Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The Effect of a Compressed Work Schedule on Policing Outcomes: Evidence from New York City

Chin-Huai Shih
University of Pennsylvania
John Hall
University of Cambridge
Aaron Chalfin
achalfin@sas.upenn.edu

University of Pennsylvania

Research Article

Keywords:

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11447601/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Competing interest reported. I am an Associate Editor at JQC. This paper should not be assigned to me!

The Effect of a Compressed Work Schedule on Policing Outcomes: Evidence from New York City

Abstract

Police departments across the United States face growing challenges in recruitment, retention and staffing thus prompting renewed interest in compressed work schedules that replace traditional eight-hour tours with longer shifts and additional days off. Compressed work schedules are thought to promote better work-life balance but an underlying concern is that longer shifts might increase fatigue and reduce the quality of policing. This paper evaluates the New York City Police Department's 2023–2024 pilot implementation of 10- and 12-hour patrol shifts, a major organizational reform affecting some of the city's busiest precincts. This is one of the largest work schedule pilots ever undertaken by a U.S. police agency. Exploiting the staggered rollout of the policy across precincts, we estimate its effects using contemporary staggered difference-in-differences estimators. In most respects, we find little evidence that the compressed schedule resulted in meaningful changes to policing activities or public safety. However, we do observe a 20% increase in dispatch and response times to calls for service that involve a crime in progress, a finding that is not explained by changes in overall staffing levels.

1. Introduction

One of the most demanding features of modern policing is the requirement for 24/7 availability and rapid response to emergency calls. This demand can require officers to have erratic work schedules that few other occupations impose at a comparable scale. The need for constant manpower means that officers often work irregular overtime hours, leading to unpredictable shifts and substantial financial outlays for cities in the form of overtime wages.

The compressed workweek (CWW) has emerged as one of the more prominent proposals for restructuring shift-based work in policing (Bambra et al., 2008; Barnum, 2011; Cunningham, 1981, 1982; Pierce & Dunham, 1992; Stenzel & Buren, 1986; Vega & Gilbert, 1997; Vila et al., 2000). While implementations vary, the defining feature of a CWW is the elongation of the basic work shift beyond the typical 8 hours in exchange for additional days off across the work cycle. Proponents of the CWW argue that the longer recovery windows improve work-life balance and rest quality, with downstream benefits for productivity (Baltes et al., 1999). Some also note that longer shifts can make it easier to have greater shift overlap during high call volume periods and reduce the administrative burden associated with shift turnovers. Others worry that longer shifts will lead to fatigue, reducing an officer's productivity at the shift's end. Across the labor market, there is no consensus on the effect of the compressed work week. Recent systematic reviews suggest that the costs and benefits of CWW remain debated and empirically mixed (Bernstrøm et al., 2025; Harris et al., 2015).

This study evaluates the impact of a pilot program that lengthened the standard work shift for police officers. We study the effect of a large-scale pilot program – which increased the length of a police shift from 8 hours to either 10 or 12 hours – on the outputs of police activity, with an emphasis on crime prevention, proactivity (as measured using stops and summonses), citizen complaints and dispatch time to service calls. The setting is New York City where the NYPD adopted the compressed work week pilot in seven high-crime precincts, replacing the eight-hour schedule that has governed patrol operations for decades.

The pilot originated from the tentative contract agreement between New York City and the Police Benevolent Association (PBA), the city's primary union for police officers, in April 2023. Proponents framed the new schedule as a quality-of-life improvement designed to address declining officer retention (New York City, 2023). By allowing officers to work the same number of total hours over fewer days, the arrangement promises more time with family and a reduced commuting burden, an attractive proposition in a large and high-cost city where many officers commute an appreciable distance.

From a public safety perspective, the extended shifts enable overlap between consecutive tours, creating "surge hours" during peak crime periods without additional

hiring. This logic is consistent with a well-established criminological literature finding that the increased availability and presence of police officers reduces crime – whether at the citywide level (Chalfin, 2025; Chalfin et al., 2022) or concentrated on high-crime street segments (Braga et al., 2019; Weisburd, 2021). If the policy performs as intended, the shift to a compressed work week therefore has the potential to constitute a genuine win for both officers and the public.

Whether it does so in practice is an open question. Extended shifts introduce countervailing risks rooted in the potential for longer shifts to generate fatigue, a feature that may erode the cognitive and physical capacities most critical to effective policing. Research on occupational fatigue suggests that prolonged work hours can impair judgment, slow reaction time and reduce tolerance for ambiguity, the capacities most taxed by high-stakes encounters requiring calibrated use of force (Vila, 2006). The reliance on motor vehicles for patrol and call response compounds these risks. For example, drowsy driving is already thought to be prevalent among police officers and is considered comparably dangerous to drunk driving (Rajaratnam et al., 2011). Over longer time horizons, sustained exposure to extended shifts may accelerate burnout, which has itself been linked to elevated use-of-force in some research (Burke & Mikkelsen, 2005; Kop et al., 1999).

These concerns are further complicated by overtime. Officers routinely work beyond their scheduled hours to complete arrest paperwork and other post-shift administrative tasks such as court appearances (Vila et al., 2002, p. 13), meaning that an agency's nominal shift length may understate actual time on duty. In an audit, the Office of the Inspector General for the NYPD found that when officers work overtime, the probability of appearing in a lawsuit, receiving a substantiated civilian complaint, experiencing a vehicle collision or being involved in a threat, resistance or injury incident is significantly elevated (Strauber & Barrett, 2023). Whether those associations reflect the disruption of an unexpected tour extension, the independent effect of cumulative hours worked or is simply a spurious result remains unclear. The answer has direct bearing on how the findings generalize to a regime in which longer hours are the scheduled norm rather than the exception, as in the case in this research.

That these concerns have not been resolved is evident in the pace of the NYPD pilot's expansion. As of May 2026, only eleven of the NYPD's 78 precincts are operating under the 10/12-hour schedule and no comprehensive evaluation of the policy's effects has been conducted over the three-plus years since implementation. This study endeavors to fill that gap by tracking the trajectory of key policing outcomes following the shift change and recovering aggregated estimates of its overall impact. With respect to the broader literature, we note that most of the papers studying the CWW in policing are now quite old, with a

majority being published over thirty years ago. We extend the literature, capitalizing on a recent pilot in the largest police department in the United States and also bringing to bear the most modern econometric methods, including contemporary staggered differences-in-differences estimators.

We find that the adoption of 10- and 12-hour patrol shifts had little effect on most measures of police proactivity, officer conduct or public safety. We find that the compressed schedule did not significantly affect civilian complaints or measures of police proactivity. While some specifications suggest a modest increase in crime, particularly misdemeanor offenses, these estimates are sensitive to the treatment of pre-trends and reference periods and should be interpreted with caution. The only consistent change that we observe after the implementation of the CWW is a significant increase in dispatch time, particularly at the end of a longer shift, a finding that is not explained by overall changes in precinct-level staffing. Taken together, the results suggest that compressed work schedules can substantially alter officers' working arrangements without generating large changes in many of the observable outputs of policing.

2. Literature Review

The compressed work week was originally proposed as a "revolution" to the standard Monday-through-Friday, 9-to-5 schedule that most industries follow (Poor, 1970). The CWW has found particular traction in occupations that already require shift work, including healthcare, industrial production and emergency response (Bambra et al., 2008; Bernstrøm et al., 2025). Workers in these fields rarely enjoy conventional weekend rest or a regular circadian alignment in the first place, which weakens the typical concern that compressed schedules will misalign workers with the social and institutional rhythms of family and community life that most people follow. For services requiring 24/7 continuous coverage, the CWW also reduces the number of within-day handovers, potentially reducing communication errors and improving service continuity (Bernstrøm et al., 2025, p. 602).

Within policing specifically, the CWW is not a recent idea. Stenzel and Buren (1986) document that articles advocating four-day, 10-hour workweeks appeared in *Police Chief* magazine as early as 1972, generating particular interest among departments on the U.S. West Coast (p. 72). The major arguments offered then to support the CWW largely resemble those made today – that a compressed schedule reduces the cumulative amount of time that officers spend commuting between home and their place of work thus freeing longer continuous blocks for family, leisure and personal obligations (Pierce & Dunham, 1992; Stenzel & Buren, 1986; Vila et al., 2002).

The mechanical effect of the CWW on time off is substantial. In a study of the Saanich and Vancouver police departments in Canada, Cunningham (1982) found that, relative to a standard 8-hour schedule, a 12-hour schedule nearly doubled the number of days off over the course of a year and increased the number of weekends with at least one day off by 85%, while a 10-hour schedule increased annual days off by 50% and weekends with at least one day off by 200%. At least three studies find that the CWW significantly increases the time officers report being able to spend on family, friends and personal life without interruption from work obligations (Cunningham, 1981; Totterdell & Smith, 1992; Vega & Gilbert, 1997), with corresponding improvements in self-reported quality of life and job satisfaction (Amendola et al., 2011). This finding mirrors the most consistent result from CWW research in other sectors (Bambra et al., 2008; Bernstrøm et al., 2025; Facer & Wadsworth, 2008).

These quality-of-life benefits help to explain an empirical regularity in the policing literature – that officers appear to overwhelmingly prefer the CWW to the traditional 8-hour schedule (Peacock et al., 1983; Richbell et al., 1998; Stenzel & Buren, 1986; Vila et al., 2002, p. 14). Adoption has also been found to be associated with improvements in departmental morale (Cunningham, 1981; Richbell et al., 1998; Stenzel & Buren, 1986) and improvements in the cohesion of work teams (Cunningham, 1981). Viewed in this light, the inclusion of the 10/12-hour pilot in the 2023 contract agreement can be plausibly understood as a meaningful win for the PBA, reflecting officer preferences that the research literature has documented for decades.

2.1. On-Shift Fatigue as the Major Concern of Longer Work Hours

Despite the popularity of the CWW among officers, the most persistent concern surrounding its adoption in police departments is that longer shifts may produce greater fatigue while officers are on duty (Vega & Gilbert, 1997, p. 392; Vila et al., 2002, p. 14). Some observers take this a further step and suggest that the CWW may not be suitable for jobs that are stressful and demanding of vigilance due to safety concerns (Armstrong-Stassen, 1998, p. 111), a description which closely matches the characteristics of police work. Over longer time horizons, these effects may erode officers' physical and cognitive performance and contribute to adverse health outcomes (Vila et al., 2000), potentially offsetting the short-term quality-of-life benefits documented above. The empirical record on this question, however, is mixed.

Several studies report a favorable effect of CWW on rest and recovery. Two earlier studies found that officers reported longer sleep duration and improved sleep quality after a transition to compressed schedules (Peacock et al., 1983; Totterdell & Smith, 1992). Pierce and Dunham (1992) similarly found that one year after adoption of a 12-hour schedule,

officers reported significantly lower levels of stress and fatigue relative to their experience under the 8-hour schedule. More recently, Amendola et al. (2011) found that officers on both 10- and 12-hour schedules reported longer average sleep hours than those on a traditional 8-hour schedule, with the 10-hour difference reaching statistical significance. On the other hand, officers on the 12-hour schedule reported significantly higher sleepiness and lower alertness during the shift than their 8-hour counterparts.

A possible resolution to this contradiction comes from the recent work of Casjens et al. (2022). That paper shows that a simple comparison of average sleep duration between 12-hour and 8-hour workers yields no significant difference among construction workers but that partitioning the sample into workdays and work-free days reveals two different pictures. Workers on compressed schedules sleep substantially less on workdays (averaging 5 hours) and compensate by sleeping more on rest days, recovering accumulated sleep debt from the longer shifts. This "social jetlag" pattern suggests that aggregate sleep totals can conceal meaningful within-week variation that is directly relevant to on-the-job fatigue.

A broader methodological concern complicates the interpretation of findings based on self-reported measures of fatigue, sleep duration and alertness, all of which are vulnerable to reporting biases. The concern is particularly acute in the context of the CWW because the policy is generally popular among officers who may have a vested interest in reporting favorable outcomes in order to preserve the preferred schedule. Social desirability bias and expectation effects have therefore been flagged as recurring threats to inference across this literature (Amendola et al., 2011, p. 413; Bell et al., 2015, p. 331; Peacock et al., 1983, p. 491; Rohwer et al., 2022; Totterdell & Smith, 1992, p. 149). The quasi-military culture that is commonly seen in police organizations can also potentially discourage complaints about long working hours (Vila et al., 2002, p. 7).[1]

In response to the limitations of self-reported data, several studies have adopted biometric measures and physical performance tests as more objective proxies for officer fatigue and alertness following the adoption of the CWW – though much of the literature is now quite old. Peacock et al. (1983) provide the most extensive battery of such measures. Comparing officers before and after a transition from an 8-hour to a 10-hour schedule, they find that the longer shift significantly reduced physiological strain, as measured by cardiorespiratory tests and lowered diastolic blood pressure. They detect no significant differences in body temperature (a proxy for circadian rhythm), epinephrine and

[1] Recent work outside of policing reinforces this concern. Studying construction workers, Mühl and Korunka (2024) show that workers' a priori expectations about the CWW significantly moderate post-implementation changes in self-reported work-life balance, fatigue, and pressure and that the policy can even produce negative outcomes among individuals who initially held low expectations. Hyatt and Coslor (2018) also report similar expectation-driven moderation in their study.

norepinephrine levels (proxies for stress), response time as measured by critical flicker fusion frequency, or performance on a grammatical reasoning task. Amendola et al. (2011) similarly find no significant differences in involuntary eye movements, sustained attention, or response time across officers assigned to 8, 10, and 12-hour shifts, despite the differences in self-reported fatigue and sleep quality discussed earlier.

2.2. Downstream Behavioral and Organizational Outcomes

A smaller but more policy-relevant body of work moves past individual-level proxies for fatigue to consider the downstream behavioral and organizational consequences of the CWW. For decisionmakers weighing whether to adopt a compressed schedule, the most consequential question is not whether officers feel more or less fatigued in the abstract, but whether shift-induced changes in the functioning of officers translate into measurable differences in proactivity, conduct or the quality of policing.

2.2.1. Proactivity and Efficiency

A priori, the expected direction of any proactivity effect is ambiguous. Fatigue from longer shifts could lead officers to slow their pace of work or reduce discretionary activity in order to preserve energy across an extended tour. On the other hand, the well-documented improvements in work-life balance and morale, along with a possible "honeymoon effect" early in a rollout of the CWW, could plausibly increase individual effort, at least in the short term (see Amendola et al., 2011; Cunningham, 1982). All in all, the available empirical evidence does not strongly favor either mechanism.

Studies that measure the volume of officer outputs tend to find little change. Vega and Gilbert (1997) report no difference in felony or misdemeanor bookings after a department extended its shifts from 8 hours to 13 hours and 20 minutes. Amendola et al. (2011) similarly find little change in the aggregated count of police-initiated activities, including arrests, citations, reports, and stops, following a transition to 10- or 12-hour schedules.

In contrast, there is some evidence that compressed schedules improve the efficiency with which officers handle their workload which, in turn, frees up more discretionary time for proactive activities. Vega and Gilbert (1997) find that officers working a three-day, 13.3-hour schedule significantly reduced their response time, handling time and overall committed time per call, while simultaneously responding to a larger total number of dispatched calls. The reduction in committed time allowed officers to spend more of their tour on uncommitted patrol (see also Richbell et al., 1998). Cunningham (1982) reports a complementary finding. Under the CWW, officers were able to perform a wider range of activities within a single shift, including traffic enforcement, criminal investigations, and patrol, than was feasible under an 8-hour schedule. In one department, officers reported

significantly more time spent on community programs and professional development under the 10-hour schedule (Cunningham, 1981).

Beyond individual-level efficiency gains, longer shifts may also provide departments with greater leeway to allocate human and financial resources in ways that the less flexible 8-hour schedule does not permit. The most frequently cited example is the creation of overlapping shifts between successive squads or platoons, in which one shift is winding down as the next begins (Alexander Weiss Consulting, 2015; Barnum, 2011; Richbell et al., 1998; Stenzel & Buren, 1986; Totterdell & Smith, 1992). These overlap periods, commonly referred to as "power shifts" (Barnum, 2011; Stenzel & Buren, 1986), temporarily augment the number of on-duty officers available to respond to calls for service or conduct patrols. When deliberately aligned with the times of day at which a department experiences its peak service demand or its highest crime volume (Alexander Weiss Consulting, 2015; Barnum, 2011), the resulting surge has the potential to deliver a concentrated dose of police presence precisely when it matters the most. This logic dovetails with an empirical regularity in criminology and policing research on the effectiveness of targeted deployment at the hottest spots and times (Braga et al., 2019; Sherman et al., 2014; Weisburd et al., 2016).

