NBER WORKING PAPER SERIES

DETERRENCE OR BACKLASH?
ARRESTS AND THE DYNAMICS OF DOMESTIC VIOLENCE

Sofia Amaral
Gordon B. Dahl
Victoria Endl-Geyer
Timo Hener
Helmut Rainer

Working Paper 30855
http://www.nber.org/papers/w30855

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2023

We thank the West Midlands Police for generously providing access to their data and expertise, without which this project would not have been possible. Amaral thanks the Leibniz Association Pact for Research and Innovation for financial support. We also thank colleagues and seminar participants at several universities and conferences for valuable feedback and suggestions. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2023 by Sofia Amaral, Gordon B. Dahl, Victoria Endl-Geyer, Timo Hener, and Helmut Rainer. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Deterrence or Backlash? Arrests and the Dynamics of Domestic Violence
Sofia Amaral, Gordon B. Dahl, Victoria Endl-Geyer, Timo Hener, and Helmut Rainer
NBER Working Paper No. 30855
January 2023
JEL No. J12,J16,K42

ABSTRACT

There is a vigorous debate on whether arrests for domestic violence (DV) will deter future abuse or create a retaliatory backlash. We study how arrests affect the dynamics of DV using administrative data for over 124,000 DV emergency calls (999 calls) for West Midlands, the second most populous county in England. We take advantage of conditional random assignment of officers to a case by call handlers, combined with systematic differences across police officers in their propensity to arrest suspected batterers. We find that an arrest reduces future DV calls in the ensuing year by 51%. This reduction is not driven by reduced reporting due to fear of retaliation, but instead a decline in repeat victimization. We reach this conclusion based on a threshold reporting model and its testable implications regarding (i) the severity of repeat DV calls and (ii) victim versus third-party reporting. Exploring mechanisms, we find that arrest virtually eliminates the large spike in re-victimization which occurs in the 48 hours after a call, consistent with arrest facilitating a cooling off period during a volatile, at-risk time. In the longer run, we estimate a sizeable deterrence effect. Substantiating this, arrest increases the probability an offender is charged with a crime. Our findings suggest that if the goal is to lower the number of domestic violence incidents, police should lower their threshold for arrest, not decriminalize domestic violence.

Sofia Amaral
ifo Institute
Center for Labor and Demographic
Economics Poschingerstraße 5
81679 Munich
Germany
amaral.sofiafernando@gmail.com

Gordon B. Dahl
Department of Economics
University of California, San Diego
9500 Gilman Drive #0508
La Jolla, CA 92093-0508
and NBER
gdahl@ucsd.edu

Victoria Endl-Geyer
ifo Institute
Center for Labor and Demographic
Economics Poschingerstraße 5
81679 Munich
Germany
endl-geyer@ifo.de

Timo Hener
Aarhus University
Department of Economics and
Business Economics
Fuglesangs Alle 4
8210 Aarhus
Denmark
thener@econ.au.dk

Helmut Rainer
University of Munich
Ludwigstrasse 28
80539 Munich
Germany
rainer@econ.lmu.de

1 Introduction

Domestic violence (DV) is a pervasive threat to the well-being of women worldwide, with one third of women reporting some form of physical or sexual violence from a partner during their lifetime (WHO, 2021). A key aspect of DV is that it is seldom a one-time occurrence, with women frequently experiencing repeat abuse by the same partner (Tjaden and Thoennes, 2000; Aizer and Dal Bo, 2009). Despite the prevalence and seriousness of DV, the question of how best to police this crime so as to break the cycle of DV is still largely unresolved. A highly controversial police response is to arrest suspects on the spot.

The nature of the controversy is multifaceted. Proponents of arrest contend that in addition to temporarily incapacitating offenders, it deters men from future abuse by signaling a high cost for repeat incidents (Berk, 1993). Opponents raise concerns about backlash effects (Schmidt and Sherman, 1993; Goodmark, 2018), arguing that while arrest offers immediate relief, it triggers an escalation of DV in the longer term as abusers retaliate against their partners. Both sides also make arguments related to whether victims will report future DV, with proponents claiming that arrest empowers women to do so and opponents saying that arrest discourages future calls for police help.

