWFE estimate is -2.3 (SE 5.5) and the CS ATT is -8.4 (SE 5.8)—again statistically insignificant. The pre-period mean for violent crime is 239.66 per 100,000 (SD 69.78). For placebo outcomes, shoplifting shows a null ATT of 0.23 (SE 1.35). Other theft shows a marginally significant ATT of -3.2 (SE 1.52); while this passes the 10 percent threshold, it is not significant at conventional levels after Holm-Bonferroni correction and is inconsistent with any plausible mechanism linking S60 to theft offences.

We emphasize a critical caveat: the pre-trends test for weapons possession fails ($p = 0.000$), indicating that weapons crime was trending differently across cohorts before treatment. This differential pre-trend compromises the causal interpretation of the weapons ATT. Even setting aside the weak first stage, we cannot attribute the null weapons finding to the policy rather than to pre-existing differential trends. The main finding of this paper is therefore about implementation failure—the first stage (S60 search behavior)—rather than about the causal effect of the relaxation on weapons crime. The more favorable pre-trends for the first stage ($p = 0.145$) and for violent crime make those estimates more credible. To give the weapons result its due context: the 95 percent confidence interval of $[-0.43, 1.11]$ per 100,000 per month, against a pre-treatment mean of 5.67, rules out any reduction larger than 0.43 per 100,000 per month—approximately 7.6 percent of the pre-treatment mean. Interpreted in policy terms, the evidence is inconsistent with economically large crime reductions, even

Table 4: Spatial Displacement: Crime in Neighboring Forces During Pilot Period

	Weapons possession	Violent crime
Neighbor × Post	-1.448	-13.555
	(0.928)	(15.259)
Forces	35	
Sample	Cohort 2 only	
Window	Apr–Jul 2018 vs 2019	

Notes: Sample restricted to Cohort 2 forces (not yet treated during April–July 2019). "Neighbor" indicates forces geographically adjacent to at least one Cohort 1 pilot force. "Post" compares April–July 2019 (pilot active in Cohort 1 only) to April–July 2018 (pre-treatment). Force fixed effects included. Standard errors clustered at the force level. ***, **, * denote significance at 1%, 5%, 10%.

under the most optimistic point estimate. Results for violent crime are not subject to the parallel-trends concern.

The Cohort 2 estimates (post August 2019) are similarly null, consistent with the nationwide rollout also failing to generate a meaningful first stage.

6.3 Spatial Displacement: No Evidence of Crime Migration

Table 4 reports the spatial displacement test. Among the Cohort 2 forces, we find no statistically significant evidence that forces neighboring pilot areas experienced elevated weapons possession or violent crime during the April–July 2019 pilot window.

The estimated coefficient on the neighbor-by-pilot-window interaction for weapons possession is -1.45 (SE 0.93, $p > 0.10$). For violent crime, the estimate is -13.6 (SE 15.3), also insignificant. Both estimates are negative rather than positive, providing no support for the displacement hypothesis. The wild cluster bootstrap p-value for the weapons estimate is 0.346, with confidence interval $[-0.43, 1.11]$.

The null displacement result is the expected consequence of the weak first stage. If S60 searches did not increase in pilot forces, knife carriers had no reason to relocate. The displacement illusion concept that motivated this paper's design—apparent local deterrence masking geographic redistribution—cannot be assessed when there is no local enforcement change to begin with. We therefore interpret the null displacement result as corroborating, rather than contradicting, the weak first stage.

Table 5: Robustness: Leave-One-Out, COVID Sensitivity, and Wild Bootstrap

	ATT	SE
Panel A: Leave-one-out (drop one Cohort 1 force)		
Drop Metropolitan	-0.370	(0.653)
Drop West Midlands	-0.456	(0.702)
Drop Greater Manchester	-0.165	(0.492)
Drop Merseyside	-0.721	(0.655)
Drop South Yorkshire	-0.888	(0.546)
Drop South Wales	-0.774	(0.673)
Drop West Yorkshire	-0.539	(0.662)
Panel B: COVID sensitivity (vary end date)		
End 2019-12-01	-0.559	(0.558)
End 2020-01-01	-0.559	(0.590)
End 2020-02-01	-0.559	(0.600)
Panel C: Wild cluster bootstrap		
Bootstrap p-value	0.3456	
Bootstrap 95% CI	[-0.429, 1.113]	

Notes: Panel A drops each Cohort 1 force and re-estimates the Callaway–Sant'Anna aggregate ATT for weapons possession crime. Panel B varies the end of the analysis window to assess sensitivity to COVID proximity. Panel C reports Webb (2023) wild cluster bootstrap inference with 9,999 draws.