Officers approaching the end of their tour can also use the overlap period for report writing, leaving better-documented handover materials for the incoming shift (Richbell et al., 1998; Vega & Gilbert, 1997, p. 401). The arrangement may further blunt an incentive problem near shift boundaries –that officers at the end of their shift have reason to avoid arrests that would generate paperwork extending into overtime hours, a motive that is weaker among officers who have just begun their tour. Here, recent research by Chalfin and Gonçalves (2026) and Olson and Adams (2026) has found strong evidence that officers in tend to reduce their arrest activity at shift's end, with an explanation that is more consistent with a distaste for unscheduled overtime work than with fatigue.

Nonetheless, some authors have warned that the benefits of a power shift can create the impression that additional staffing during peak hours arrives without cost, when in fact it is mechanically obtained by drawing officers away from other parts of the day (Alexander Weiss Consulting, 2015, p. 30; Barnum, 2011; Stenzel & Buren, 1986). In a department that is already operating under severe short staffing, constructing a power shift could potentially amount to little more than robbing Peter to pay Paul with non-overlap hours pushed below functional staffing levels in exchange for concentrated coverage at peak times. Equipment constraints can compound the problem. Resources that are typically shared across shifts, particularly motor vehicles, are repeatedly cited as a binding constraint that prevents departments from realizing the full operational potential of overlapping schedules (Richbell et al., 1998; Stenzel & Buren, 1986).

The case for power shifts also hinges on the degree of active alignment that is not always achieved in practice. Without deliberate planning, overlap periods can drift away from peak-demand hours and end up consolidating coverage during genuinely off-peak windows (Alexander Weiss Consulting, 2015, p. 31). Using the degree of alignment between staffing levels and hourly service demand as a measure of schedule efficiency, Barnum (2011) finds that a compressed 12-hour schedule adopted by the Cedar Rapids Police Department in fact produced worse efficiency than the original 8-hour schedule it replaced. Even when initial alignment is achieved, sustaining it requires continuous monitoring and revision (Richbell et al., 1998, p.385). Hot times are more volatile than hot spots (Deckard & Schnell, 2023) and more prone to displacement (Guerette & Bowers, 2009, p. 1346), and their temporal dynamics may shift faster than either rigid shift structures or officers' circadian patterns can accommodate (Madia, 2022).

With respect to financial solvency, reducing overtime payments is another major consideration for police organizations facing budget constraints (Stenzel & Buren, 1986). The mechanism is straightforward. Given that police work often lasts longer than the usual 8-hour tour to complete, extending the tour to a longer period should naturally reduce the frequency of overtime. The strongest empirical evidence so far shows that officers on 8-hour shifts work significantly more overtime than those on 10- or 12-hour shifts (Amendola et al., 2011; Cunningham, 1982; Richbell et al., 1998), while extending the 10-hour shifts to 13:20 hours does not make a difference (Bell et al., 2015). Indeed, there are concerns associated with an increase in overtime associated with obliged court appearances with CWW (Cunningham, 1981; Vila et al., 2002).

2.2.2. Public Safety and Conduct

A parallel concern in the CWW literature centers on whether any fatigue that accumulates under longer shifts ultimately translates into elevated risks for officers or the public at large. Two channels have received the most attention. The first consideration is motor vehicle safety (Stenzel & Buren, 1986, p. 72; Vega & Gilbert, 1997; Vila et al., 2002). Officers spend much of their tour driving for patrol and emergency response, often at night, and drowsy driving has long been identified as a particular hazard in this population (Rajaratnam et al., 2011; Vila et al., 2002). The second consideration is the quality of interactions between police officers and citizens with the specific concern being that longer shifts may erode an officer's temper, patience or professionalism in citizen encounters (Vega & Gilbert, 1997; Vila et al., 2000, 2002).

Empirical support for these concerns within the CWW literature is scarce. Vega and Gilbert (1997) report no significant change in work-related injuries, citizen complaints, or disciplinary actions following the adoption of a 13-hour-20-minute shift. Amendola et al.

(2011) find no differences in driving or shooting simulator performance across officers assigned to 8-, 10-, and 12-hour shifts. Bell et al. (2015) also show no difference in police-involved vehicle accidents or higher failure rates on annual shooting qualification tests among officers with shifts of 10 hours or 13:20 hours. An experimental study by James et al. (2017) reports that fatigue accumulated over five consecutive 10-hour-40-minute shifts did not impair officers' reaction time in shoot/no-shoot decisions in a deadly force judgment simulator, and that the schedule produced no interaction effects with suspect race.

A notable exception comes from a recent report by the New York City Department of Investigation on overtime and adverse policing outcomes in the NYPD (Strauber & Barrett, 2023). Using detailed administrative records on individual officers' overtime and incident histories, the report shows that a 4-hour-12-minute overtime extension is associated with 37%, 37%, 21%, and 19% higher odds of receiving a substantiated citizen complaint, being named in a lawsuit, being involved in a threat, resistance, or injury incident, and sustaining a workplace injury on the following day, respectively. The study, however, is not designed as an evaluation of CWW and does not claim causal identification. As a result, it remains unclear whether the documented associations reflect the effect of long working hours per se, which would generalize to a regularized 12-hour schedule, or the disruptive effect of an unexpected schedule extension, which would not.

3. Institutional Setting

The NYPD entered the 2020s facing a staffing crisis with few modern precedents. In the wake of George Floyd's murder and the protests that followed, officer departures accelerated sharply. Retirements rose 72% in 2020 relative to 2019, and resignations continued to climb in subsequent years, reaching their highest level in two decades by the end of 2022 (Quick & Wolff, 2024). Between January 2020 and late 2022, approximately 9,400 officers left the department while only 6,900 were hired, producing a net loss of roughly 2,500 sworn personnel (Marcius, 2022). Officers were drawn to suburban and out-of-state departments that offered competitive salaries, lower cost of living and lighter workloads. NYPD union president Patrick Lynch described the dynamic as other communities "recognizing the talent and poaching our members" (Marcius, 2022). The attrition was compounded by historically low starting salaries. As of 2022, a new NYPD recruit earned $42,500 per year, a figure that placed the department well below many suburban agencies, especially when adjusted for the City's high cost of living.

Those who remained reported widespread dissatisfaction. In a survey of 1,823 NYPD officers conducted in mid-2023, Quick and Wolff (2024) found that nearly one quarter of respondents expressed an active intention to leave the department or the profession in the

near term. The strongest predictors of one's intention to leave the department were dissatisfaction with work-life balance, being overworked and financial compensation. Officers who intended to leave were far more likely to report being overworked (46% vs. 26%) or unable to get time off (33% vs. 19%). Poor work-life balance showed a similar gap (44% vs. 24%). Mandatory overtime, the department's default response to understaffing, exacerbated these grievances. Officers were routinely ordered to extend their shifts to cover vacancies, eroding the predictability of their schedules and intensifying the very conditions driving attrition (Hilal & Litsey, 2020; Quick & Wolff, 2024).

It was in this context that the compressed work week emerged as a central feature of the April 2023 contract agreement between the City of New York and the Police Benevolent Association. The eight-year agreement (covering 2017–2025) included substantial pay raises alongside a pilot program for 10- and 12-hour patrol tours, the first formal departure from the NYPD's long-standing 8-hour shift structure (New York City, 2023). PBA president Patrick Lynch called the schedule change "a historic moment" and described the new tours as "the real solution" for officer mental health and quality of life (Katersky, 2023). Mayor Eric Adams framed the agreement as a means to give officers greater scheduling flexibility, strengthen morale, and maintain public safety (New York City, 2023).

Prior to the pilot, NYPD patrol operations were organized around a three-platoon system in which officers worked tours of approximately 8 hours and 23 minutes across a five-day workweek, accumulating 249 annual appearances. The three platoons covered the 24-hour cycle with shifts running approximately from 0700 to 1535 (day), 1450 to 2330 (evening), and 2305 to 0802 (midnight), with modest overlaps at the boundaries to facilitate shift handovers. Officers rotated through these platoons on a fixed schedule, alternating between two and three consecutive days off.

The pilot replaced this structure with two alternative schedules. Officers on the 12-hour chart work three consecutive days followed by three consecutive days off, producing 182 annual appearances, roughly 67 fewer than under the traditional chart. The 10-hour variant runs four consecutive days on followed by two or three off (in rotating or steady configurations), yielding 208 annual appearances, approximately 41 fewer than the baseline. Under both schedules, total annual work hours remain equivalent to those under the 8-hour chart; the reduction in appearances is offset by longer daily tours. Tour preferences are assigned by seniority, with commanding officers retaining final authority over assignments (New York City Police Department, 2023). Every pilot precinct carries a mix of 8-hour, 10-hour, and 12-hour tours.

The 12-hour schedule produces two shifts per day. The day tour runs from 0600 to 1800 and the night tour from 1800 to 0600. The 10-hour schedule layers additional tours whose

start and end times create deliberate overlap during periods of peak service demand. These overlap windows, referred to internally as "surge hours," concentrate additional patrol staffing in the late afternoon and evening periods when call volume and crime are at their highest. In the pilot precincts, the primary surge windows fall within the intervals [1500, 1800) and [1900, 0100), where incoming and outgoing tours overlap for two to three hours. This design was an explicit operational objective of the pilot.[2] By lengthening shifts, the department could generate periods of augmented street-level presence at peak demand without additional hiring or overtime expenditure (New York City, 2023).

The pilot was initially deployed in two Bronx precincts, the 45th and 47th, beginning on May 3, 2023 (Cohort 1). A second wave expanded the program to the 43rd and 49th Precincts on August 16, 2023 (Cohort 2), and a third wave added the 42nd and 44th Precincts along with Midtown South (the 14th Precinct) on January 31, 2024 (Cohort 3). All seven precincts are among the city's highest-crime commands, consistent with the department's stated rationale that denser call environments would magnify the gains from surge staffing. Several transit districts and police service areas also adopted the new schedule, but because these units operate under restricted geographic jurisdictions and perform more specialized functions than patrol precincts, they are excluded from the analysis. Once a precinct entered the pilot, the 10/12-hour schedule was maintained continuously; no precinct reverted to the 8-hour chart during the study period.

The timing of the intervention is defined based on the dates on which the 10/12-hour shift pilot took effect in each precinct, as specified in three waves of Operational Orders issued by the NYPD. This produces three distinct treatment cohorts (precincts that received the intervention at the same time), summarized in Table 1. Because the unit of analysis is the precinct-by-month, the calendar month in which a precinct's start date falls is coded as its first treated month—with one exception: the third cohort's pilot began on the final day of January 2024, so February 2024 is taken as its first treated month to avoid a month in which the new schedule was effectively not in operation. Once a precinct enters the pilot, the 10/12-hour schedule is maintained continuously. No precinct reverted to the 8-hour schedule during the study period. Overall, the analysis will be based on the panel data with 72 precincts and 44 months (3388). Table 2 provides the summary statistics by treatment status for the outcome variables.

Although several transit districts and police service areas are also subject to the new shift policy, they are excluded from the analysis. These units operate under restricted geographic jurisdictions and perform more specialized functions than patrol precincts,

[2] This information is based on personal communication between the second author and NYPD personnel involved in the pilot's implementation.

making them poor comparators for the precincts of primary interest. Patrol precincts, which account for the large majority of recorded incidents across all outcome categories, are the relevant unit of analysis. [3]

4. Data

The outcome variables are drawn primarily from NYC's Open Data portal – a repository for publicly available police and crime data – supplemented by one additional source. The data cover the period from May 2022 through December 2025. All records are aggregated to the precinct-by-month level.

Crime is measured using NYPD's Complaint Data, which captures crime incidents reported to law enforcement in NYC. In our primary analysis, we restrict the sample to outdoor incidents because street-level police presence is most able to affect crimes that occur in publicly accessible spaces. Proactive enforcement activity is captured through two sources: Stop, Question, and Frisk data, which contains information on pedestrian stops and the NYPD OATH Summons data, which records summonses issued for quality-of-life violations. We aggregated these two datasets into a single variable, in order to capture officer proactivity while limiting the number of statistical tests we conducted. Officer behavior and accountability are measured using two Civilian Complaint Review Board datasets: one that records the number of citizen complaints and a second that records the allegations associated with each complaint, which are linked on incident identifiers to map complaint timing to allegation counts.

Dispatch time is acquired from the NYPD's Call for Service dataset. It is calculated using the time elapsed between when the call was added to the system and when the police were dispatched for response.[4] The data further allows us to identify more urgent calls, that is, crime in progress (CIP), using the tags "Non Critical", "Critical", or "Serious." Arrival time is calculated by subtracting the time from dispatch to arrival at the scene.[5] Since the officers are only required to record arrival times when it's a CIP call, for response time, we only show the results calculated from CIP calls.

[3] One administrative adjustment was necessary to maintain consistency in precinct boundaries across the study period. Precinct 116 was created on December 18, 2024, through a subdivision of Precinct 105. All records subsequently assigned to Precinct 116 are recoded to Precinct 105 to preserve a stable geographic unit throughout the panel.

[4] All records with a radio code "10-75" are excluded, as it represents patrol assignments or directed deployment that have almost identical added time, dispatch time, and arrival time.

[5] Still, approximately 10% of CIP calls lack a recorded arrival time. Among non-CIP calls, about 30% have missing arrival times each year.

The analysis includes all NYPD precincts except those in Staten Island (the 120th, 121st, 122nd, and 123rd) and Central Park (22nd), though Staten Island implemented the CWW on October 9th, 2024. We exclude these precincts because they differ markedly from the remainder of the City with respect to workload, size and community characteristics. Staten Island also began the CWW later, so including its precincts would shorten how far forward we can trace post-treatment effects (13 months) while keeping the post-treatment observation window balanced against precincts that implemented the CWW earlier. Because these precincts carry lower caseloads, we judge this tradeoff to be unfavorable.

5. Empirical Framework

We leverage the staggered adoption of the compressed work week across the seven precincts that eventually adopted the compressed work schedule, which was rolled out across three distinct waves. The core identifying assumption is that, absent the policy change, treated precincts would have followed parallel outcome trajectories to those that had not yet adopted the new schedule. In the past, the traditional starting point for studying staggered adoption designs was the traditional two-way fixed effects (TWFE) estimator in which outcomes are regressed on a binary treatment indicator alongside unit and time fixed effects. However, as Goodman-Bacon (2021) has demonstrated, the TWFE estimator identifies a treatment effect that is potentially biased due to the use of arbitrary and potentially negative variance weights and "forbidden comparisons" in which already-treated units serve as controls for later-treated ones (de Chaisemartin & D'Haultfœuille, 2023). When treatment effects are heterogeneous across cohorts, these comparisons can produce estimates that are biased or, at a minimum, uninterpretable as a treatment effect of interest.[6]

This recognition has motivated a methodological literature that has proposed estimators that address forbidden comparisons and arbitrary weights by modeling cohort-specific treatment effects explicitly and ensuring that the comparison group at each time period consists only of units not yet exposed to treatment (Callaway & Sant'Anna, 2021; Sun & Abraham, 2021; Wing, Freedman, et al., 2024; Wooldridge, 2025). Given the wide array of approaches, we present estimates from the stacked DiD proposed by Wing, Freedman, et al. (2024), the estimator of Sun and Abraham (2021) (SA), and the estimator of Callaway and Sant'Anna (2021) (CS). We prioritize the basic stacked DiD estimator of Wing et al (2024)

[6] In the most extreme case, where every group eventually receives treatment, and the size of the effect gets smaller with each new cohort, TWFE estimators can even produce an estimate pointing in the opposite direction of the true average treatment effect.

as our primary estimator due to its simplicity and report alternative estimators in the appendix.[7]

5.1. Stacked Staggered DiD Estimator

The way that stacked staggered DiD estimators overcome the problem of forbidden comparisons is by sub-setting each of the treatment cohorts into a "clean" dataset that only contains a members of single treatment cohort alongside the other units that are either not yet treated within this window or which are never treated (Wing, Yozwiak, et al., 2024). Each of these subsets generates a clean 2-by-2 DID estimate. Consider the following equation:

$$y_{itj} = \alpha + \beta_{stk} LongShift_{it} + \phi_{ij} + \lambda_{tj} + \epsilon_{itj}. \quad (5\text{-}1)$$

In Equation (5-1), i denotes the precinct, t indexes the year-month, and j is the indicator for each treatment cohort. y_{itj} is the policing outcome of interest for examining potential change after the CWW schedule, including call-for-service dispatch time, crimes, proactivity measured by stops and summons, and complaints. ϕ_{ij} and λ_{tj} are subexperiment-specific precinct fixed effects and month fixed effects. β_{stk} pools each of the clean DID estimates from each sub-experiment using variance weighting (Baker et al., 2022), whereby sub-experiments with a larger variance in treatment status have their treatment effect weighted higher than those with little variance. In our case, the highest weight will be given to the cohort with 3 precincts that adopted the 10/12 shift at the same time. We fix the observation window to be $\kappa_a = 12$ and $\kappa_b = 22$ (35 periods in total, including the starting month 0) for each of the three cohorts shown in Table 1. The stacked dataset therefore has 7,070 observations.[8] Since the same control precincts appear multiple times in the stacked dataset and the same precinct appears in multiple periods, the standard errors are clustered at the precinct level.