Against the backdrop of these debates, this paper asks how arrest affects the dynamics of DV. Estimating the consequences of arrest on future victimization is challenging for two reasons. The first is the scarcity of large-scale datasets which include information on DV incidents, police responses, and repeat victimization. In most datasets, DV is not directly identified as a crime category, but must be inferred based on the characteristics of the incident. Moreover, researchers often only observe cases where criminal charges are filed, but in DV cases, the victim is often reluctant to press charges. And finally, it is difficult to link repeat victimization to a prior incident; even in datasets which track identities, they are usually only recorded if there is a documented criminal charge.

The second estimation challenge is that arrest is endogenous, leading to selection bias. The problem is that there are likely to be characteristics of a case which are observable to police officers, but not the researcher. For example, if cases which result in arrest are (unobservably) more serious and hence more likely to be positively related to the underlying risk of repeat violence, OLS will underestimate any potential deterrent effect or possibly yield a positive estimate even if there is no backlash.

We address both the data and endogeneity challenges in the context of emergency 999 calls in the United Kingdom (similar to 911 calls in the US). On the data front, we observe the universe of all 999 emergency calls for West Midlands, the second most populous county in England, over 10 years. Due to the priority placed on domestic abuse in the UK in recent years, DV is singled out as a separate category by call handlers. We follow an incident from the time the call is placed until the first-response police officers arrive and complete their on-scene intervention (including whether they arrest a suspected offender on the spot). We merge these records with data on whether a criminal investigation is opened by an investigative officer and, if so, whether offenders are charged with a crime.

To create a linked panel of DV incidents over time, we exploit information on the precise geo-location of where the incident occurred. This takes advantage of the fact that most DV occurs at home and that most police intervention originates via a 999 emergency call (HM Inspectorate of Constabulary, 2014). The key benefit of this approach is that it allows us to track repeat DV even if there is not a formal criminal charge filed. For this reason, we are likely to pick up a much higher fraction of repeat cases compared to other panel datasets which only have information related to formal charges. We provide several checks showing that the scope for misclassification errors using this measure (e.g., residential moves after being exposed to DV) is limited.[1]

To address endogeneity, we exploit the conditional random assignment of police officers to 999 DV calls combined with heterogeneity in officers' propensity to arrest.[2] Since it is difficult to precisely predict when and where demands for police resources will emerge, the availability of patrol officers which can be dispatched to a DV incident is as good as random after conditioning on time, geography, and the priority level assigned by the call handler. Since some officers are more likely to arrest than others, the average arrest propensity in other cases can be used as an instrument for arrest in the current case. As multiple patrol officers can be dispatched to an incident (there are usually 2 officers in a patrol car), we use the weighted average arrest rate

[1] For the subsample of cases which result in a criminal investigation (and hence for which we observe victim ID), we show that an arrest does not impact whether an observation is identified as a repeat DV incident using geo-location. We also show that our geo-location based results are robust to focusing on areas with single family homes as opposed to multifamily units.

[2] Our IV strategy is inspired by the quasi-random assignment of criminal cases to judges to investigate the consequences of juvenile incarceration (Aizer and Doyle Jr, 2015), adult incarceration (Kling, 2006; Bhuller et al., 2020), pre-trial detention (Dobbie et al., 2018), and electronic monitoring (Di Tella and Schargrodsky, 2013). In contexts other than crime, quasi-random assignment of caseworkers or judges has been used to examine the effects of disability insurance receipt (Maestas et al., 2013; Dahl et al., 2014; French and Song, 2014; Autor et al., 2019), child protection (Doyle Jr, 2007, 2008), and consumer bankruptcy (Dobbie and Song, 2015).

of dispatched team members to DV calls.[3] As we show, the instrument is highly predictive of arrest in the current case, but uncorrelated with observable case characteristics. Tests support the monotonicity assumption. To address concerns about the exclusion restriction, we show that the IV estimate for arrest is unlikely to be biased by other actions taken by responding officers on the scene or the characteristics of responding officers.

Our main finding is that arrest reduces the probability of a repeat DV emergency call within the ensuing 12 months by 51%. In sharp contrast, OLS finds a precisely estimated zero effect. The implication is that not accounting for selection bias would lead one to erroneously conclude that arrest has no affect on DV call trajectories.