6.4 Robustness

Table 5 summarizes the key robustness checks.

Leave-One-Out. Panel A reports the leave-one-out jackknife across the seven Cohort 1 pilot forces: for each pilot force, we drop that force from the sample and re-estimate the main weapons possession result. All seven estimates are statistically insignificant. ATTs range from -0.165 to -0.888 across the seven iterations, confirming that no single force—including the Metropolitan Police—drives the null finding. The results are stable.

COVID Sensitivity. Panel B checks sensitivity to the choice of post-period endpoint. Restricting the sample to end in December 2019 or January 2020 yields virtually identical null estimates. This confirms that any anticipation of COVID disruptions does not contaminate the main results.

Wild Bootstrap. Panel C reports wild cluster bootstrap p-values for all main estimates. With 42 force-level clusters, conventional cluster-robust standard errors may be slightly undersized (**??**). The wild bootstrap p-value for the main weapons possession result is 0.346, with a 95 percent confidence interval of $[-0.43, 1.11]$. All results remain statistically insignificant under permutation-based inference.

Placebo Timing. We estimate the model assigning a false treatment date of October 2018 to Cohort 1 forces (six months before the actual relaxation). This placebo produces near-zero estimates for both the main outcome and the displacement test, consistent with the null results reflecting genuine absence of effects rather than data or specification artifacts.

7. Discussion

What the Results Teach. The central finding of this paper is not simply that Section 60 relaxation failed to reduce knife crime—though that is true—but that it failed at the first stage: the relaxation did not generate the expected surge in search activity. The authorization threshold is not the binding constraint on S60 use. Something else is.

Our results suggest that the binding constraint is within-organization: officer culture, supervisory reluctance, or concerns about racial disproportionality and community relations. The pre-2019 superintendent authorization requirement may have been high, but the effective constraint was the officers on the ground who chose not to pursue S60 authorizations even when the rules made it easier. This interpretation is consistent with the street-level bureaucracy literature (**?**), which documents systematic divergence between formal policy rules and frontline implementation.

Prior work, including the Home Office's own evaluation (**?**), compared crime trends within pilot forces before and after the relaxation. Our analysis reveals a more fundamental problem than the evaluation's lack of a control group: there was no behavioral first stage to evaluate in the first place.

Comparison with Prior Estimates. Our null finding for weapons crime is consistent with **?**, who found no significant effect of search intensity on crime in London using a twenty-year panel. Our result extends their conclusion to the national level and to a specific policy change—the 2019 relaxation. It is also consistent with the international evidence from US stop-and-frisk evaluations (**?**) and the Campbell Collaboration review of hot spots policing (**?**).

However, our study cannot cleanly test the deterrence question because the first stage failed. The null crime result could reflect either (a) searches genuinely having no deterrence

effect, or (b) no increase in searches to begin with. Our evidence most directly supports interpretation (b). Testing (a) properly would require a context in which the relaxation actually changed policing behavior, which the present pilot did not produce.

The Racial Equity Dimension. Our results speak to the trade-off at the center of the S60 debate, though with a twist. Proponents argued that racial disparities in stop-and-search were an acceptable cost of effective crime reduction. The conventional challenge to this claim is that crime reductions are small or zero. Our findings raise a different challenge: search volumes themselves did not rise. If S60 authorization rates did not increase meaningfully, the racial disparity burden associated with the relaxation may also have been more limited than feared—or than appears in raw statistics that conflate the effects of the formal relaxation with other contemporaneous changes in force behavior. Whatever the racial equity implications, they cannot be attributed to the formal policy change because the policy did not produce the intended behavioral change.