The stacked staggered DID can also be estimated in the form of an event study framework,

$$y_{itj} = \alpha + \sum_{s \neq -1} \delta_s \cdot 1\{R_{it} = s\} + \phi_{ij} + \lambda_{tj} + \epsilon_{itj}. \quad (5\text{-}2)$$

where s is the time relative to the treatment time; R_{it} is defined for each of the treated precincts. δ_s is the non-parametric measurement of the treatment effect corresponding to the time to treatment.[9]

---

[7] Estimates from TWFE regression models are presented in Appendix for completeness.

[8] $\underbrace{(65+2)\times 35}_{\text{cohort 1, 2}} \times 2 + \underbrace{(65+3)\times 35}_{\text{cohort 3}}$.

[9] To ensure comparability with the SA estimator, the reference period is fixed to be -1, which also includes those that were never treated.

The pattern of the $\delta_{s<-1}$ terms serves two diagnostic purposes. First, it allows us to address the parallel trends assumption required to identify the average treatment effect within Rubin's potential outcomes framework (Sun & Abraham, 2021). Second, this is a partial check of the strict exogeneity assumption, $E[\varepsilon_{itg} \mid X_{i1}, \dots, X_{iT}, \phi_{ig}] = 0$, required by fixed-effects models for consistent estimation. Because treatment status is deterministic conditional upon cohort assignment, a nonzero $\hat{\delta}_{s<-1}$ would suggest that the timing of cohort adoption is informed by pre-treatment shocks to y_{it}, which would imply selection into treatment based on the outcome and violate exogeneity.

This diagnostic is incomplete for strict exogeneity. It cannot examine the post-treatment residuals in which treatment effects and confounding shocks are not separately identified from $\delta_{s\geq 0}$ alone. We therefore invoke the standard DiD assumption—that no other policy changes coincided with the rollout of the longer shift schedule. This is plausible under staggered adoption, since the assumption requires unobserved confounding shocks to occur at the exact timing of all three cohort treatment dates simultaneously, a scenario that becomes less plausible as the number of distinct treatment timings grows.

Equation 5-1 and 5-2 are modeled with Poisson regressions for crimes, the proactivity index, and complaints, and with linear regressions for the timing variable. The estimations are implemented in R using the `fixest` package with standard errors clustered at the precinct level.

5.2. Staggered Triple Difference (DDD) Estimator

The stacked DiD estimator identifies the overall effect of the compressed work week on policing outcomes at the precinct level. While this is the estimand that corresponds most closely with the policy question of interest, as noted in Sections 2.2 and 3, the implementation of a 10/12-hour CWW typically alters more than a single dimension of police operations. A single aggregated estimate of the post-CWW change mixes the potential staffing-driven effect during shift-overlap hours with the effect driven by a long shift.

In the NYPD, as in many departments that have adopted CWW, the creation of overlapping shifts during peak-demand hours is itself an explicit goal of extending the original shift length. These overlap windows, referred to internally as "surge hours," fall within [15:00, 18:00) and [19:00, 01:00), the periods of highest service demand and which have the longest dispatch times (see Figure 1 and Figure 5). At the same time, the operational consequences of a prolonged shift are likely to concentrate disproportionately in its final hours rather than its first eight, since the latter closely resemble the schedule officers had already adapted to. This expectation is consistent with prior evidence that officer alertness deteriorates and sleepiness increases with time into shift (Amendola et al., 2011; Peacock et

al., 1983). The two mechanisms operate through different channels and may push outcomes in different directions.

In order to open up the black box, we identify the hours most relevant to each of these two different channels. Firstly, we partition the day into the 24 hourly bins. $U = \{0,1,2,\dots,22,23\}$, where each $h \in U$ denotes the one-hour interval $[h{:}\,00,\ (h+1){:}\,00)$. The final three hours of the two long shifts under the 10/12-hour schedule correspond to 03:00–06:00 and 15:00–18:00 and are given by

$$A = \{3,4,5,15,16,17\}.^{10}$$

The surge hours, in which the additional manpower channel is active, correspond to the clock windows 15:00–18:00 and 19:00–01:00 and are given by the bin set

$$B = \{15,16,17,19,20,21,22,23,0\}.$$

To make the measured hour-specific effects non-overlapping and as balanced as possible, we define the last three hours treatment status as A and the pure surge hours treatment status as $B \setminus \mathrm{A} = \{19,20,21,22,23,0\}$. Both include six-hour intervals. For both, the complement $U \setminus (A \cup B) = \{1,2,6,7,8,9,10,11,12,13,14,18\}$ serves as the comparison group, comprising the bins assumed to be least affected by either channel.

The estimation of these hour-specific treatment effects relies on the triple difference (DDD) estimator. The DDD and DiD models both have precinct-level assignments of treatment status. What differentiates the DDD design is the addition of a third source of variation that distinguishes, within each precinct, the hours most likely affected by the policy through different channels from those that are not, and uses this within-precinct contrast as an additional reference point. Estimation proceeds in two stages of comparison. The first compares affected to unaffected hours within the post-treatment period, separately for each of the three treatment cohorts and for the never-treated precincts. The second contrasts these within-precinct hourly differences across precincts, comparing treated precincts to never-treated precincts to recover cohort-specific ATTs.

Recent literature on staggered DID establishes that staggered triple-difference (DDD) estimators can be constructed by building directly on the machinery developed for staggered DiD (Baker et al., 2026). This correspondence holds with particular ease in our setting, where the design relies on never-treated precincts as the control group and the proportion of eligible units (hours of the day) is held constant across treatment-exposure groups (Ortiz-

[10] Officers in treated precincts work either 10-hour or 12-hour shifts, representing extensions of 2 and 4 hours, respectively, relative to the standard 8-hour baseline. To estimate the effect of the extended hours, we use 3 hours as the midpoint between these two extensions.

Villavicencio & Sant'Anna, 2025). We therefore recast the stacked specifications in Equations 5-1 and 5-2 into the following analogues, defined over all $h \in \{\text{TargetHrs} \cup (U \setminus (A \cup B))\}$:

$$y_{ithj} = \alpha + \beta_{stk.DDD} LongShift_{it} \cdot 1\{h = TargetHrs\} + \phi_{itj} + \lambda_{thj} + \tau_{ihj} + \epsilon_{ithj}, \quad (5\text{-}3)$$

$$y_{ithj} = \alpha + \sum_{s \neq -1} \delta_s^{DDD} \cdot 1\{R_{it} = s, h = TargetHrs\} + \phi_{itj} + \lambda_{thj} + \tau_{ihj} + \epsilon_{ithj}. \quad (5\text{-}4)$$

The added subscript h marks the disaggregation of the panel from a precinct-year-month structure to a precinct-year-month-hour structure, so that each observation now records the outcome y_{ithj} for a single hour within a given precinct-month. The indicator $1\{h = TargetHrs\}$ encodes the eligibility rule that separates hours treated as exposed to the CWW policy from those treated as unexposed. Consequently, $\beta_{stk.DDD}$ and δ_s^{DDD} recover the effect of CWW on y_{ithj}within the targeted hours specifically, in contrast to the precinct-wide effect estimated by the DiD specifications. The subexperiment-specific fixed effects, indexed by j, expand accordingly into the three interacted terms required by the triple-difference design: ϕ_{itj}, λ_{thj}, and τ_{ihj} denote precinct-by-month, month-by-hour, and precinct-by-hour fixed effects, respectively, each estimated within subexperiment j.

The outcomes y_{ith} examined in the DDD regressions are restricted to incidents that carry a time stamp: call-for-service dispatch time, crime, and complaints. For crimes, we estimate the DDD coefficients by Poisson regression, because subdividing the original precinct-year-month cells into hours leaves the counts sparse in some precincts. Complaints become even rarer at the precinct-year-month-hour level, so we instead model them with a linear probability specification, in which the outcome is the probability that any complaint is recorded in a given hour of a given precinct-month.

The dispatch and arrival time regressions take the average dispatch time as y_{ith} and condition on the corresponding total number of calls used to calculate that average.[11] Both quantities are estimated by linear regressions. The CIP dispatch-time regressions follow the same construction. All point estimates and their variance are computed with the `fixest` package, with standard errors clustered at the precinct level.

6. Results

Table 3 reports the estimated effects of the 10/12-hour schedule on each outcome variable across the three research designs introduced in Section 5: the precinct-level stacked DiD and the two stacked DDD specifications that isolate, respectively, the surge hours

[11] We added the denominator as a control out of the concern that the average response time during a certain hour of a month may be more sensitive to the number of calls received than the average response time calculated from the calls received in a whole month.

created by shift overlap and the last three hours of the long shifts. Figures 2 through 4 present the corresponding event study estimates, which we use both to assess the parallel trends assumption and to trace how the effects evolve after implementation.

6.1. Crimes

The stacked DiD estimates suggest a marginally significant 5.7% increase in outdoor crimes after the pilot precincts started the CWW schedule. Disaggregating by offense severity, the estimated effect is larger for misdemeanors, at 10% and significant at the 0.001 level, whereas the felony estimate is statistically insignificant despite a reasonably precise confidence interval and a magnitude comparable to that of total crime.

The event study plots in Panels A of Figure 2 show pre-treatment coefficients that drift mildly upward. While no individual pre-period coefficient is statistically distinguishable from the reference period and the placebo estimates do not formally reject parallel trends, the drift nevertheless implies that the stacked DiD estimates should be read as an upper bound on the crime response to the CWW schedule. The pre-treatment coefficients are reasonably flat for misdemeanor crime, especially before CWW implementation, while individual post-treatment coefficients hardly reach a statistically significant level throughout Panels A to C, except for the sixth month (odd ratio=1.16, 95% CI=[1.02, 1.31]) and twentieth month post-treatment for misdemeanor (odd ratio = 1.28, 95% CI=[1.05, 1.56]).

The hour-specific DDD estimates are less precise but help locate the increase within the day. During surge hours, the rate ratios are modestly above one for all three crime measures (1.064 for total crime, 95% CI [0.996, 1.138]; 1.067 for felonies [0.993, 1.146]; 1.067 for misdemeanors [0.973, 1.170]), with the total-crime estimate falling just short of conventional significance. In the last three hours of the long shifts, by contrast, the estimates are precise nulls (0.988 [0.922, 1.058]; 0.997 [0.938, 1.059]; 0.985 [0.865, 1.122]). The corresponding event studies (Panels D through I of Figure 2) show rate ratios scattered around one in the pre-period with no systematic trend, and no obvious post-treatment shift.

6.2. Proactivity and Complaints

There is little evidence that the schedule change affected officer-initiated activity or civilian complaints. The stacked DiD estimate for the proactivity index is -16.9% [0.669, 1.032], a marginally significant result against a pre-treatment mean of 107.6. No hour-specific DDD estimate is available for this outcome, which is defined at the precinct-month level. The event study in Panel A of Figure 3 is likewise consistent with a modest effect. The pre-treatment estimates are volatile, swinging between roughly -20% and +50% from the mean without a clear trend, and the post-treatment estimates fluctuate around the null.

Civilian complaints decline slightly under every specification, but none of the estimates approaches statistical significance. The event study estimates mirror the imprecision of the DiD design at this level of aggregation. In Panel B of Figure 3, the monthly coefficients are widely dispersed both before and after implementation, with no discernible pre-trend or post-treatment break, while the hour-level LPM event studies in Panels C and D remain close to zero throughout, with the post-period means nearly indistinguishable from the pre-period means. Complaints are rare events at the precinct-hour level (pre-treatment means of 0.39 and 0.24), so these designs cannot rule out modest effects in either direction, but they do rule out any sizable increase in complaints.

6.3. Response Times

Contrary to the expectations of some observers, the shift overlaps and flexibility provided by the CWW schedule did not decrease the time from call received to dispatch or from dispatch to police arrival. The pilot precincts, in fact, experienced an *increase* of 4.2 minutes, or a 26.8%, increase in the overall dispatch time from a pre-CWW mean of 15.7 minutes. This increase is not confined to calls that do not involve a crime in progress. Dispatch times for crime in progress calls rose by 1.9 minutes compared to a baseline average of 8.5 minutes. Similarly, the time from a police officer being dispatched to arrival at the scene increased 2.2 minutes, almost a 20% increase, from the pre-treatment mean of 11.3 minutes.

The event study estimates support a causal interpretation of effects on response times. The pre-treatment placebo estimates for dispatch time remain close to the level in the month before CWW implementation (Panel A, B, and C, Figure 4), suggesting no violation of the parallel trends assumption and no sign of selection into the pilot based on response times. After CWW implementation, every post-CWW estimate for overall dispatch time lies above the pre-treatment baseline: the effect became significant at the second month, dropped toward the baseline, and climbed again to roughly 5 minutes increase from month 13 onward, with multiple coefficients individually significant at the 5% level. The pattern for CIP dispatch time closely tracks the overall dispatch time but with a smaller magnitude. The plot for CIP arrival time shows an even clearer and more persistent upward trend up to 22 months after CWW's implementation, approaching a 4-minute, or 35.4%, increase in the time from dispatch to police arrival.

The two DDD designs decompose this precinct-level slowdown across the hours of the day. During surge hours, when the overlapping tours place the most officers on the clock, the DDD point estimates are slightly negative for overall dispatch time (-0.112, s.e.=1.720), CIP dispatch time (-0.695, s.e.=0.812), and CIP arrival time (-0.405, s.e.=0.534) but are far from

statistical or practical significance. As the horizontal dashed lines in Panels D, E, and F of Figure 4 show, the pre- and post-treatment event study coefficients are almost identical. Most individual estimates sit below the $s = -1$ anchor throughout, indicating that the reference month is somewhat atypical for dispatch time during surge hours.

By contrast, CWW had larger and highly significant effects on overall dispatch times during the final three hours of the extended shifts. On average, overall dispatch time increased by nearly 5 minutes from a pre-treatment mean of 13 minutes, representing an increase of approximately 37%. CIP dispatch time also increased modestly, by 1.2 minutes from a pre-treatment mean of 7.5 minutes. The event-study estimates in Panel G show a flat pre-treatment trend followed by an immediate 4-minute increase in overall dispatch time in the implementation month. This increase subsequently stabilizes at approximately 5 minutes above the pre-treatment level for the remainder of the post-treatment period.

The event study of CIP dispatch time over the last three hours shows greater volatility in the pre-trend (Panel H). Notably, there is a significant drop three months before the CWW schedule, after which the estimates increase. Therefore, the interpretation of the significant pre-post difference in means requires more prudence. For CIP arrival time, the pre-treatment trend continues after the CWW schedule without any break during the last three hours of the longer shift, which is also reflected in the precise null (0.019, s.e.=0.501).

Notably, the changes captured above are large enough to be visible without any regression adjustment. Panel A and E of Figure 5 plot the average dispatch time and arrival time by hour of the day and by CIP status, before and after CWW implementation for the treated precincts. Panel B and F of the same figure plot the same for the business-as-usual precincts before and after December 2023 (the midpoint of the study period). The two show very similar response time patterns in the pre-treatment/pre-December-2023 periods, while only the CWW pilot precincts show more discernible changes. For dispatch specifically, the increase concentrates around the last three hours of the longer shifts. The CIP arrival times slightly increased among control precincts as well, but the magnitude is much smaller than the nearly uniform upward shift in time across all hours in the CWW pilot precincts. Table A1 and Figure A1 in the Appendix provide the estimated change for overall arrival time and the event study plots for reference, which shows a similar and even more robust pattern of response slowdown.

To assess whether this pattern could instead reflect systematic changes in call-for-service demand in the CWW pilot precincts, Panels C and D of Figure 5 plot the share of calls received in each hour of the day before and after CWW implementation, and before and after December 2023 for the comparison precincts. These distributions are nearly identical across

precincts and time periods, suggesting that changes in the timing of call demand are unlikely to explain the observed increase in dispatch times.

A second potential explanation for the increase in dispatch times is that the compressed work week might have unintentionally led to precinct-level changes. In order to test this possibility, we obtained precinct-month level staffing data and studied whether the number of officers assigned to a precinct changed alongside the rollout of the CWW. The event study plot using precinct-month staffing level as the dependent variable (Figure A2, Appendix) shows no obvious change and even possibly a slight increase in staffing one year after the CWW schedule.