Whether the reduction in repeat DV calls after an arrest is good or bad from a policy perspective depends on whether it is driven by a reduction in incidence or merely a reduction in reporting. To disentangle these explanations for our main result, we use a simple threshold reporting model. In the model, women who experience backlash from an arrest raise the level of abuse they are willing to tolerate before reporting in the future. In contrast, women who are empowered due to the deterrent effect of arrest will report future abuse at a lower threshold. Using a measure for the severity of repeat DV calls, we find that reporting thresholds drop on average after an arrest: there is a large reduction in severe DV calls and an increase in less severe DV calls. This compositional shift is statistically significant. Viewed through the lens of the model, it implies the decline in future DV calls is not driven by a change in reporting behavior, but rather a reduction in abuse. As a second exercise, we use differences in victim versus third-party reports of repeat DV calls. The idea behind this comparison is that any reluctance to report a repeat case due to fears of retaliation should be substantially larger for the victim than a third party. Yet we find, if anything, the opposite: a larger (but statistically insignificant) reduction in third party versus victim reporting. This is also consistent with a drop in actual abuse.

We next turn to mechanisms. In the absence of an arrest, we estimate that 23% of compliers will experience repeat DV within 48 hours. An arrest prevents virtually all of this re-victimization. This is consistent with arrest facilitating a cooling-off period, which could happen if the offender is removed from the scene to be processed or held in temporary custody. Additional reductions occur over the following year, consistent with a longer-term deterrence effect. Con-

[3] We use the initially dispatched officers to construct the instrument, as additional officers may be called to the scene if backup is needed or if the initially dispatched officers are not able to respond.

sistent with a deterrence effect, we estimate that an arrest leads to a higher probability of being formally charged with a crime. For non-arrested compliers, the probability of being formally charged with a crime is just 2%, while for arrested compliers, this probability rises to 12%. These findings go against the claim that arrest has a weak dosage of legal sanctions and therefore will be ineffective.

Taken together, our findings provide a cautionary tale for recent policy debates calling for a decriminalization of DV. Goodmark (2018) argues that the criminal legal system has been ineffective in deterring DV and has had detrimental consequences for victims, offenders, and their communities. Arrest in particular has been singled out as having negligible effects on DV recidivism. Indeed, our OLS estimates show no association between arrest and future violence. However, our causal estimates indicate a strong deterrent effect of arrest on domestic assault for compliers. This suggests that decriminalizing DV would increase the total amount of domestic violence, while lowering the police threshold for arrests would reduce it.

In the existing literature, there is sparse causal evidence on the effect arrest has on repeat victimization. Much of what we know comes from the 1981 Minneapolis Domestic Violence Experiment (Sherman and Berk, 1984) and its replications in Omaha (Dunford et al., 1990), Charlotte (Hirschel and Hutchison III, 1992), Milwaukee (Sherman et al., 1992), Dade County (Pate and Hamilton, 1992), and Colorado Springs (Berk et al., 1992). In these social experiments, the research design called for one of three treatments to be randomly enforced by patrol officers encountering a DV situation: arrest, separate, or advise. Despite being highly innovative at the time, the significance of these trials is limited in that they produced inconclusive findings,[4] had to contend with small samples, and suffered from flawed designs. A major concern was non-compliance with random assignment: patrol officers frequently deviated from the treatment called for by random assignment, and their delivered treatments had a substantial behavioral component (Angrist, 2006). These studies highlight some of the challenges involved in using randomized control trials for the empirical analysis of crime (Pinotti, 2020). More recently, a related literature examines how the staggered introduction of arrest laws in the US affected state-level intimate partner homicide rates. Iyengar (2009) finds an increase in homicides after a mandatory arrest while Chin and Cunningham (2019) revisit the issue and find no effect for mandatory arrest laws but a reduction in homicides for discretionary arrest laws.

[4] In Omaha, Charlotte, and Milwaukee, offenders assigned to the arrest group showed higher levels of repeat offending while in the other three jurisdictions, a modest reduction in recidivism was found among batterers assigned to arrest (Schmidt and Sherman, 1993).

Our paper is also related to a small but growing number of papers which study the role of law enforcement more generally in the prevention of DV. Miller and Segal (2019) provide evidence that when female representation increases among police officers in an area, DV crimes are reported at higher rates, intimate partner homicide falls, and non-fatal domestic abuse declines. Aizer and Dal Bo (2009) find that the adoption of no-drop policies in the US, which require prosecution even if DV victims request charges be dropped, resulted in an increase in reporting of DV. Golestani et al. (2021) show that victims in cases assigned to specialized DV courts are less likely to be re-victimized and more likely to cooperate with police and prosecutors. In a contemporary working paper, Black et al. (2022) use inverse propensity-score weighting to study the role of criminal charges against offenders and protective services for victims, finding that the former substantially reduces the risk of DV recidivism. Grogger et al. (2021) apply machine learning methods to forecast recidivism in DV cases, laying out how their approach can be used to prioritize DV emergency calls.