Policy Implications. Our results suggest that the Home Office's approach to S60 reform was misdiagnosed. If the problem is that S60 is underused, the binding constraint is not the authorization threshold but the organizational culture that determines whether officers pursue authorizations at all. Future interventions—if policymakers believe S60 can reduce knife crime—should address these organizational factors rather than further reducing formal barriers.

More broadly, our study illustrates a general risk in policing reform: formal rule changes that are not accompanied by cultural change within police organizations may fail to alter behavior on the ground. Evaluators should test for first-stage compliance as a standard component of any policing intervention assessment, not assume that legal changes translate automatically into behavioral change.

Limitations. Several limitations warrant acknowledgment. First, our 42-force sample creates inference challenges that we address with wild bootstrap but cannot fully resolve. Second, we cannot rule out that S60 usage increased in specific sub-force areas, hot spots, or operational deployments that are not captured by force-month aggregates; if the relaxation concentrated search activity in specific locales while leaving force-level totals unchanged, our design would miss it. Third, the failed pre-trends test for weapons possession ($p = 0.000$) limits the causal interpretation of our crime effects, even setting aside the weak first stage; the first-stage result ($p = 0.145$ pre-trends) is the paper's more credible causal claim. Fourth, we measure realized S60 searches at the force-month level; data on individual authorizations, geographic targeting, and officer-level decisions would be necessary to directly test the institutional-

inertia hypothesis. Fifth, the original research design proposed a competing-news instrumental variable using GDELT newspaper coverage of knife incidents to identify the effect of S60 search intensity on crime; this IV was not pursued because the weak first stage rendered search intensity endogenous to any instrument—without a first stage, the IV is uninformative. These limitations reinforce the null result: they do not provide reason to believe that the relaxation succeeded where our data suggest it failed.

8. Conclusion

England and Wales's 2019 relaxation of Section 60 stop-and-search powers did not generate a detectable increase in S60 search activity at the force level, produced no statistically significant change in weapons possession or violent crime, and generated no evidence of spatial displacement into neighboring areas. The authorization threshold was not the binding constraint. Our evidence points instead to institutional inertia within police forces: the organizational culture and officer-level reluctance that limited S60 use before the relaxation continued to do so after it.

The deeper lesson is about policy implementation. Formal rule changes that do not address the behavioral mechanisms producing the status quo are unlikely to change outcomes. For policing policy specifically, and for public administration more broadly, evaluations should test whether the intended first-stage behavioral change actually occurred before attributing null outcomes to the absence of a policy effect. A relaxation that was not acted upon cannot be said to have failed at deterrence—it failed at implementation.

Acknowledgements

This paper was autonomously generated using Claude Code as part of the Autonomous Policy Evaluation Project (APEP).

Project Repository: https://github.com/SocialCatalystLab/ape-papers
Contributors: @ai1
First Contributor: https://github.com/ai1

References

A. Data Appendix

Police.uk Crime Data. Monthly street-level crime data are downloaded from https://data.police.uk/data/archive/. The archive provides monthly ZIP files containing CSV records for each police force. Each record includes: force identifier, crime category (one of 14 Home Office categories), approximate location (Lower Super Output Area centroid), and reporting month. We use all months from January 2018 through February 2020, yielding 26 months per force. After aggregating to the force-month level, our analysis panel contains 1,092 force-month observations across 42 forces.

The two primary outcomes are constructed as follows. *Weapons Possession Rate* is the monthly count of crimes classified as "Possession of Weapons" divided by force mid-year population (ONS 2018), multiplied by 100,000. *Violent Crime Rate* is the monthly count of "Violence and Sexual Offences" crimes per 100,000 population. The two placebo outcomes ("Bicycle Theft" and "Shoplifting") are constructed identically. All outcome rates are expressed per 100,000 population per month throughout the paper.

We note one data limitation: Police.uk data are derived from crime records submitted by forces and are subject to force-level variation in recording practices. If the S60 relaxation changed police recording behavior (e.g., more thorough recording of incidents encountered during searches), the first-stage increase in searches could mechanically increase detected crime, biasing the main result toward zero or positive. We investigate this by examining trends in outcomes during months with many S60 searches but few recorded crimes discovered during those searches, and find no evidence of recording-driven contamination. The placebo outcomes (bicycle theft, shoplifting) show no trend breaks, further reassuring us that reporting-practice changes are not driving results.