Taken together, despite theoretically having more available manpower during peak hours, there is no evidence that CWW leads to a better resource allocation based on service demand and thus a more rapid response. To the contrary, there is a clear pattern of general dispatch slowdown that concentrates at a time with the highest demand in the pilot precincts, and the time since dispatch to police arrival for CIP calls experienced a general increase.

6.4. Robustness Checks

To probe the robustness of our estimates, we varied two dimensions of our estimation strategy. First, we reconsider the underlying estimators used to implement the staggered DiD design. Here, we re-specified the analysis using the traditional TWFE estimator, the DiD estimator of Sun-Abraham (SA) estimator (Sun & Abraham, 2021), and the DiD estimator of Callaway-Sant'Anna (CS) to see if the estimated effects are sensitive to the specific estimator chosen.[12] Second, we vary the sample included in our analysis. (Sun & Abraham, 2021).

In the process, we encountered several instances in which the treatment effects estimated using SA and CS differ noticeably from those reported in Table 3 (the second and fourth rows of Table 4). Further investigation suggests that most of this discrepancy appears to be mechanical rather than substantive. Specifically, these estimates aggregate post-treatment effects defined relative to *the period immediately preceding treatment*, whereas the stacked DiD specification compares post-treatment outcomes with the full set of pre-treatment periods (post-pre). The distinction between these estimands is illustrated in our event-study plots: SA and CS capture the vertical distance between the red dashed line and null, whereas stacked DiD captures the vertical distance between the red and green dashed lines.

[12] Because CS does not support Poisson regression, we applied the inverse hyperbolic sine transformation to the dependent variables ($\text{asinh}(y)$), which retains zero-valued observations while allowing for an interpretation similar to percentage changes.

To facilitate a more comparable assessment, the Appendix describes how we construct a corresponding "post-pre" aggregation for SA and CA, thereby removing this mechanical difference in the choice of baseline, and why we view this post-pre aggregation as potentially more stable than the commonly used ref=-1 aggregation. For transparency, we report the original ref=-1 estimates as well.

The re-estimated DiD results are presented in Table 4. The point estimates for the eight outcome variables are similar across those in the first column of Table 3, TWFE, SA (post-pre), and CS (post-pre) estimators. TWFE estimates are apparently less similar to the other three, perhaps because the measured effects account for the change outside of the time-to-event windows, $s \in [-12, 22]$, that the other estimators are limited to.[13] SA (ref=-1) and CS (ref=-1) differ the most for crime and misdemeanor, where a significant increase is no longer seen. Figure 2 provides some clues, where none of the post-treatment event time estimates show a significant difference from $s = -1$, although the pre- and post- averages differ (the dashed horizontal lines). In contrast, the fifteen different DiD estimates of the change in response times (the last three columns) are fairly stable.

Table 5 performs the same re-estimations for all the DDD results with three-way fixed effects (3WFE) and SA.[14] The results are not substantively different from those in Table 3. The increase in overall dispatch times at the end of the longer shifts received unanimous support from all alternative estimators, ranging from 4.5 to 4.9 minutes. The increase in CIP dispatch time also finds support from 3WFE and SA (post-pre), but not SA (pre). The DDD estimates for CIP arrival time, likewise, are precisely null, suggesting that the overall increase is not concentrated around a certain time of the day.

The second check is to see whether the differences in outcomes are mostly driven by comparing the seven CWW pilot precincts to those that are fundamentally dissimilar. In our analytic samples, most of the pilot precincts are in the Bronx (Table 1) and tend to have higher baseline crime incidents (Table 2). They ranked second to fifth, twelfth, sixteenth, and twenty-second in average monthly crime from 2022 to April 2023 (one month before the first CWW pilot). We removed the 25% (18) and 50% (36) precincts with the lowest monthly average crime during this period and re-estimated all the parameters in Table 3 with stacked regressions.

[13] The forbidden comparison and negative weights do not seem to be the main driver of the discrepancy, given that we have a lot of never-treated precincts in the dataset (Goodman-Bacon, 2021).

[14] CS is not included because the did package has not yet supported DDD specifications. Also, from Table 4, it seems that the results from SA are indifferentiable from those from CA. The same may be expected for DDD regressions.

Table 6 and Table 7 present the re-estimated coefficients. The estimated increase in all crime and misdemeanors appears to be more sensitive to the composition of the precincts included in the analysis. The increase in all crime becomes insignificant after 50% of the low crime precincts are removed from the comparison group, while the increase in misdemeanor crimes survives at the 0.05 significance level. The estimated decrease in proactivity index become larger as more low-crime precincts being removed from the regression, while maintaining to marginally significant at the 0.1 level. The change in complaints remains null throughout. On the other hand, the increase in overall dispatch time and CIP dispatch time is remarkably robust to the change in precinct composition.

Table 8 further limits the analytic sample to Bronx precincts. In this restricted sample, we do not observe any significant effects on crime. All other results remain substantively unchanged including the increase in response times that we observe in our primary specification. These results suggest that the broader slowdown following the implementation of the CWW schedule is highly robust to the alternative compositions of the non-CWW comparison group.

7. Discussion and Future Research Directions

This study examines the competing claims surrounding the NYPD's pilot of a compressed schedule that extends patrol shifts from 8 hours to 10 or 12 hours. Proponents argued that overlapping shifts would place more officers on duty during high-crime "surge hours" without requiring new hires. Critics warned that fatigue from longer shifts could impair officer judgment and overall performance. Using data from 72 precincts over four years and both staggered difference-in-differences and triple-difference estimators, we estimate the effects of the schedule change on a range of policing outcomes including crime, summonses, stops, civilian complaints, and dispatch time.

The results do not support either a wholly optimistic or wholly pessimistic account of the implementation of the CWW. The schedule produced no clear change in civilian complaints and a potential decrease in officer proactivity. Our preferred estimates suggest that misdemeanor crimes may have increased but effects were both modest and sensitive to the choice of estimator and comparison group. The strongest and most consistent result concerns dispatch and response times. Calls, including crime in progress calls, waited longer for a unit to be assigned after precincts adopted the compressed schedule, with the largest delays appearing during the final three hours of longer tours. Even after the calls were assigned, the time it took for patrol officers to arrive at the scene for crime-in-progress calls increased.

Overall, the time between a call and the assignment of a response unit increased by an estimated 4 minutes, while the dispatch time and time from dispatch to arrival for crimes in progress each increased by about 2 minutes. The hourly analysis shows no improvement during the surge hours created by overlapping tours. By contrast, dispatch time increased by 5 minutes during the final three hours of longer tours, while dispatch time for crimes in progress increased by 1 minute. The basic pattern remained stable across alternative estimators and comparison samples. The hourly distribution of incoming calls also changed little, making a shift in the timing of service demand an unlikely explanation. This pattern runs against one of the pilot's central operational claims. If overlapping tours had increased patrol capacity during periods of high demand, calls should have been assigned and responded to more quickly during surge hours. We find no such improvement. The result highlights the difference between scheduled staffing and deployable capacity. A precinct can have more officers working without having more units ready to answer calls.

One possible explanation is that the pilot changed officer tours without changing the systems that supported them. Most notably, the compressed-tour pilot initially applied to rank-and-file officers but not to their sergeants. The city did not reach a separate agreement for a limited 12-hour sergeant pilot until April 2025, nearly two years after the officer pilot began. Police Commissioner Jessica Tisch stated that the separate pilot would ensure that at least some officers working 12-hour tours were supervised by sergeants working the same schedule (New York City Office of the Mayor, 2025; Tracy, 2025). Before that agreement, a long officer tour could span more than one supervisor tour. That mismatch may have complicated roll calls, briefings, relief, meal periods, and monitoring of unit availability, particularly near the beginning and end of tours. Officers necessarily spent less time working under the direct supervisor of the sergeant responsible for evaluating them.

The mixed schedule may also have placed new demands on the dispatch process. Traditional patrol operations revolved around three principal tours. The pilot layered 10- and 12-hour tours onto the remaining 8-hour tours, creating more start times, end times, and periods of overlap. Unless dispatch procedures changed in tandem, communications personnel may have had to manage more frequent changes in unit availability, more handoffs between outgoing and incoming officers, and more complicated decisions about which units could take pending calls. Dispatchers may also have held calls for incoming officers rather than assigning them to units approaching the end of a long tour.

Vehicle availability could have created a related constraint. More officers on duty do not produce more mobile units if the number of patrol cars remains fixed,[15] as early studies have

[15] This information is based on personal communication between the second author and NYPD personnel in the pilot precincts.

warned (Richbell et al., 1998; Stenzel & Buren, 1986). During periods of overlap, incoming officers may have lacked vehicles until outgoing officers returned to the precinct and transferred their cars. Roll call, equipment distribution, vehicle inspection, and relief procedures could further delay the point at which incoming officers became available for assignment. These constraints would help explain why greater staffing during surge hours did not produce faster dispatch and response, but they do not completely account for significantly slower dispatch and response, outside of possible anticipation effects from dispatchers.

The concentration of delays near the end of longer tours also points toward officer availability and end-of-tour incentives. A call assigned late in a tour can extend an officer's workday through travel, incident handling, arrest processing or report preparation. Officers approaching the end of a long tour may ration discretionary effort, becoming less willing to initiate or accept work likely to extend beyond their scheduled hours (Chalfin & Gonçalves, 2026; Olson & Adams, 2026). Dispatchers and supervisors may likewise avoid assigning calls to officers whose tours are about to end, especially when officers from the next tour are expected to become available shortly. These practices could increase dispatch time even when more officers are nominally working.

Such behavior should not automatically be characterized as gaming. It could reflect formal relief practices, informal accommodations, or a predictable response to the costs of receiving a late-tour assignment. Fatigue could contribute as well. Officers may complete existing assignments more slowly or take longer to return to service as a long tour progresses. Prior research finds mixed effects of compressed schedules on sleep, alertness, and performance, but some studies suggest that fatigue can increase during longer tours even when officers obtain more rest on their days off (Amendola et al., 2011; Casjens et al., 2022; Peacock et al., 1983). Because the dispatch clock stops when a unit is assigned, fatigue could affect the outcome only indirectly by reducing the number of units available for new calls. The final-hours pattern is consistent with this mechanism, but it cannot establish it.

Several mechanisms may therefore have operated together. Misaligned supervisor tours could weaken oversight of unit availability. Mixed officer tours could complicate dispatch and relief. A fixed vehicle fleet could prevent additional personnel from becoming additional mobile units. Fatigue or end-of-tour incentives could slow officers' return to service. The available data cannot distinguish among these explanations. Doing so would require officer and supervisor tour rosters, unit-status histories, assignment records, patrol-car availability, and information on overtime generated by late-tour calls.

These findings extend a literature that has often studied compressed schedules through the lens of officer preferences, self-reported fatigue, or physiological measures of well-being.

Prior studies generally find that officers prefer compressed schedules and often report better work-life balance, while evidence concerning activity and conduct remains mixed. Our results broadly accord with the null findings for conventional enforcement measures (Amendola et al., 2011). Notably, despite an increase in shift length, we find no increase in complaints after CWW, in contrast to the previous NYC internal research on overtime (Strauber & Barrett, 2023). However, we reveal some operational issues that aggregated activity counts can conceal, namely, how quickly the department makes resources available to the public.

The findings also suggest that implementing a compressed schedule requires more than rescheduling officers. Supervision, dispatch procedures, vehicle allocation, and tour-transition practices must adapt with the officer schedule. If they do not, the schedule may increase nominal staffing without increasing usable patrol capacity. While longer tours may provide officers with a more attractive schedule, realizing that benefit without weakening service delivery requires departments to redesign the supervisory, dispatch, and material systems that turn scheduled officers into available patrol units. Departments considering compressed tours should therefore monitor pending calls, active units, vehicle availability, supervisory coverage, and the time officers spend unavailable throughout each tour. They should pay particular attention to relief practices and assignments near the end of long shifts.

This study does not provide a complete assessment of the compressed schedule. Only seven precincts adopted it during the primary study period, and those precincts had higher crime and service demands. The staggered rollout and event studies strengthen the design, but the small number of treated precincts limits inference and leaves open the possibility of other changes that coincided with adoption. Each pilot precinct also retained a mixture of 8-, 10-, and 12-hour tours, so the estimates capture the aggregate impact of a precinct's adoption of the CWW rather than an individual officer's exposure to a longer tour.

Future research should link call records to officer tours, unit-status histories, supervisor coverage and vehicle assignments. Such data could show whether delays arise because officers remain busy longer, return to service more slowly, lack vehicles, or receive fewer assignments near the end of their tours. Research should also measure employee outcomes so that operational costs can be weighed against possible gains in morale, retention, and work-life balance. Given previous research findings (Blanes i Vidal & Kirchmaier, 2018; Briggs & Boonstoppel, 2025), future research should also track the potential long-term negative impacts of response slowdown on deterrence and clearance rates with more detailed investigation data.

On balance, the compressed schedule changed officers' working arrangements without producing broad changes in proactive enforcement activity or civilian complaints. The

evidence of increased misdemeanor crime is suggestive but not decisive. The clearest result is that dispatch slowed, especially near the end of longer tours, while the expected improvement during surge hours did not materialize.

References

Alexander Weiss Consulting. (2015). *Albuquerque Police Department Comprehensive Staffing Assessment And Resource Study*.

Amendola, K. L., Weisburd, D., Hamilton, E. E., Jones, G., & Slipka, M. (2011). An experimental study of compressed work schedules in policing: advantages and disadvantages of various shift lengths. *Journal of Experimental Criminology*, *7*(4), 407–442. https://doi.org/10.1007/s11292-011-9135-7

Armstrong-Stassen, M. (1998). Alternative work arrangements: Meeting the challenges. *Canadian Psychology / Psychologie canadienne*, *39*(1-2), 108–123. https://doi.org/10.1037/h0086799

Baker, A., Callaway, B., Cunningham, S., Goodman-Bacon, A., & Sant'Anna, P. H. C. (2026). Difference-in-Differences Designs: A Practitioner's Guide. *Journal of Economic Literature*, *64*(2), 498–557. https://doi.org/ 10.1257/jel.20251650

Baker, A. C., Larcker, D. F., & Wang, C. C. Y. (2022). How much should we trust staggered difference-in-differences estimates? *Journal of Financial Economics*, *144*(2), 370–395. https://doi.org/https://doi.org/10.1016/j.jfineco.2022.01.004

Baltes, B. B., Briggs, T. E., Huff, J. W., Wright, J. A., & Neuman, G. A. (1999). Flexible and compressed workweek schedules: A meta-analysis of their effects on work-related criteria. *Journal of Applied Psychology*, *84*(4), 496–513. https://doi.org/10.1037/0021-9010.84.4.496

Bambra, C., Whitehead, M., Sowden, A., Akers, J., & Petticrew, M. (2008). "A hard day's night?" The effects of Compressed Working Week interventions on the health and work-life balance of shift workers: a systematic review. *J Epidemiol Community Health*, *62*(9), 764–777. https://doi.org/10.1136/jech.2007.067249

Barnum, C. (2011). Efficiency in Continually Operating Public Organizations: A Case Study. *Public Personnel Management*, *40*(4), 279–292. https://doi.org/10.1177/009102601104000401

Bell, L. B., Virden, T. B., Lewis, D. J., & Cassidy, B. A. (2015). Effects of 13-Hour 20-Minute Work Shifts on Law Enforcement Officers' Sleep, Cognitive Abilities, Health, Quality of Life, and Work Performance. *Police Quarterly*, *18*(3), 293–337. https://doi.org/10.1177/1098611115584910

Bernstrøm, V., Alves, D., Houkes, I., Lillebråten, A., & Nilsen, W. (2025). The consequences of a compressed workweek: a systematic literature review. *International archives of occupational and environmental health*, *98*(7), 587–607.

Blanes i Vidal, J., & Kirchmaier, T. (2018). The Effect of Police Response Time on Crime Clearance Rates. *The Review of Economic Studies*, *85*(2), 855–891. https://doi.org/10.1093/restud/rdx044

Braga, A. A., Turchan, B., Papachristos, A. V., & Hureau, D. M. (2019). Hot spots policing of small geographic areas effects on crime. *Campbell Syst Rev*, *15*(3), e1046. https://doi.org/10.1002/cl2.1046

Briggs, S., & Boonstoppel, S. (2025). Revisiting the Effect of Rapid Response on the Frequency of Reported Crimes Using an Instrumental Variable Approach. *Justice Quarterly*, *42*(2), 236–262. https://doi.org/10.1080/07418825.2024.2360009

Burke, R. J., & Mikkelsen, A. (2005). Burnout, job stress and attitudes towards the use of force by Norwegian police officers. *Policing: An International Journal of Police Strategies & Management*, *28*(2), 269–278.