While our focus is on how DV can be prevented, there is an important line of work examining its determinants. Bhalotra et al. (2021) estimate the impacts of job loss and unemployment benefits while work by Aizer (2010), Tur-Prats (2021), and Erten and Keskin (2021) has focused on local labor market conditions. Cultural factors have also been linked to DV (e.g., Alesina et al. (2021), González and Rodríguez-Planas (2020), and Tur-Prats (2019)). Our paper also relates to a small literature that examines the effects of DV on victims and their children. Aizer (2011) and Currie et al. (2022) find that domestic violence during pregnancy significantly increases the incidence of negative birth outcomes.

The remainder of the paper is organized as follows. In Section 2 we discuss our setting and research design and in Section 3 we assess our instrument. Section 4 presents the main results on how arrest affects repeat DV calls. Section 5 tests for changes in incidence versus reporting behavior. Section 6 explores mechanisms, followed by further robustness checks and heterogeneity in Section 7. The final section concludes.

2 Research Design

In this section we outline our research design. We begin by briefly describing our setting, followed by an explanation of how DV emergency calls are handled. We then discuss how the conditional random assignment of police officers with different arrest propensities can be used to estimate the effect of arrests on repeat DV.

2.1 Domestic Violence in West Midlands County

Our setting is the policing of domestic violence which occurs in West Midlands County in England. The prevalence of DV in the UK mirrors that of other high-income countries. The World Health Organization estimates that the lifetime incidence of intimate partner violence is 24% in the UK; for comparison, it is 22% in Europe more broadly and 26% in the US (WHO, 2021).

West Midlands is the second largest county in England, with a population of 3 million. The county includes the cities of Birmingham, Coventry, and Wolverhampton. The area is ethnically diverse, with 79% Whites, 11% Asians, 3% Blacks, and 7% Other (Office for National Statistics, 2020). Using survey data, the estimated annual intimate partner victimization rate in the police force area is 6.6%, a percentage which is somewhat higher compared to England as a whole (5.8%).[5]

There are several features of DV which distinguish it from other types of crimes. First, most DV occurs at home (90% of cases in England according to HM Inspectorate of Constabulary (2014)). Second, repeat victimization is common (49% within a year in our data). Moreover, the majority of victims facing repeat episodes of DV are women (89% according to Walby and Allen (2004)). At its most severe, DV can result in mental health breakdowns, bodily injury, or even death (WHO, 2002). In the UK, 44% of female homicide victims are killed by their partners or ex-partners (Office for National Statistics, 2013). These patterns for England in general, and the West Midlands in particular, mirror those for other countries (WHO, 2002, 2021).

2.2 Handling of DV Emergency Calls

In England, 999 is the official telephone number which allows citizens to contact the police for emergency assistance (similar to a 911 call in the US). Figure 1 outlines how 999 emergency

[5] Based on own calculations using British Crime Survey data. The reported numbers are average victimization rates (for both physical and non-physical DV) over the survey years 2004/05 to 2009/10.

calls for DV work their way through the system, beginning with the call itself and ending with the resolution of the case.

Call handlers are the first step in the process. In the West Midlands, 999 calls arrive in one of two main control rooms of a centralized call center. Due to the seriousness of DV offenses, the government requires call handlers to identify and separately classify DV calls. DV includes incidents of threatening behavior, violence, or abuse between intimate partners. DV calls make up 9% of all emergency calls. The call handler records the identity of the caller (victim 33% vs. third party 67%), confirms the GPS location of the incident,[6] and grades the priority level of the incident. This call grade priority ranking determines how quickly officers are expected to respond based on the severity and immediacy of the situation; the goal is to respond within 15 minutes for priority 1 (immediate response) and within 60 minutes for priority 2 (priority/early response). 60% of DV calls are given a priority ranking of 1 and 31% are given a priority ranking of 2. The fact that few DV calls are given a priority ranking of 3 or below is consistent with the mandate that police officers need to defuse DV situations before they escalate.

In the next step, the information collected by call handlers is transferred electronically to the dispatcher. The role of dispatchers is to coordinate first response police teams (i.e., police officers driving in the same patrol car) in the field. To that end, they monitor an electronic map with the real-time locations of all police vehicles within each catchment area. Based on the severity ranking of the call, and the availability of response teams and their proximity to the incident's location, the dispatcher assigns a team of first response officers to the incident.