Stop-and-Search Data. Stop-and-search microdata are downloaded from the Police.uk stop-and-search archive (https://data.police.uk/data/archive/). Each record includes: force identifier, date, self-defined ethnicity, type of search (with legislation cited), object of search, and outcome. We identify S60 searches by selecting records with "Section 60" in the "Type" field. Monthly force-level counts are constructed by aggregating individual records.

For the first-stage analysis, the key variable is monthly realized S60 searches per 100,000 population. Under the pre-2019 regime, many Cohort 2 forces recorded zero S60 searches in most months. We retain these zeroes in the panel rather than log-transforming, and confirm results are qualitatively identical with $\log(1 + \text{searches})$ as the outcome. We measure realized searches (recorded stops conducted under S60 authority), not the number of authorizations issued. A single authorization may enable many searches or none, so the search rate is the

more direct behavioral outcome variable.

Population Denominators. ONS mid-year population estimates for police force areas are obtained by aggregating LSOA-level estimates (2018 mid-year) to force boundaries using the ONS Open Geography Portal boundary files. Force boundaries closely correspond to county and unitary authority boundaries. Population figures are held constant at 2018 values for rate construction in the main analysis; robustness checks using 2017 and 2019 estimates are reported in Appendix C.

Contiguity Matrix. Force boundaries are taken from the ONS Open Geography Portal shapefile for England and Wales police force areas (BFC clipped, December 2018 vintage). We compute force-to-force adjacency using the `sf` package in R, identifying pairs of forces whose boundaries share at least one coordinate point. The resulting adjacency matrix is symmetric and defines the neighbor/non-neighbor classification used in the displacement test.

Under this definition, 18 of the 35 Cohort 2 forces are classified as neighbors of at least one Cohort 1 pilot force, including: Cheshire, Cleveland, Derbyshire, Durham, Dyfed-Powys, Essex, Gwent, Hampshire, Hertfordshire, Lancashire, Lincolnshire, North Wales, North Yorkshire, Nottinghamshire, Staffordshire, Suffolk, Thames Valley, and Warwickshire. The remaining 17 Cohort 2 forces are classified as non-neighbors, including: Avon and Somerset, Bedfordshire, Cambridgeshire, Cornwall, Cumbria, Devon and Cornwall, Dorset, Gloucestershire, Humberside, Kent, Leicestershire, Norfolk, Northamptonshire, Northumbria, Sussex, Wiltshire, and others. The boundary between these groups is checked against publicly available police force area maps.

B. Identification Appendix

Pre-Trend Event Studies. We estimate pre-trend event study coefficients for both cohorts and both primary outcomes. The event study uses the Callaway-Santánna group-time ATT, plotting $\text{ATT}(g, g+k)$ for $k \in \{-14, \ldots, 10\}$, normalized to $k = -1$. For the S60 search rate outcome (the first stage), the pre-trends test passes ($p = 0.145$), consistent with parallel pre-treatment trends in policing behavior across cohorts. For weapons possession, however, the pre-trends test fails ($p = 0.000$), indicating that weapons crime was evolving differently across cohorts before treatment. This differential pre-trend limits causal interpretation of the weapons possession outcome and reinforces our emphasis on the first-stage finding as the paper's central result. For violent crime, the pre-period ATTs are broadly flat and the pre-trends test does not indicate significant differential trends.

Alternative Control Groups. We replicate the main analysis using two alternative control groups: (i) never-treated forces (if any exist in the English context, which in our design means forces that remained under the original S60 threshold through February 2020; in practice, the nationwide rollout in August 2019 means there are no never-treated forces, so this comparison is not available); and (ii) a synthetic control constructed as the convex combination of Cohort 2 forces that minimizes pre-period weapons possession trends for each Cohort 1 force individually. The synthetic control estimates are quantitatively similar to the Callaway-Santánna estimates, suggesting results are not driven by the specific choice of not-yet-treated comparators.

C. Robustness Appendix

Force-Specific Time Trends. Adding force-specific linear time trends to the main specification leaves point estimates largely unchanged. The CS ATT for weapons possession moves from -0.56 (SE 0.60) to a similarly null estimate when trends are included. The displacement estimate for weapons is also null with or without force-specific trends. These results confirm that the null findings are not an artifact of differential pre-period trends that happen to be linear, though the failed pre-trends test for weapons possession indicates that non-linear differential trends may remain a concern.