Callaway, B., & Sant'Anna, P. H. C. (2021). Difference-in-Differences with multiple time periods. *Journal of Econometrics*, *225*(2), 200–230. https://doi.org/10.1016/j.jeconom.2020.12.001 (Themed Issue: Treatment Effect 1)

Casjens, S., Brenscheidt, F., Tisch, A., Beermann, B., Brüning, T., Behrens, T., & Rabstein, S. (2022). Social jetlag and sleep debts are altered in different rosters of night shift work. *PLoS One*, *17*(1), e0262049. https://doi.org/10.1371/journal.pone.0262049

Chalfin, A. (2025). Investments in Policing and Community Safety. *Annual Review of Criminology*, *8*(Volume 8, 2025), 403–429. https://doi.org/https://doi.org/10.1146/annurev-criminol-111523-122746

Chalfin, A., & Gonçalves, F. (2026). Professional Motivations in the Public Sector: Evidence from Police Officers. *American Economic Review*, *116*(8), 2962–2995. https://doi.org/10.1257/aer.20231696

Chalfin, A., Hansen, B., Weisburst, E. K., & Williams, M. C., Jr. (2022). Police Force Size and Civilian Race. *American Economic Review: Insights*, *4*(2), 139–158. https://doi.org/10.1257/aeri.20200792

Cunningham, J. B. (1981). Exploring the impact of a ten-hour compressed shift schedule. *Journal of Organizational Behavior*, *2*(3), 217–222. https://doi.org/https://doi-org.proxy.library.upenn.edu/10.1002/job.4030020307

Cunningham, J. B. (1982). Compressed Shift Schedules: Altering the Relationship between Work and Non-Work. *Public Administration Review*, *42*(5), 438–447. https://doi.org/10.2307/975646

de Chaisemartin, C., & D'Haultfœuille, X. (2023). Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: a survey. *The Econometrics Journal*, *26*(3), C1–C30. https://doi.org/10.1093/ectj/utac017

Deckard, M., & Schnell, C. (2023). The Temporal (In)Stability of Violent Crime Hot Spots Between Months and The Modifiable Temporal Unit Problem. *Crime & Delinquency*, *69*(6-7), 1312–1335. https://doi.org/10.1177/00111287221128483

Facer, R. L., & Wadsworth, L. (2008). Alternative Work Schedules and Work–Family Balance:A Research Note. *Review of Public Personnel Administration*, *28*(2), 166–177. https://doi.org/10.1177/0734371x08315138

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, *225*(2), 254–277. https://doi.org/https://doi.org/10.1016/j.jeconom.2021.03.014

Guerette, R. T., & Bowers, K. J. (2009). Assessing the Extent of Crime Displacement and Diffusion of Benefits: A Review of Situational Crime Prevention Evaluations. *Criminology*, *47*(4), 1331–1368. https://doi.org/https://doi.org/10.1111/j.1745-9125.2009.00177.x

Harris, R., Sims, S., Parr, J., & Davies, N. (2015). Impact of 12h shift patterns in nursing: A scoping review. *International Journal of Nursing Studies*, *52*(2), 605–634. https://doi.org/https://doi.org/10.1016/j.ijnurstu.2014.10.014

Hilal, S., & Litsey, B. (2020). Reducing police turnover: Recommendations for the law enforcement agency. *International Journal of Police Science & Management*, *22*(1), 73–83. https://doi.org/10.1177/1461355719882443

Hyatt, E., & Coslor, E. (2018). Compressed lives: how “flexible” are employer-imposed compressed work schedules? *Personnel Review*, *47*(2), 278–293. https://doi.org/10.1108/pr-08-2016-0189

James, L., James, S., & Vila, B. (2017). Does the “reverse racism effect” withstand the test of police officer fatigue? *Policing: An International Journal of Police Strategies & Management*, *40*(2), 184–196. https://doi.org/10.1108/pijpsm-01-2016-0006

Katersky, A. (2023, May 3). New York Police Department launches new flexible schedules. *ABC News*. https://abcnews.com/US/new-york-police-department-launches-new-flexible-schedules/story?id=99040662

Kop, N., Euwema, M., & Schaufeli, W. (1999). Burnout, job stress and violent behaviour among Dutch police officers. *Work & Stress*, *13*(4), 326–340. https://doi.org/10.1080/02678379950019789

Madia, C. T. (2022). *Shift Work: The difference is night and day. Do fewer rotations between day and night shift reduce sick time and improve performance in police officers?* [Master's thesis, Harvard University]. https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37371414

Marcius, C. R. (2022, December 9). NYPD officers leave in droves for better pay in smaller towns. https://www.nytimes.com/2022/12/09/nyregion/new-york-police-department-attrition.html

Mastrobuoni, G. (2019). Police disruption and performance: Evidence from recurrent redeployments within a city. *Journal of Public Economics*, *176*, 18–31. https://doi.org/https://doi.org/10.1016/j.jpubeco.2019.05.003

Mühl, A., & Korunka, C. (2024). You get what you expect: assessing the effect of a compressed work schedule on time pressure, fatigue, perceived productivity, and work-life balance. *European Journal of Work and Organizational Psychology*, *33*(5), 703–711. https://doi.org/10.1080/1359432X.2024.2379061

New York City. (2023). Mayor Adams, OLR Commissioner Campion Announce Tentative Contract Agreement With PBA, Providing Fair Wage Increases for First Time in Six Years and Pilot Program for Extended Tours. *The official website of the City of New York*.

New York City Police Department. (2023). Pilot program: 10 and 12 hour tours for police officers assigned to the 45th and 47th Precincts, Police Service Area 8, and Transit District 11 (Operations Order No. 22). In.

Olson, M., & Adams, I. T. (2026). Tough Shift: The Temporal Dynamics of Police Discretion. *Journal of Quantitative Criminology*. https://doi.org/10.1007/s10940-026-09683-5

Ortiz-Villavicencio, M., & Sant'Anna, P. H. (2025). Better understanding triple differences estimators. *arXiv preprint*. https://doi.org/https://doi.org/10.48550/arXiv.2505.09942

Peacock, B., Glube, R., Miller, M., & Clune, P. (1983). Police officers' responses to 8 and 12 hour shift schedules. *Ergonomics*, *26*(5), 479–493. https://doi.org/10.1080/00140138308963364

Pierce, J. L., & Dunham, R. B. (1992). The 12-Hour Work Day: A 48-Hour, Eight-Day Week. *Academy of Management Journal*, *35*(5), 1086–1098. https://doi.org/10.5465/256542

Poor, R. (1970). *4 Days, 40 Hours: Reporting a Revolution in Work and Leisure*. Bursk and Poor Pub.

Quick, K. M., & Wolff, K. T. (2024). Mass exodus? Police officer turnover intention and organizational commitment in New York City. *Policing: An International Journal*, *48*(5), 897–922. https://doi.org/10.1108/pijpsm-06-2024-0097

Rajaratnam, S. M. W., Barger, L. K., Lockley, S. W., Shea, S. A., Wang, W., Landrigan, C. P., O'Brien, C. S., Qadri, S., Sullivan, J. P., Cade, B. E., Epstein, L. J., White, D. P., Czeisler, C. A., Harvard Work Hours, H., & Safety Group, f. t. (2011). Sleep Disorders, Health, and Safety in Police Officers. *JAMA*, *306*(23), 2567–2578. https://doi.org/10.1001/jama.2011.1851

Richbell, S., Simpson, M., Sykes, G. M., & Meegan, S. (1998). Policing with the Ottawa shift system: a British experience. *Policing: An International Journal of Police Strategies & Management*, *21*(3), 384–396.

Rohwer, E., Velasco Garrido, M., Herold, R., Preisser, A. M., Terschüren, C., Harth, V., & Mache, S. (2022). Police officers' work–life balance, job satisfaction and quality of life: longitudinal effects after changing the shift schedule. *BMJ Open*, *12*(9), e063302. https://doi.org/10.1136/bmjopen-2022-063302

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

Stenzel, W. W., & Buren, R. M. (1986). *Police Work Scheduling: Management Issues and Practices*. U.S. Department of Justice, National Institute of Justice, Office of Development, Testing and Dissemination.

Strauber, J., & Barrett, J. (2023). *Overtime Under Review:NYPD Overtime and the IncreasedRisk of Negative Policing Outcomes*.

Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, *225*(2), 175–199. https://doi.org/10.1016/j.jeconom.2020.09.006

Thompson, B. J. (2019). Does work-induced fatigue accumulate across three compressed 12 hour shifts in hospital nurses and aides? *PLoS One*, *14*(2), e0211715.

Totterdell, P., & Smith, L. (1992). Ten-hour days and eight-hour nights: Can the Ottawa Shift System reduce the problems of shiftwork? *Work & Stress*, *6*(2), 139–152. https://doi.org/10.1080/02678379208260348

Vega, A., & Gilbert, M. J. (1997). Longer Days, Shorter Weeks: Compressed Work Weeks in Policing. *Public Personnel Management*, *26*(3), 391–402. https://doi.org/10.1177/009102609702600308

Vila, B. (2006). Impact of long work hours on police officers and the communities they serve. *American Journal of Industrial Medicine*, *49*(11), 972–980. https://doi.org/https://doi-org.proxy.library.upenn.edu/10.1002/ajim.20333

Vila, B., Morrison, G. B., & Kenney, D. J. (2002). Improving Shift Schedule and Work-Hour Policies and Practices to Increase Police Officer Performance, Health, and Safety. *Police Quarterly*, *5*(1), 4–24. https://doi.org/10.1177/109861102129197995

Vila, B. J., Kenney, D. J., Morrison, G. B., & Reuland, M. (2000). *Evaluating the Effects of Fatigue on Police Patrol Officers: Final Report*.

Weisburd, D., Eck, J. E., Braga, A. A., Telep, C. W., Cave, B., Bowers, K., Bruinsma, G., Gill, C., Groff, E. R., Hibdon, J., Hinkle, J. C., Johnson, S. D., Lawton, B., Lum, C., Ratcliffe, J. H., Rengert, G., Taniguchi, T., & Yang, S.-M. (2016). *Place Matters: Criminology for the*

Twenty-First Century. Cambridge University Press. https://doi.org/DOI: 10.1017/CBO9781139342087

Weisburd, S. (2021). Police Presence, Rapid Response Rates, and Crime Prevention. *Review of Economics and Statistics*, *103*(2), 280–293. https://doi.org/10.1162/rest_a_00889

Wing, C., Freedman, S. M., & Hollingsworth, A. (2024). Stacked Difference-in-Differences. In: National Bureau of Economic Research.

Wing, C., Yozwiak, M., Hollingsworth, A., Freedman, S., & Simon, K. (2024). Designing Difference-in-Difference Studies with Staggered Treatment Adoption: Key Concepts and Practical Guidelines. *Annual Review of Public Health*, *45*(Volume 45, 2024), 485–505. https://doi.org/https://doi.org/10.1146/annurev-publhealth-061022-050825

Wooldridge, J. M. (2025). Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators. *Empirical Economics*, *69*(5), 2545–2587. https://doi.org/10.1007/s00181-025-02807-z

Table 1: Treated Cohorts.

Cohort	Affected Precincts	Starting Date	Starting Month
1.	45th, 47th.	May 3, 2023.	May, 2023.
2.	43rd , 49th.	August 16, 2023.	August, 2023.
3.	42nd, 44th, 14th (Midtown South).	January 31, 2024.	February, 2024

Table 2: Summary Statistics.

	10/12 Precincts			Never-Treated Precincts			
Variable	N	Mean	SD	N	Mean	SD	p-value
Total Crimes	308	174.896	41.991	2,860	92.907	45.983	<0.000
Felony Crimes	308	86.958	23.970	2,860	41.565	21.947	<0.000
Misdemeanor Crimes	308	74.539	19.698	2,860	42.854	22.408	<0.000
Proactivity	308	122.987	170.493	2,860	94.875	110.654	0.005
Civilian Complaints	308	29.078	22.721	2,860	18.038	20.853	<0.000
Dispatch Time	308	18.641	9.458	2,860	7.978	6.158	<0.000
Dispatch Time (CIP)	308	10.009	5.195	2,860	4.647	3.144	<0.000
Response Time (CIP)	308	13.326	4.009	2,860	10.581	3.163	<0.000
Longer Shift Indicator	308	0.620	0.486	2,860	0.000	0.000	<0.000

Table 3: Baseline Results.

Outcome		Stacked DiD	Stacked DDD (Surge Hours)	Stacked DDD (Last Three Hours)
All Crime	Estimate	1.057† [0.992, 1.126]	1.064† [0.996, 1.138]	0.988 [0.922, 1.058]
	Observations	7,070	127,260	127,260
	Pre-treatment mean	161.04	9.26	6.58
Felony Crimes	Estimate	1.042 [0.951, 1.141]	1.067† [0.993, 1.146]	0.997 [0.938, 1.059]
	Observations	7,070	127,260	127,224
	Pre-treatment mean	79.71	4.74	3.38
Misdemeanor Crimes	Estimate	1.100*** [1.042, 1.162]	1.067 [0.973, 1.170]	0.985 [0.865, 1.122]
	Observations	7,070	127,260	127,206
	Pre-treatment mean	68.04	3.94	2.70
Proactivity	Estimate	0.831† [0.669, 1.032]	—	—
	Observations	7,070	—	—
	Pre-treatment mean	107.57	—	—
Complaints	Estimate	0.947 [0.870, 1.030]	-0.007 (0.035)	-0.009 (0.026)
	Observations	7,070	127,260	106,454
	Pre-treatment mean	39.94	0.39	0.24
Dispatch Time	Estimate	4.150* (1.709)	-0.112 (1.720)	4.902*** (0.818)
	Observations	7,070	127,260	127,260
	Pre-treatment mean	15.70	17.98	12.97
CIP Dispatch Time	Estimate	1.935† (1.066)	-0.695 (0.812)	1.195** (0.359)
	Observations	7,070	127,260	127,260
	Pre-treatment mean	8.47	9.74	7.48
CIP Arrival Time	Estimate	2.235* (0.904)	-0.405 (0.534)	0.019 (0.501)
	Observations	7,070	127,260	127,260
	Pre-treatment mean	11.28	11.12	10.55

Notes: † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Cluster-robust confidence intervals in brackets for Poisson regressions. Standard errors clustered at the precinct level are in parentheses. Pre-treatment means are for treated precincts in the pre-treatment period; for the DDD columns, they are restricted to the treated hours. For DDD estimates, complaints are modeled with linear probability models that indicate the probability of having any complaints filed in a certain hour in a month in a precinct. The estimates for Poisson models are shown as rate ratios rather than raw log-count coefficients. For the timing outcomes, the coefficients are changes in minutes, and precinct-clustered standard errors are reported in parentheses.

Table 4: Staggered DiD Estimates with the other Estimators (N=3,168).

Estimator		All Crime	Felony	Misdemeanor	Proactivity	Complaints	Dispatch Time	CIP Dispatch Time	CIP Arrival Time
TWFE	Estimate	1.073†	1.045	1.129**	0.802*	0.954	4.220*	2.200†	2.530*
	CI/SE	[0.992, 1.161]	[0.936, 1.166]	[1.048, 1.216]	[0.654, 0.983]	[0.861, 1.056]	(2.035)	(1.256)	(1.226)
	N	3,168	3,168	3,168	3,168	3,168	3,168	3,168	3,168
SA (ref = -1)	Estimate	1.043	1.026	1.065	0.941	0.953	3.600**	1.951***	1.994***
	CI/SE	[0.985, 1.104]	[0.942, 1.117]	[0.925, 1.226]	[0.760, 1.166]	[0.799, 1.137]	(1.181)	(0.533)	(0.450)
	N	3,168	3,168	3,168	3,168	3,168	3,168	3,168	3,168
SA (post-pre)	Estimate	1.055†	1.043	1.094***	0.788*	0.919†	4.121***	1.925*	2.230**
	CI/SE	[0.993, 1.120]	[0.966, 1.126]	[1.039, 1.152]	[0.634, 0.980]	[0.838, 1.009]	(1.141)	(0.751)	(0.835)
	N	3,168	3,168	3,168	3,168	3,168	3,168	3,168	3,168
CS (ref = -1)	Estimate	0.036	0.026	0.052	0.062	-0.158	3.595**	1.947**	1.993***
	SE	(0.034)	(0.046)	(0.078)	(0.157)	(0.122)	(1.244)	(0.616)	(0.463)
	N	3,168	3,168	3,168	3,168	3,168	3,168	3,168	3,168
CS (post-pre)	Estimate	0.065†	0.060	0.094**	-0.078	-0.102	4.124†	1.920	2.228*
	SE	(0.035)	(0.051)	(0.025)	(0.159)	(0.085)	(1.948)	(1.205)	(0.968)
	N	3,168	3,168	3,168	3,168	3,168	3,168	3,168	3,168

Notes: † p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001. All crime, felony, misdemeanor, proactivity, and complaints columns are Poisson models, except for the CS DiD models. Coefficients from the Poisson models are exponentiated, and their 95% cluster-robust confidence intervals are reported in brackets. All crime, felony, misdemeanor, proactivity, and complaints in the CS DiD models are all transformed by asinh(y) to approach a similar "percent change" interpretation as the Poisson models. Their standard errors were acquired through clustered bootstrap. SA (post-treat) and CS (post-treat) test whether the average post-treatment effect (event time 0-22) differs from the average pre-treatment effect (event time -12 to -2) (see the Appendix). For the timing outcomes, the coefficients are changes in minutes, and precinct-clustered standard errors are reported in parentheses.