First response teams are not specialized, but rather respond to all types of emergency 999 calls. The primary purpose of the first response team in DV cases is to protect victims and prevent further harm. To achieve this goal, police officers can separate the offender from the victim, provide advice, and collect information. Officers make a short-term safety plan with the victim, often making them aware of support organizations and safe places in their communities. The most drastic immediate action that can be taken at the scene is to arrest the perpetrator. Arrest is a discretionary action which occurs if there are reasonable grounds to suspect a crime has been committed, if the victim's safety is at risk, or if it will help facilitate the collecting of

[6] Location is recorded with an accuracy of ten by ten meters using the Ordnance Survey National Grid reference system (a 12 digit point reference). The determination of geo-location is a semi-automated process, with call handlers entering address information when the incident occurs at a different location compared to the origination of the cell phone call.

information. Our database records the most severe action taken by the first response team, with (in descending order of severity) arrest occurring 3.1% of the time, a recommended criminal investigation 42.0% of the time, advice/warning 9.3% of the time, and no further action 45.6% of the time. Since only the most severe action is recorded in the dataset, an arrest could also include a recommended criminal investigation or advice.

After completing their on-scene work, first response officers transfer the responsibility of the case to an investigative officer who specializes in domestic violence. The investigative officer decides whether to open a formal investigation, which prompts an entry in a national crime database with unique person identifiers for the victim and perpetrator. Police can search this database when there is a future incident, and eligible parties can make information requests according to the Domestic Violence Disclosure Scheme (also known as Clare's Law). Criminal investigations occur in 59.1% of incidents. During the investigation, further evidence can be gathered and information shared with a prosecutor. The prosecutor works with the investigative officer to determine whether the offender should be formally charged with a crime and summoned to court (1.4% of incidents).[7] During this process, suspects can be taken into custody for a maximum of 24 hours, with a possibility of extended custody up to 96 hours for more serious crimes.[8]

Our analysis is possible due to detailed data which allows us to observe what happens at each stage of the process depicted in Figure 1. We combine information from five different police registries for (i) incidents, (ii) officer deployments, (iii) investigations, (iv) charges, and (v) police personnel records.

2.3 Conditional Random Assignment

The way DV emergency calls are handled leads to conditional random assignment. Dispatchers receive only limited information from the call handlers, making dispatch decisions based on the priority grading of the call and the availability of nearby officers on patrol. While we do not directly observe the availability of officers, we control for predictable staffing needs using a rich set of time and geographic variables. Specifically, we use variables for year, calendar month, day of week, time of day, and bank holidays as well as ward (146 geographic areas

[7] The charging decision usually lies with the prosecutor. However, for summary offenses (those with a maximum punishment of 6 months in prison), investigative officers can directly charge the offender.

[8] The custody decision is made by custody officers and not directly by the first response or investigative officers.

which are subsets of police officer catchment areas). We also control for call grade, as higher ranked calls jump to the top of the dispatch queue. After controlling for these variables flexibly, precisely when and where residual demands for police resources will emerge should be as good as random, something we verify empirically.

Several features of the dispatch process make it unlikely that dispatchers selectively assign response teams to incidents after conditioning on this information. First, the guiding principle in DV incidents is to ensure a speedy police response, with the first available officers being sent to the scene. In our baseline specification, we restrict our sample to DV cases with a call grade of 1 or 2 (91% of all DV calls), as in these cases the mandate is to arrive on the scene as soon as possible. Calls with a grade of 1 are classified as ″immediate response″ and include calls where there is ″A danger to life/use or immediate threat of use of violence/serious injury to a person.″ The directive is for the police officers to arrive within 15 minutes of the call; we observe a mean response time of 8 minutes. Calls with a grade of 2 are classified as ″priority response/early response″ and include calls where there is a ″concern for someone's safety.″ The directive is to arrive within 60 minutes; we observe a mean response time of 26 minutes.[9]

Second, the limited information available to dispatchers argues against non-random assignment. On their electronic map with patrol cars' real-time locations, dispatchers only observe officers' identification numbers without information on their names or background. Dispatchers also never have any direct contact with the individual who calls for help, ruling out that they have extra information on specific victim needs.