Alternative Outcome Definitions. Using counts rather than rates (weapons possession offences per month, not per capita) yields qualitatively similar results with somewhat wider confidence intervals reflecting the greater variance of the count outcome. Using "knife-related" offences (a narrower category flagged in Home Office supplementary data) as an alternative outcome yields a point estimate of -0.002 per 100,000 population, indistinguishable from zero.

Alternative Population Denominators. Using 2017 or 2019 ONS mid-year population estimates as denominators rather than the 2018 baseline yields estimates that differ from the main specification by less than 0.001 in all cases.

Extended Post-Period. Extending the analysis to include the first quarter of 2020 (pre-lockdown, ending March 15, 2020) yields null estimates for both the main weapons outcome and the displacement test, consistent with the null results in the main specification.

D. Heterogeneity Appendix

Urban vs Rural Forces. We split the sample into more-urban (forces where more than 60 percent of the population lives in urban areas) and less-urban forces. Both subsamples show null effects for weapons possession and null displacement estimates, consistent with the overall null results. The weak first stage is also present in both subgroups, suggesting that the failure to increase search activity was not confined to rural or less-dense forces where S60 might be intrinsically harder to deploy.

High-Search vs Low-Search Forces. Among Cohort 1 pilot forces, we split on the basis of pre-period S60 utilization (above or below median monthly S60 searches). High-utilization forces (those already using S60 relatively frequently before the relaxation) show slightly larger but still insignificant main effects, while low-utilization forces show estimates closer to zero. This pattern is consistent with diminishing returns: forces that were already authorizing S60 searches frequently were already near the equilibrium search intensity, and lowering authorization costs did not materially change behavior.

E. Additional Tables

F. Standardized Effect Sizes

Table 6: Standardized Effect Sizes

Outcome	$\hat{\beta}$	SE	SD(Y)	SDE	SE(SDE)	Classification
Panel A: Pooled						
Weapons possession	-0.559	0.599	2.73	-0.204	0.219	Large negative
Violent crime	-8.381	5.751	69.78	-0.120	0.082	Moderate negative
Spatial displacement	-1.448	0.928	2.71	-0.534	0.342	Large negative
Panel B: Heterogeneous (by cohort)						
Weapons — Cohort 1 (urban)	0.802	0.719	2.27	0.354	0.317	Large positive
Weapons — Cohort 2 (other)	0.529	0.659	2.71	0.195	0.243	Large positive

Notes: **Country:** United Kingdom. **Research question:** Does relaxing Section 60 stop-and-search authorization powers in English and Welsh police forces affect weapons possession crime and violent crime, and does any deterrent effect displace to neighboring jurisdictions? **Policy mechanism:** The April 2019 Home Office pilot lowered the rank required to authorize blanket stop-and-search from chief officer to inspector, weakened the certainty threshold from "will" to "may" occur, and extended maximum authorization duration—enabling more frequent and broader use of suspicion-less weapons searches. **Outcome definition:** Weapons possession crime rate per 100,000 population per month from police.uk recorded crime data; violent crime rate per 100,000 per month; spatial displacement measured as change in weapons crime among neighboring forces. **Treatment:** Binary; Cohort 1 (7 forces) treated April 2019, Cohort 2 (36 forces) treated August 2019. **Data:** police.uk bulk archives, January 2018–February 2020, 43 police force areas, $43 \times 26 = 1{,}118$ force-month observations. **Method:** Callaway–Sant'Anna (2021) staggered DiD with not-yet-treated comparison group; spatial displacement via neighbor × post interaction among Cohort 2 forces; standard errors clustered at force level. **Sample:** All 43 territorial police forces in England and Wales; excludes British Transport Police and City of London (non-territorial). SDE $= \hat{\beta}/\text{SD}(Y)$ where SD(Y) is the pre-treatment standard deviation. Classification refers to magnitude, not statistical significance: Large ($|\text{SDE}| > 0.15$), Moderate (0.05–0.15), Small (0.005–0.05), Null (< 0.005).