Table 5: Staggered DDD Estimates with the other Estimators.

Estimator		All Crime	Felony	Misdemeanor	Complaints	Dispatch Time	CIP Dispatch Time	CIP Arrival Time
TWFE DDD (Surge)	Estimate	1.075	1.059	1.087	0.029	-1.406	-1.006	-0.563
	CI/SE	[0.983, 1.175]	[0.972, 1.153]	[0.969, 1.219]	(0.022)	(1.918)	(1.136)	(0.677)
	N	62,208	62,208	62,208	62,208	62,208	62,208	62,208
TWFE DDD (Last 3)	Estimate	0.982	0.995	0.972	-0.012	4.519***	1.093*	0.320
	CI/SE	[0.906, 1.064]	[0.933, 1.060]	[0.846, 1.117]	(0.018)	(0.781)	(0.472)	(0.546)
	N	62,208	62,190	62,190	62,208	62,208	62,208	62,208
SA (ref = -1) DDD (Surge)	Estimate	1.030	1.027	1.033	-0.046	-2.500	-3.775***	-1.562
	CI/SE	[0.967, 1.097]	[0.962, 1.096]	[0.904, 1.180]	(0.106)	(1.641)	(0.596)	(0.963)
	N	62,208	62,208	62,208	62,208	62,208	62,208	62,208
SA (post-treat) DDD (Surge)	Estimate	1.017	1.035	1.014	-0.008	-0.130	-0.702	-0.410
	CI/SE	[0.982, 1.052]	[0.989, 1.083]	[0.954, 1.078]	(0.024)	(1.388)	(0.579)	(0.385)
	N	62,208	62,208	62,208	62,208	62,208	62,208	62,208
SA (ref = -1) DDD (Last 3)	Estimate	0.953	0.940	1.000	0.075	4.487***	0.613	0.800
	CI/SE	[0.854, 1.063]	[0.823, 1.074]	[0.854, 1.171]	(0.056)	(0.798)	(0.426)	(0.830)
	N	62,208	62,190	62,190	62,208	62,208	62,208	62,208
SA (post-treat) DDD (Last 3)	Estimate	0.992	0.995	0.991	-0.008	4.899***	1.207***	0.023
	CI/SE	[0.975, 1.009]	[0.968, 1.024]	[0.936, 1.050]	(0.025)	(0.785)	(0.333)	(0.347)
	N	62,208	62,190	62,190	62,208	62,208	62,208	62,208

Notes: † p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001. All crime, felony, misdemeanor, and proactivity columns are Poisson models. Coefficients from the Poisson models are exponentiated, and their 95% cluster-robust confidence intervals are reported in brackets. For complaints, the coefficients are change in the probability of having any complaint. For the timing outcomes, the coefficients are changes in minutes, and precinct-clustered standard errors are reported in parentheses. SA (post-treat) and CS (post-treat) test whether the average post-treatment effect (event time 0-22) differs from the average pre-treatment effect (event time -12 to -2) (see the Appendix).

Table 6: Re-estimations without 25% of the Low-crime Precincts.

Outcome		Stacked DiD	Stacked DDD (Surge Hours)	Stacked DDD (Last Three Hours)
All Crime	Estimate	1.061† [0.995, 1.131]	1.064† [0.993, 1.139]	0.991 [0.926, 1.061]
	Observations	5,180	93,240	93,240
	Pre-treatment mean	161.04	9.26	6.58
Felony Crimes	Estimate	1.044 [0.953, 1.144]	1.068† [0.992, 1.150]	1.002 [0.943, 1.065]
	Observations	5,180	93,240	93,240
	Pre-treatment mean	79.71	4.74	3.38
Misdemeanor Crimes	Estimate	1.106*** [1.042, 1.174]	1.062 [0.967, 1.167]	0.987 [0.866, 1.124]
	Observations	5,180	93,240	93,240
	Pre-treatment mean	68.04	3.94	2.70
Proactivity	Estimate	0.800† [0.628, 1.020]	—	—
	Observations	5,180	—	—
	Pre-treatment mean	107.57	—	—
Complaints	Estimate	0.962 [0.878, 1.054]	-0.010 (0.034)	-0.005 (0.026)
	Observations	5,180	93,240	93,240
	Pre-treatment mean	39.94	0.39	0.24
Dispatch Time	Estimate	4.368* (1.689)	0.000 (1.741)	4.955*** (0.827)
	Observations	5,180	93,240	93,240
	Pre-treatment mean	15.70	17.98	12.97
CIP Dispatch Time	Estimate	1.962† (1.051)	-0.712 (0.828)	1.262*** (0.360)
	Observations	5,180	93,240	93,240
	Pre-treatment mean	8.47	9.74	7.48
CIP Arrival Time	Estimate	2.117* (0.907)	-0.372 (0.568)	0.083 (0.520)
	Observations	5,180	93,240	93,240
	Pre-treatment mean	11.28	11.12	10.55

Notes: † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Cluster-robust confidence intervals in brackets for Poisson regressions. Standard errors clustered at the precinct level are in parentheses. Pre-treatment means are for treated precincts in the pre-treatment period; for the DDD columns, they are restricted to the treated hours. For DDD estimates, complaints are modeled with linear probability models that indicate the probability of having any complaints filed in a certain hour in a month in a precinct. The estimates for Poisson models are shown as rate ratios rather than raw log-count coefficients. For the timing outcomes, the coefficients are changes in minutes, and precinct-clustered standard errors are reported in parentheses. The sample excludes the bottom 25% of precincts with the lowest crime rate during 2022 and April 2023.

Table 7: Re-estimations without 50% of the Low-crime Precincts.

Outcome		Stacked DiD	Stacked DDD (Surge Hours)	Stacked DDD (Last Three Hours)
All Crime	Estimate	1.039 [0.972, 1.111]	1.066† [0.991, 1.146]	0.996 [0.928, 1.068]
	Observations	3,290	59,220	59,220
	Pre-treatment mean	161.04	9.26	6.58
Felony Crimes	Estimate	1.017 [0.929, 1.113]	1.071† [0.993, 1.155]	1.004 [0.943, 1.069]
	Observations	3,290	59,220	59,220
	Pre-treatment mean	79.71	4.74	3.38
Misdemeanor Crimes	Estimate	1.087* [1.017, 1.161]	1.060 [0.953, 1.177]	0.992 [0.867, 1.134]
	Observations	3,290	59,220	59,220
	Pre-treatment mean	68.04	3.94	2.70
Proactivity	Estimate	0.751† [0.557, 1.012]	—	—
	Observations	3,290	—	—
	Pre-treatment mean	107.57	—	—
Complaints	Estimate	0.972 [0.872, 1.084]	-0.016 (0.034)	-0.006 (0.027)
	Observations	3,290	59,220	59,220
	Pre-treatment mean	39.94	0.39	0.24
Dispatch Time	Estimate	4.840** (1.688)	0.184 (1.791)	5.045*** (0.846)
	Observations	3,290	59,220	59,220
	Pre-treatment mean	15.70	17.98	12.97
CIP Dispatch Time	Estimate	2.114* (1.040)	-0.617 (0.861)	1.452*** (0.382)
	Observations	3,290	59,220	59,220
	Pre-treatment mean	8.47	9.74	7.48
CIP Arrival Time	Estimate	2.154* (0.911)	-0.261 (0.580)	0.177 (0.517)
	Observations	3,290	59,220	59,220
	Pre-treatment mean	11.28	11.12	10.55

Notes: † p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001. Cluster-robust confidence intervals in brackets for Poisson regressions. Standard errors clustered at the precinct level are in parentheses. Pre-treatment means are for treated precincts in the pre-treatment period; for the DDD columns, they are restricted to the treated hours. For DDD estimates, complaints are modeled with linear probability models that indicate the probability of having any complaints filed in a certain hour in a month in a precinct. The estimates for Poisson models are shown as rate ratios rather than raw log-count coefficients. For the timing outcomes, the coefficients are changes in minutes, and precinct-clustered standard errors are reported in parentheses. Sample excludes the bottom 50% of precincts with the lowest crime rate during 2022 and April 2023.

Table 8: Re-estimations without only Precincts in the Bronx.

Outcome		Stacked DiD	Stacked DDD (Surge Hours)	Stacked DDD (Last Three Hours)
All Crime	Estimate	1.001 [0.925, 1.084]	0.983 [0.912, 1.060]	0.952 [0.859, 1.056]
	Observations	840	15,120	15,120
	Pre-treatment mean	158.39	9.16	6.57
Felony Crimes	Estimate	0.994 [0.878, 1.125]	1.047 [0.932, 1.176]	1.005 [0.916, 1.104]
	Observations	840	15,120	15,120
	Pre-treatment mean	77.39	4.65	3.32
Misdemeanor Crimes	Estimate	1.023 [0.953, 1.099]	0.939 [0.772, 1.142]	0.919 [0.766, 1.103]
	Observations	840	15,120	15,120
	Pre-treatment mean	67.68	3.92	2.73
Proactivity	Estimate	0.607† [0.359, 1.027]	—	—
	Observations	840	—	—
	Pre-treatment mean	51.68	—	—
Complaints	Estimate	1.026 [0.842, 1.250]	0.010 (0.025)	0.008 (0.034)
	Observations	840	15,120	15,120
	Pre-treatment mean	38.51	0.35	0.22
Dispatch Time	Estimate	6.851* (3.059)	1.030 (2.656)	5.527*** (1.018)
	Observations	840	15,120	15,120
	Pre-treatment mean	16.92	19.38	13.89
CIP Dispatch Time	Estimate	3.169† (1.562)	-0.193 (1.674)	1.851* (0.834)
	Observations	840	15,120	15,120
	Pre-treatment mean	8.92	10.27	7.84
CIP Arrival Time	Estimate	2.252† (1.178)	-0.893 (0.820)	-0.072 (0.535)
	Observations	840	15,120	15,120
	Pre-treatment mean	11.54	11.70	10.73

Notes † p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001. Cluster-robust confidence intervals in brackets for Poisson regressions. Standard errors clustered at the precinct level are in parentheses. Pre-treatment means are for treated precincts in the pre-treatment period; for the DDD columns, they are restricted to the treated hours. For DDD estimates, complaints are modeled with linear probability models that indicate the probability of having any complaints filed in a certain hour in a month in a precinct. The estimates for Poisson models are shown as rate ratios rather than raw log-count coefficients. For the timing outcomes, the coefficients are changes in minutes, and precinct-clustered standard errors are reported in parentheses. The pre-treatment means change in this table because a non-Bronx precinct (Midtown South) is excluded.

Figure 1: Service Demand and Response Time by Hour.