A first response team is usually 2 police officers in a patrol car. We measure the team arrest propensity as the total number of DV arrests by all officers in the team divided by the total number of DV cases handled by all officers on a team. In other words, we use the weighted average of individual officers' arrest propensities, where the weights are proportional to the number of cases handled by an officer. To construct the instrument, we use all DV cases an officer has been involved in (minus the current case and any cases at the same address), including both past and future cases, and not just those cases which appear in our estimation sample.

In the construction of our instrument for team arrest propensity, we only consider police officers who are initially dispatched to a DV incident before any other officers arrive, leaving out

[9] Potential harm to individuals is not the only principle used to assign call grades. For example, call grade 1 includes cases where the crime is in progress and call grade 2 includes cases where evidence is likely to be lost if there is a delay.

those who are called to the scene later for reinforcement. Doing so ensures that information about the case collected from officers who are the first to arrive at the scene does not lead to selective reinforcements. We also use the initially dispatched officers to construct the instrument, even if they are not able to respond and a different team is assigned in their place. On average there are 2.6 officers initially dispatched to the scene. When using the instrument, we always condition flexibly on call grade, time, and geography fixed effects. We further require at least 400 total cases for a team so as to minimize any mechanical bias; the rationale for requiring a large number of cases is that arrest occurs rarely (3% of cases). As we will show, results are robust to higher or lower thresholds.

Appendix Table A1 details the sample sizes for various components of our analysis. Our baseline estimation sample includes 124,216 observations, comprised of cases classified as DV by call handlers between 2011-2016, with at least 400 DV cases for the dispatched team, and with a call grade of 1 or 2. The construction of our instrument uses 631,834 officer-case level observations between 2010-2019.

Table 1 tests whether first response teams in our baseline sample are randomly assigned to DV calls. We start by showing which characteristics predict whether an arrest will be made. The first column regresses arrest in the current case on predetermined characteristics, controlling flexibly for call grade, time, and geography. The estimates reveal that a prior DV case, a prior arrest, and a prior criminal charge all strongly predict whether an arrest will be made in the current case. Who initiates the call, the gender of the call handler, and the experience of the call handler have no statistically significant effect. The joint F statistic for all of these variables is highly significant (p-value < 0.001).

We next show that despite the predictive power of these case characteristics, they are uncorrelated with our instrument. The second column regresses team arrest propensity on the same set of characteristics, controlling flexibly for call grade, time, and geography. The estimates are all close to zero and statistically insignificant, both individually and as a group (p-value=.478). This provides empirical support for first response teams being conditionally randomly assigned to DV calls. Note that for this table, we multiplied the arrest and team propensity variables by 100 so that the table would be more readable (i.e., to avoid estimates of 0.000). We do not multiply these variables by 100 anywhere else in the paper.

2.4 Regression Model

We are interested in estimating the effect of arrests on future victimization, which we model with the regression

$$DV_{i,t+1} = \beta_0 + \beta_1 A_{i,t} + X'_{i,t}\delta + \epsilon_{i,t} \tag{1}$$

where the outcome variable $DV_{i,t+1}$ measures whether repeat domestic violence occurs in the next period. $A_{i,t}$ is an indicator for whether an arrest is made in the current case, $X'_{i,t}$ is a vector of flexible controls for call grade, time, and geography, and $\epsilon_{i,t}$ is an error term. Specifically, in our preferred specification, we include fixed effects for call grade (1 versus 2), year, calendar month, day of week, time of day (6 hour intervals), and bank holidays, each interacted with ward fixed effects.

For our measure of repeat victimization $DV_{i,t+1}$, we use the officers' coding of future DV, even if the future case was not initially classified as DV by the call handler. Note, however, that when constructing our instrument we only use calls initially classified as DV by the call handler to avoid the possible endogeneity of officers' classifications.

Our research design recognizes that arrests are unlikely to be random. More serious incidents of domestic violence are likely to both increase the probability of arrest and lead to future victimization. Failure to account for the endogeneity of arrest may therefore lead to an underestimate of any deterrent effect or could yield a positive estimate even in the absence of backlash. We therefore use the average arrest propensity of the first response team as an instrument for arrest in the current case. The intuition for our team arrest propensity instrument is that a suspected batterer will more likely be arrested if the police officers handling the current case have a higher average arrest rate in other DV cases they handle.