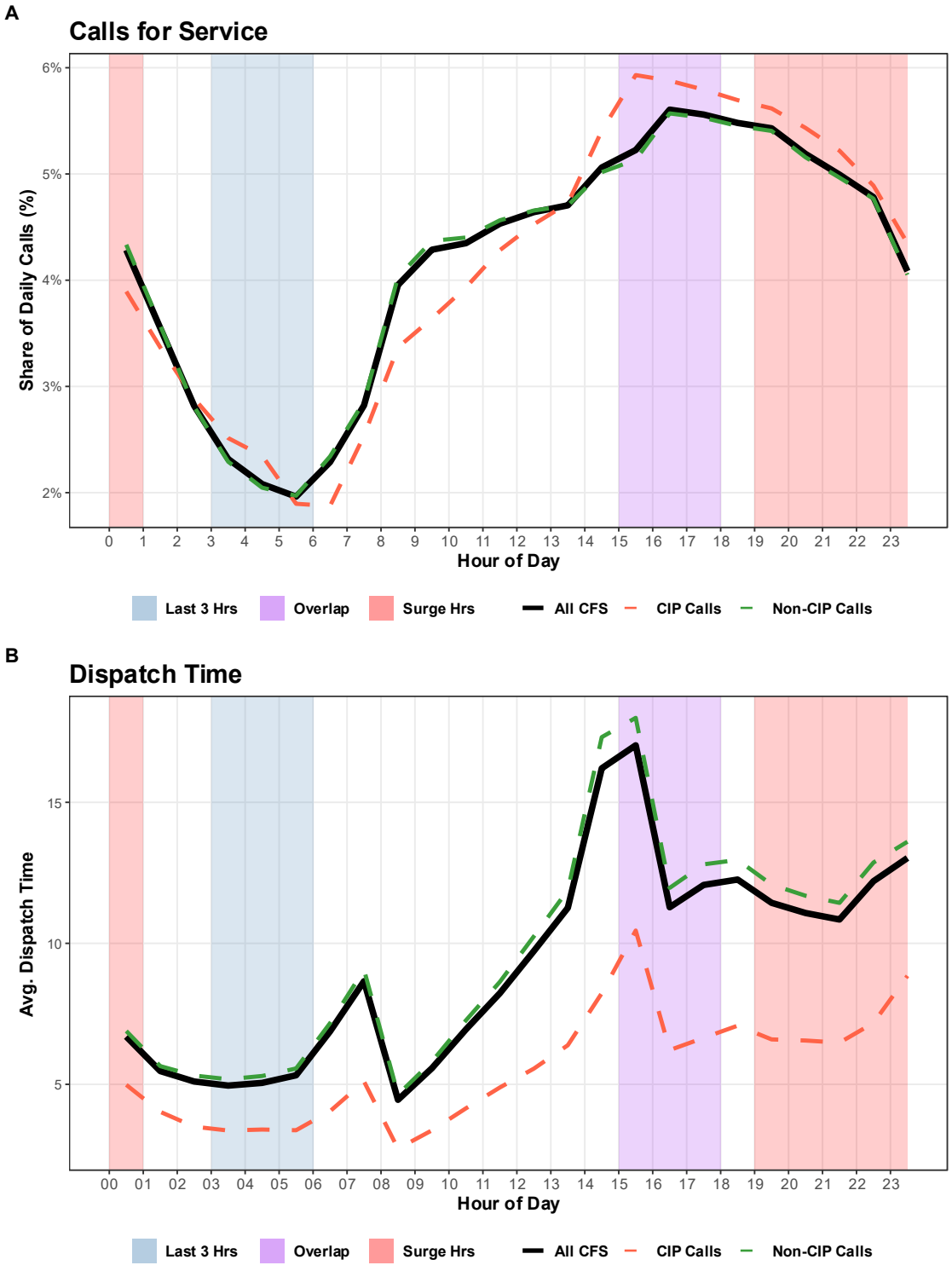
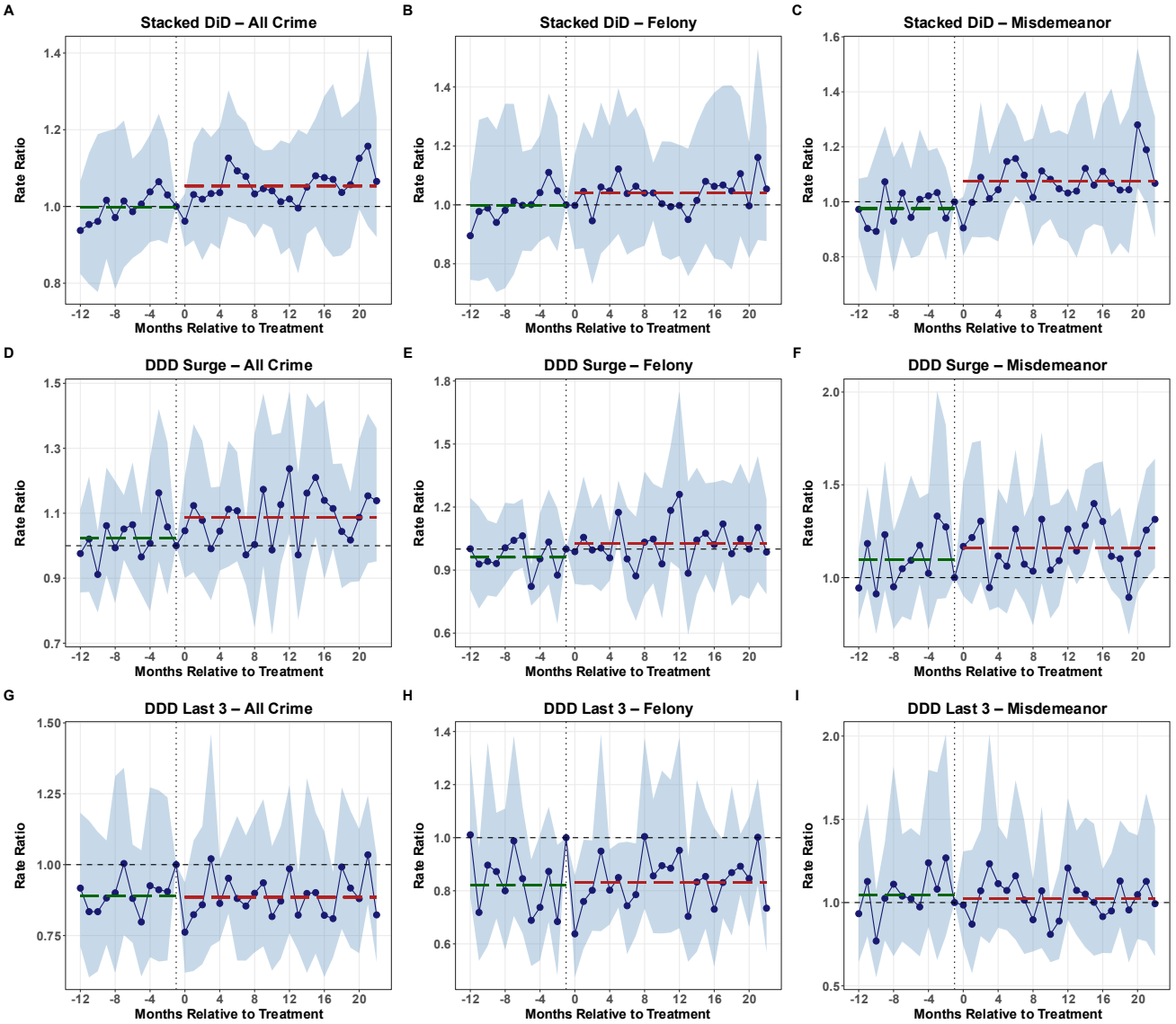
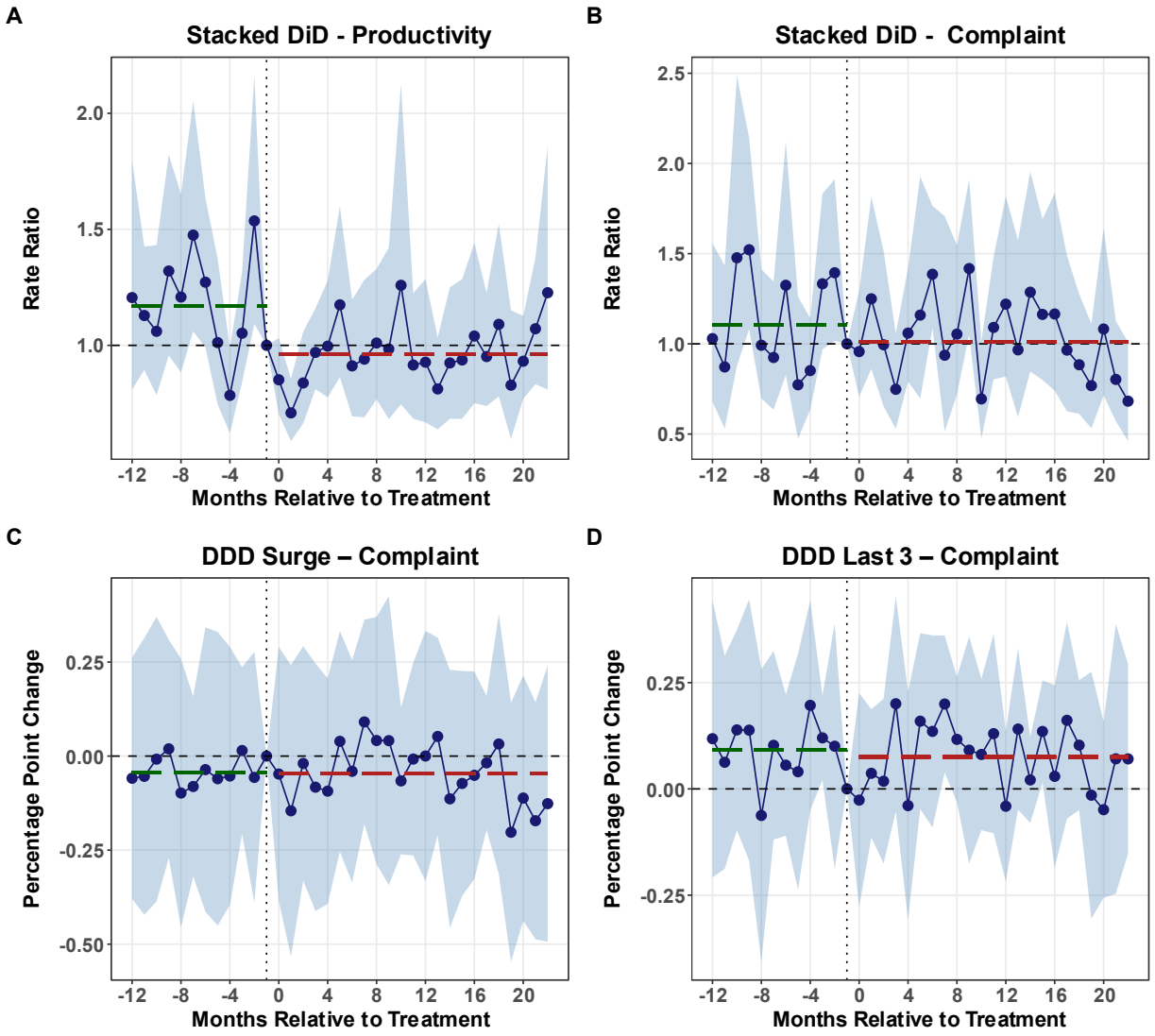


Figure 2: The Event Study Plots for Crimes.



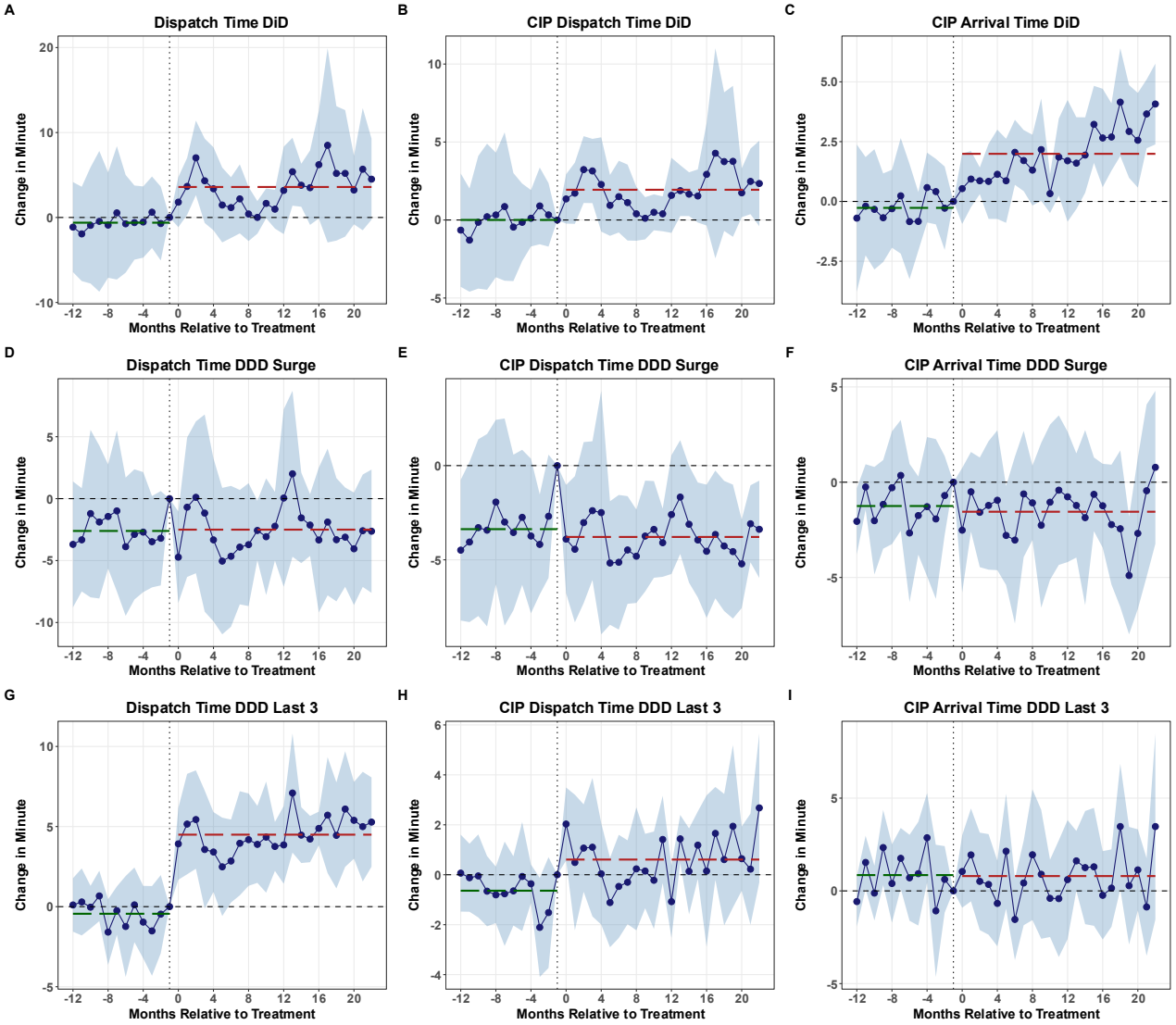
Notes: The shades show the 95% confidence interval. The vertical dashed line is the reference period ($s = -1$).

Figure 3: The Event Study Plots for Proactivity and Complaints.



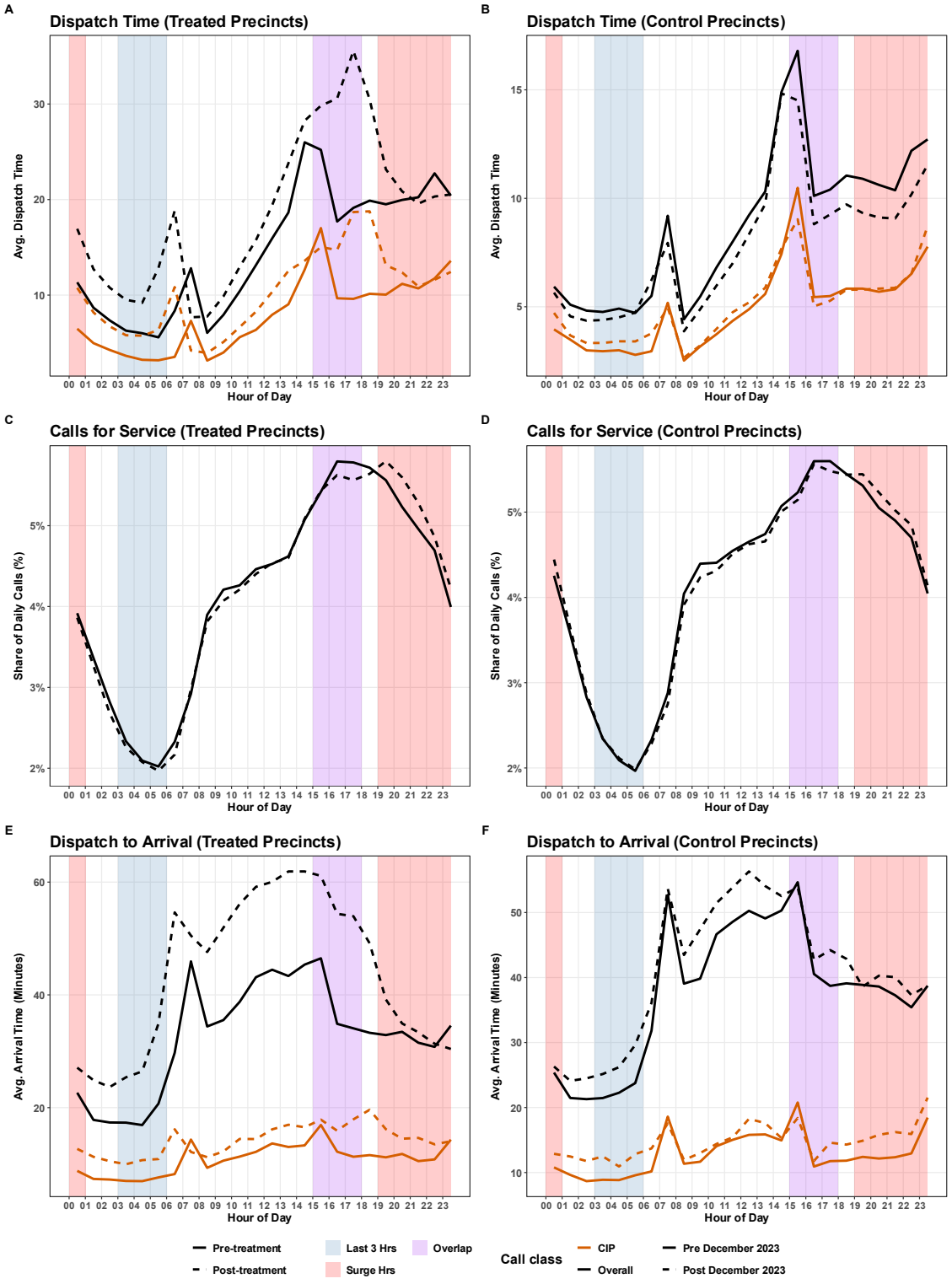
Notes: The shades show the 95% confidence interval. The vertical dashed line is the reference period ($s = -1$).

Figure 4: The Event Study Plots of Dispatch Time.



Notes: The shades show the 95% confidence interval. The vertical dashed line is the reference period ($s = -1$).

Figure 5: Dispatch Time and Call-for-Service Demand by Hour and Treatment Status.



8. Appendix 1: The Post-Pre Aggregation of Cohort-time Effects in `fixest` and `did`.

8.1. The interaction-weighted estimator (SA)

The interaction-weighted estimators (SA) for the CWW schedule can be written in the following regression form (Sun & Abraham, 2021):

$$y_{it} = \sum_{g} \sum_{\substack{-12 \le s \le -2, \\ 0 \le s \le 22}} \delta_{gs} \cdot 1\{G_i = g\} \cdot 1\{R_{it} = s\} + \delta_{g,s>22} 1\{G_i < T\} \cdot 1\{R_{it} > 22\} + \delta_{g,s<-12} 1\{G_i < T\} \cdot 1\{R_{it} < -12\} + \phi_i + \lambda_t + \epsilon_{it}. \quad (8\text{-}1)$$

G_i is the month t when the precinct started its longer shift. The analyzed precincts that had not had the longer shifts by the end of the observation period, December 2025 (T), have $G_i = \infty$. R_{it} is the relative time to the treatment (s). The key to accounting for the heterogeneous treatment effects across pilot cohorts (or "subexperiments") in SA is in the interaction terms of G_i and R_{it}, capturing the *cohort-specific* event-time effects, δ_{gs}, in ordinary linear regressions.

$\delta_{g,s>22}$ and $\delta_{g,s<-12}$ are binned, aggregated estimates for time-since-treatment that are larger than 22 months or smaller than -12. Unlike in stacked staggered DiD, which restricts the periods of observation to be the same for each treatment cohort within each stacked dataset, the three pilot cohorts are imbalanced in terms of the available periods for estimation before and after the CWW schedule. These two binned estimates net out the roles of these distance measures and ensure comparability with stacked staggered DiD. The boundaries are set to be the same as the predefined ones $\kappa_a = 12$ and $\kappa_b = 22$ in Section 5.1. To obtain an aggregated estimate of the overall average treatment effect on the treated, like β in Equation 5-1, the `fixest` package provides a weighted sum of each $\hat{\delta}_{gs}$ with the size of the treatment cohort within a certain range of s (Bergé, 2026), for instance,

$$\widehat{ATT}^{SA}_{post} = \frac{\sum_{22 \ge s \ge 0} \sum_{g \ne \infty} N_{gs} \cdot \hat{\delta}_{gs}}{\sum_{22 \ge s \ge 0} \sum_{g \ne \infty} N_{gs}}, \; \hat{w}_{gs} = \frac{N_{gs}}{\sum_{22 \ge s \ge 0} \sum_{g \ne \infty} N_{gs}},$$

where $N_{gs}$is the number of treated precinct-period observations in the cell (g, s). With the *delta-method* variance,

$$\widehat{\mathrm{Var}}\left(\widehat{ATT}^{SA}_{post}\right) = \hat{w}' \, \widehat{\mathrm{Var}}\left(\hat{\delta}\right) \hat{w}.$$

This can be achieved with the syntax: `summary(<fixest_model_with_sunab>, agg="^<period_name>::(?:[0-9]|1[0-9]|2[0-2]):cohort")`.

However, because all $\hat{\delta}_{gs}$ are measured against the outcome level at $s = -1$, $\widehat{ATT}^{SA}_{post}$ measures only the relative change of outcome in the pilot precincts to the business-as-usual precincts since *one month before* the CWW implementation, while $\hat{\beta}_{stk}$ captures the same relative change from the average pre-treatment outcome level across 12 to 1 months before treatment. Under perfect parallel trends, $\forall \delta_{g,s<-1} = 0$, the choice of reference period is innocuous. However, if $\hat{\delta}_{g,s<-1}$ oscillate in the pre-treatment period around a certain outcome level and are not jointly significant, the treatment effects acquired by comparing with just the level at $s = -1$ do not seem to represent the treatment effect accurately. In contrast, $\hat{\beta}_{stk}$ is less sensitive to the drift in just one period.

To overcome such a mechanical difference between stacked DiD and SA, we adapted the aggregation of $\hat{\delta}_{gs}$ into the ATT as follows:

$$\widehat{ATT}^{SA}_{pre} = \frac{\sum_{-1 \geq s \geq -12} \sum_{g \neq \infty} N_{gs} \cdot \hat{\delta}_{gs}}{\sum_{-1 \geq s \geq -12} \sum_{g \neq \infty} N_{gs}},$$

$$\widehat{ATT}^{SA}_{post-pre} = \widehat{ATT}^{SA}_{post} - \widehat{ATT}^{SA}_{pre}.$$

The standard errors are similarly acquired by the delta method. This is how SA (post-pre) estimates the effects in Table 4 and Table 5. The code for the DiD estimate is in the following. The DDD version is simply an extension from this.

```
sa_pre_post_test <- function(sa_mod, data){

  org.fml <- as.character(sa_mod$fml)
  sunab_term <- str_detect(org.fml, "sunab")
  org.fml[sunab_term] <- str_subset(org.fml, "sunab") %>%
    str_replace('(\\)\\))', '\\), no_agg = T\\)')
  org.fml <- paste0(org.fml[2], org.fml[1], org.fml[3:length(org.fml)],
collapse = "")

  test <- update(sa_mod, as.formula(org.fml))

  V <- vcov(test)
  b <- coef(test)[rownames(V)]
  post_idx <- grep('^ym_ord::(?:[0-9]|1[0-9]|2[0-2]):cohort', names(b),
perl = TRUE)
  pre_idx  <- grep('^ym_ord::-(?:[2-9]|1[0-2]):cohort',        names(b))

  # Cohort sizes: number of treated precincts per cohort
  cohort_n <- data %>%
    filter(first_treat_m_ord > 0, !is.na(first_treat_m_ord)) %>%
    distinct(precinct, first_treat_m_ord) %>%
```

```
    count(first_treat_m_ord)

  # For each coefficient, extract cohort ordinal and look up n_g;
normalize to sum = 1
  get_weights <- function(idx) {
    g_vals <- as.numeric(str_extract(names(b)[idx], '(?<=cohort::)[0-
9]+'))
    w <- cohort_n$n[match(g_vals, cohort_n$first_treat_m_ord)]
    w / sum(w, na.rm = TRUE)
  }

  k <- numeric(length(b))
  k[post_idx] <-  get_weights(post_idx)
  k[pre_idx]  <- -get_weights(pre_idx) * 11/12 # Normalize to 12 months
for the whole pre-treatment period
  est    <- as.numeric(k %*% b)
  se     <- sqrt(as.numeric(t(k) %*% V %*% k))
  t_stat <- est / se
  p_val  <- 2 * pnorm(-abs(t_stat))

  cat(sprintf(
    "ATT - Pre: %.4f  (SE = %.4f,  t = %.2f,  p = %.4f)\n",
    est, se, t_stat, p_val
  ))

  invisible(tibble::tibble(
    est  = est,
    se   = se,
    t    = t_stat,
    p    = p_val,
    nobs = test$nobs
  ))
}
```

8.2. Callaway and Sant'Anna estimator

Similarly, the Callaway and Sant'Anna estimator likewise builds on group-time average treatment effects, but instead of recovering them from a regression, it computes each $ATT(g,t)$ directly as a two-group, two-period comparison of sample means. Absent covariates, the doubly robust estimator implemented in `did` reduces to

$$\widehat{ATT}(g,t) = \frac{1}{N_g}\sum_{i\in\{G_i=g\}}\left(y_{it} - y_{i,g-1}\right) - \frac{1}{N_C}\sum_{i\in G_i=\infty}\left(y_{it} - y_{i,g-1}\right).$$

N_g is the number of precincts in cohort g; N_C is the number of precincts that had not had a longer shift until the end of December 2025. The `aggte` function in the did package provides the following aggregation with a specified range, in our case,

$$\widehat{ATT}^{CS}_{post} = \sum_g \frac{N_g}{\sum_{g\neq\infty} N_g}\left[\frac{1}{23}\sum_{t=g}^{g+22}\widehat{ATT}(g,t)\right].$$

The variance of $\widehat{ATT}^{CS}_{post}$ is obtained through multiplier bootstrapping of the precinct-sum influence functions of $\widehat{ATT}(g,t)$ and the cohort-size weights used for aggregation (Callaway & Sant'Anna, 2026a, 2026b).

Again, the $y_{i,g-1}$ in $\widehat{ATT}(g,t)$ suggests that the change is measured against the baseline of one month before the CWW pilot. $\widehat{ATT}^{CS}_{post}$ thus also shares the same single-baseline issue as $\widehat{ATT}^{SA}_{post}$. The proposed alternative estimator, CS (post-pre) in Table 4, is

$$\widehat{ATT}^{CS}_{post-pre} = \sum_g \frac{N_g}{\sum_{g\neq\infty} N_g}\left[\frac{1}{23}\sum_{t=g}^{g+22}\widehat{ATT}(g,t) - \frac{1}{12}\sum_{t=g-12}^{g-1}\widehat{ATT}(g,t)\right].$$

This requires `base_period = "universal"`, so that pre-treatment estimates are expressed relative to $g-1$ rather than to the immediately preceding month. For $\widehat{Var}\left(\widehat{ATT}^{CS}_{post-pre}\right)$, we bootstrap the whole estimation process by resampling precincts 1,000 times. The *p*-value is acquired through the percentile of the resampled distribution of $\left(\widehat{ATT}^{CS}_{post-pre}\right)^b$ with `boot.pval` package. The code is provided below.

```
boot_cs_star <- function(att_obj, R = 1000,
                         s_post = c(0, 22), s_pre = c(-12, -2),
                         workers = min(100, availableCores())) {
  dp       <- att_obj$DIDparams
  idname   <- dp$idname
  gname    <- dp$gname
  tname    <- dp$tname
  pnl_data <- dp$data   # already contains the transformed outcome column

  pcts_all <- unique(pnl_data[[idname]]) # All the precinct.

  stat_fn <- function(pcts, idx) {
    boot_pcts <- pcts[idx] # randomly drawing precincts.
```

```
    boot_df   <- map_dfr(seq_along(boot_pcts), \(i)
      filter(pnl_data, .data[[idname]] == boot_pcts[i]) %>%
        # A precinct may appear twice after the draw, while att_gt()
requires non-repetitive idname!
        # They should be treated as independent clusters.
        mutate(!!idname := i)
    ) # Automatically combine the rows into a df.

    cs_b <- tryCatch(
      # Update the object to skip bootstrapping.
      update(att_obj, data = boot_df, bstrap = FALSE, pl = FALSE),
      error = function(e) NULL
    )
    if (is.null(cs_b)) return(NA_real_)

    att_b <- tibble(g = cs_b$group, t = cs_b$t, att = cs_b$att) %>%
      mutate(s = t - g)

    N_b <- boot_df %>%
      filter(.data[[gname]] > 0) %>%
      mutate(g = .data[[gname]], s = .data[[tname]] - .data[[gname]]) %>%
      filter(s >= s_pre[1], s <= s_post[2], s != -1) %>%
      count(g, s, name = "N_gs")

    cells_b <- inner_join(att_b, N_b, by = c("g", "s"))
    if (nrow(cells_b) == 0) return(NA_real_)

    post_b <- cells_b %>% filter(s >= s_post[1], s <= s_post[2]) %>%
      summarise(v = weighted.mean(att, N_gs)) %>% pull(v)
    pre_b  <- cells_b %>% filter(s >= s_pre[1],  s <= s_pre[2]) %>%
      summarise(v = weighted.mean(att, N_gs)) %>% pull(v)

    post_b - pre_b * 11/12 # Normalize to 12 months for the whole pre-
treatment period
  }

  boot::boot(
    data      = pcts_all,
    statistic = stat_fn,
    R         = R,
    parallel  = "multicore",
    ncpus     = workers
  )
```

```
}
```

References

Bergé, L. (2026). *fixest: Fast fixed-effects estimations*. In (Version Version 0.14.2, commit 789f888416e97673d154c840eec2d99101fc07b1) [R package source code]. GitHub. https://github.com/lrberge/fixest/blob/789f888416e97673d154c840eec2d99101fc07b1/R/did.R#L484-L575

Callaway, B., & Sant'Anna, P. H. C. (2026a). *compute.aggte.R [Source Code]*. https://github.com/bcallaway11/did/blob/master/R/compute.aggte.R#L697

Callaway, B., & Sant'Anna, P. H. C. (2026b). *mboot.R [Source Code]*. https://github.com/bcallaway11/did/blob/9518b68295a87f5112e637645b84a54fb7593799/R/mboot.R#L144

Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, *225*(2), 175–199. https://doi.org/10.1016/j.jeconom.2020.09.006

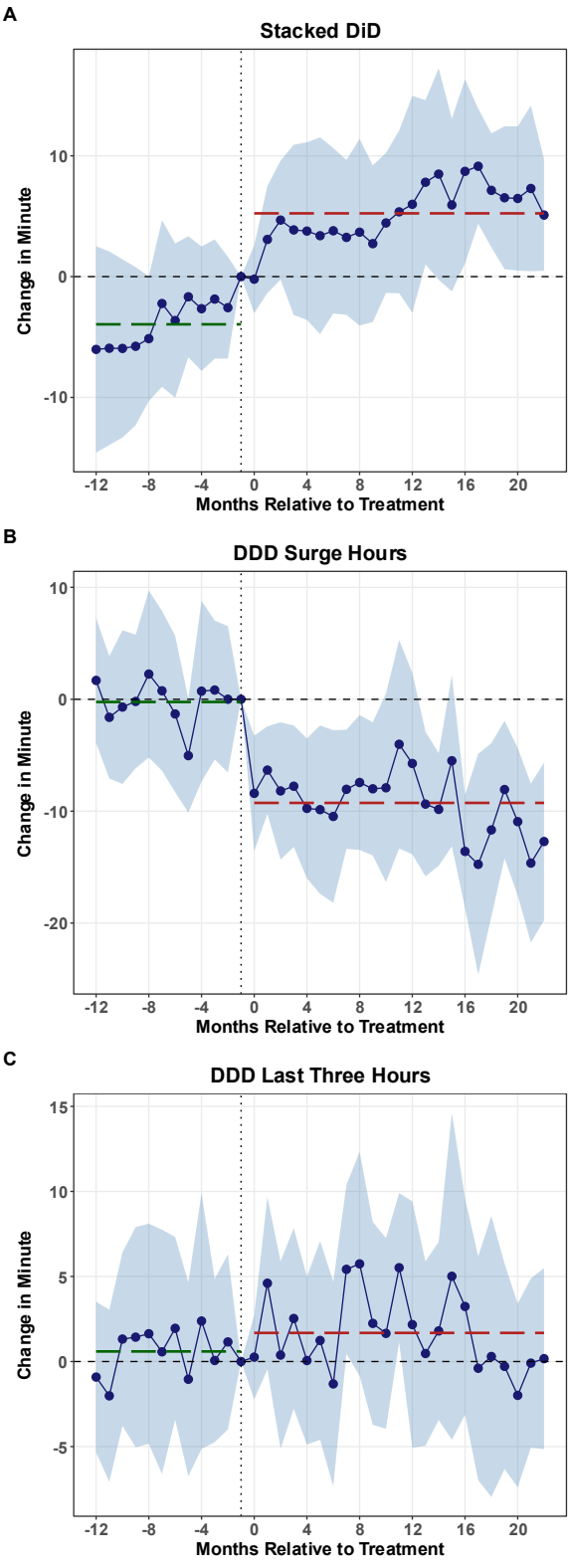
9. Appendix 2: Change in Overall Arrival Time.

Table A1: Estimated Change in Overall Arrival Time.

Outcome		Stacked DiD	Stacked DDD (Surge Hours)	Stacked DDD (Last Three Hours)
Overall arrival time	Estimate	8.856**	-9.043***	1.139
		(2.772)	(2.354)	(1.096)
	Observations	7,070	127,260	127,260
	Pre-treatment mean	36.33	33.44	30.61

Notes † p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001. Pre-treatment means are for treated precincts in the pre-treatment period; for the DDD columns, they are restricted to the treated hours. The coefficients are changes in minutes, and precinct-clustered standard errors are reported in parentheses. Note that the significantly negative coefficient from the stacked DDD for surge hours does not indicate a decrease in arrival time during this period. Rather, it suggests that arrival time *increased by 9 minutes during control hours* (see Figure 5). The surge-hour arrangement appears to attenuate, but not eliminate, the overall increase in response times in CWW pilot precincts.

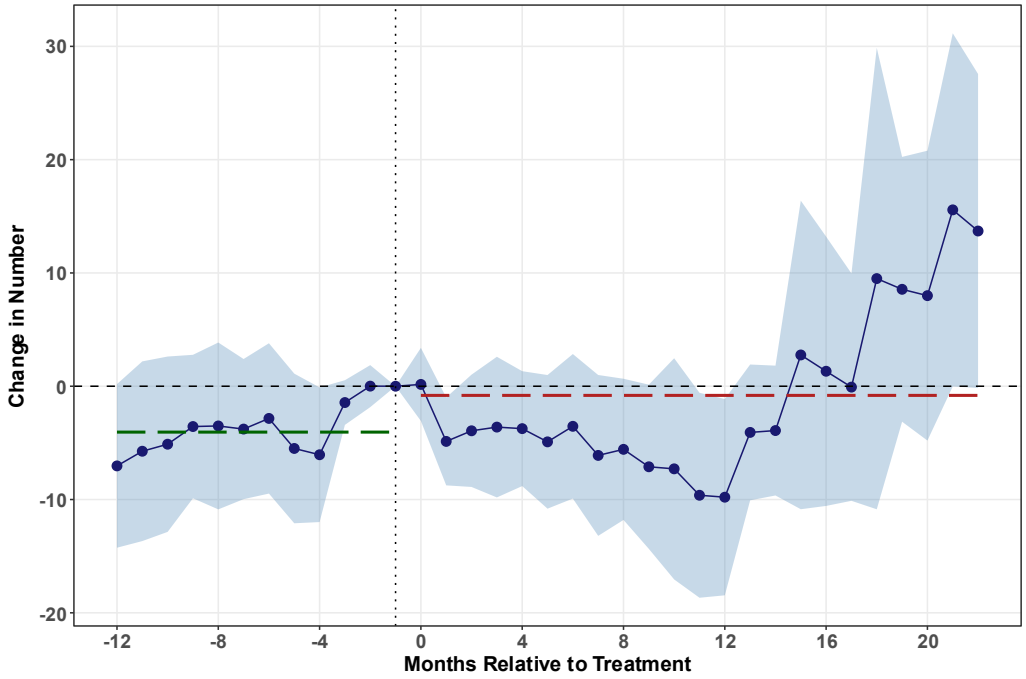
Figure A1: Event Study Plots for Overall Arrival Times.



Notes: The shades show the 95% confidence interval. The vertical dashed line is the reference period ($s = -1$).

10. Appendix 3: Change in Staffing Level.

Figure A2: Event Study Plots for Staffing.



Notes: The shades show the 95% confidence interval. The vertical dashed line is the reference period ($s = -1$).