The first stage regression can be written as

$$A_{i,t} = \alpha_0 + \alpha_1 Z_{j(i)} + X'_{i,t}\gamma + \eta_{i,t} \tag{2}$$

where the scalar $Z_{j(i)}$ denotes the arrest propensity of the first response team j assigned to case i, as defined in Section 2.3. We use instrumental variable regression based on equations (1) and (2) to estimate the causal effect of arrest on future victimization. We cluster standard errors at the level of the dispatched officer on a team with the most domestic violence cases.

3 Assessing the Instrument

3.1 Relevance

Key to our design is that not only are first response teams conditionally randomly assigned, but also that they differ in their arrest propensities. The histogram in Figure 2 displays the distribution of arrest propensities. In constructing this histogram, we first regress out the conditioning variables for call grade, time, and geography to match the variation used in our analysis. The mean of the instrument is 0.030 with a standard deviation of 0.012. The figure reveals substantial heterogeneity in arrest propensities, with the 1-99 percentile range spanning arrest rates of 0.012 to 0.060.[10]

Figure 2 also graphs the probability the suspected offender will be arrested in the current case as a function of the arrest propensity of the first response team. The solid line plots estimates from a local linear regression, and hence represents a flexible analog of the first stage. The predicted line is monotonically increasing in the instrument and close to linear. First stage estimates based on equation (2) are reported in the bottom panel of Table 3, and are all highly significant. The Kleibergen-Paap Wald F statistics reveal that our instrument is not weak. Column (4) includes our preferred set of conditioning variables, and indicates that being assigned a first response team with a 1 percentage point higher arrest rate increases the probability of being arrested by 0.722 percentage points.

Note that the first stage estimate need not mechanically equal one as the number of cases per team goes to infinity for several reasons: (i) the sample of cases used to calculate the instrument is not the same as the estimation sample, (ii) there are covariates (i.e., call grade, time, and geography controls), and (iii) the instrument is calculated by weighting the arrest propensities of different officers in a team, where teams are not held constant over time. Hence, there is no reason to expect a coefficient of one in our setting.

[10] A natural question is why some police officers are more prone to arrest. A simple regression reveals that teams with fewer female officers are more likely to arrest, while average officer age (which proxies for experience) has no impact. Later, we show our findings are robust when also including these average team characteristics in our regressions. Importantly, other factors which we do not measure account for the overwhelming share (99%) of the residual variation in arrest propensities (after regressing out call grade, time, and geography).

3.2 Validity

For the instrument to be valid, arrest propensities must be conditionally independent of case characteristics that affect the likelihood of repeat DV. In Section 2.3, we argued the assignment of DV emergency calls to first response teams should be random after conditioning flexibly on call grade, time, and geography, and tested for balance on predetermined observable characteristics in Table 1. As a second test, we add in these same predetermined characteristics to the first stage. If teams are randomly assigned, these characteristics should be uncorrelated with the instrument, and hence not significantly change the estimate. As expected, the first stage coefficient does not change appreciably (0.721 versus 0.722).

3.3 Monotonicity

If the effect of an arrest is homogeneous across DV incidents, then conditional random assignment and exclusion are enough for IV to capture the causal effect of arrest. If effects are heterogeneous, then the instrument also needs to satisfy monotonicity. With monotonicity, IV can be interpreted as the local average treatment effect (LATE) for compliers – i.e., the arrest effect for cases which would have a different arrest outcome if they had been assigned a first response team with a higher or lower arrest propensity. These compliers are particularly relevant for policy, as any changes in arrest guidelines for police officers are likely to be targeted towards these marginal DV cases.

In our setting, monotonicity requires that if an arrest is made by a lenient officer team (i.e., a team with a low arrest propensity), then it would also have been made by a stricter team, and vice versa when an arrest does not occur. One testable implication is that the first stage estimates should be the same sign for any subsample of the data. To implement this test, we construct the instrument using the entire sample of cases, but estimate the first stage on specific subsamples: by incident order, by DV hotspot area, and by time of day. Panel A of Table 2 reveals that the estimated first stage coefficients are all positive and statistically significant, as expected if monotonicity holds.

Panel B of Table 2 conducts the ”reverse instrument” test for monotonicity proposed by Bhuller et al. (2020). The test is based on the implication that response teams which have higher arrest propensities for one case type (e.g., first time callers) should also have higher arrest propensities for other case types (e.g., higher order callers). To test this, we break the data into the same subsamples as in panel A, but redefine the instrument for each subsample to be the arrest propensity of the team for cases outside the subsample. Consistent with the monotonicity requirement that officer teams which are more arrest prone in one case type also being more arrest prone in other case types, the first stage estimates are all positive.

3.4 Exclusion

The exclusion restriction for the IV estimate requires that a first response team with a higher arrest propensity only affects repeat DV by increasing the probability of an arrest. After discussing our main results, we present tests related to two concerns about the exclusion restriction: (i) that teams with a higher arrest propensity affect recidivism not only through the arrest channel but also through the other actions they take as part of their on-scene police work and (ii) that a team's arrest propensity correlates with other team characteristics that affect the probability of repeat victimization directly. These tests indicate that the exclusion restriction is likely to hold.

4 Effect of Arrest on Repeat DV Calls

We now examine how arrest affects repeat emergency DV calls. In our data, we do not observe the gender of the victim, so our analysis concerns victims of both genders. However, since most victims of DV are female,[11] for simplicity we sometimes refer to the victim as a woman and the offender as a man.

Column (1) of Table 3 starts by reporting results using OLS, where the outcome variable is a repeat 999 DV call within the ensuing 12 months. For comparability with our preferred IV specification, the OLS regression includes the same set of call grade, time, and geography controls. The OLS estimate of the effect of arrest on a repeat DV call is a precisely estimated zero.

The IV estimates in Table 3 stand in sharp contrast to OLS. Regardless of how we control for call grade, time, and geography across columns (2)-(4), there is a roughly 50 percentage point reduction in repeat DV calls after an arrest. Our preferred specification is found in column (4),

[11] According to Walby and Allen (2004), women are the victim in DV 89% of the time in the UK.

which includes the most flexible set of controls for call grade, time, and geography. Focusing on our baseline estimate, if a suspected batterer is arrested on the scene, the probability of a repeat emergency call for DV falls by 48.8 percentage points. To help interpret this magnitude, the table reports the estimated control complier mean of the outcome (i.e., the mean for compliers where there is not an arrest), which is 96.2%.[12] Hence, our IV estimate represents a sizable 51% reduction in repeat calls for compliers.

The estimated control complier mean is substantially larger compared to the overall mean of 49.2%, but not unexpectedly so. The reason is that a large fraction of DV emergency calls are relatively minor in terms of not meriting an arrest from any response team. We calculate that never takers (cases where no police officer would make an arrest) comprise 94.3% of our sample. The cases where arrest could be considered a commensurate response are those which are serious and likely to result in repeat victimization. But not all police officer teams choose to arrest in these cases, as teams with low arrest propensities may believe that arrests almost always do more harm than good in terms of the severity of repeat victimization, for example. In these cases without an arrest, the probability of revictimization will be very high.

This argument is further supported by the estimated characteristics of compliers in our sample. We estimate the fraction of compliers to be 4%. In Appendix Table A2 we characterize compliers by their past DV history.[13] Compared to the overall population, compliers are more likely to be serial offenders (column 1) who have committed more serious crimes (column 4). These complier cases appear to part of increasing pattern of DV, with these offenders having evaded arrest at a higher rate in the past (column 2).

One possible explanation for the divergence between the IV and OLS estimates in Table 3 is that the effect of arrest is heterogeneous, differing for compliers versus the entire population. To explore this possibility, we characterize compliers by the four characteristics describing past DV history appearing in Table 1: whether there was a prior DV case, a prior arrest, a prior investigation, and a prior charge.[14] The OLS estimate for this complier reweighted sample is

[12] We estimate the control complier mean following the Technical Appendix to Dahl et al. (2014). We use the 1st and 99th percentiles of team arrest propensity to define the least and most stringent teams, combined with the estimated coefficients from a linear first stage, to calculate the fraction of compliers, always takers, and never takers. Further details are found in Appendix B.

[13] Our characterization of compliers adapts the binary-instrument methodology proposed by Marbach and Hangartner (2020) to a setting with a continuous instrument. See Appendix B for details.

[14] We first split the sample into subgroups based on combinations of these prior characteristics, recognizing that not all combinations are possible (e.g., there must be a prior case to have a prior arrest). We combine prior cases where there was an arrest but no investigation with prior cases where there was an arrest and an investigation, as the former is rare. This yields 6 different subgroups. We then estimate the first stage separately by subgroup so

0.040 (s.e.=0.014), suggesting that the difference in estimates is not driven by heterogeneous effects, at least based on the observable characteristics available to us.

An alternative explanation for the divergence between the OLS and IV estimates is selection bias. The OLS estimate is likely to be biased upwards, as cases which have characteristics associated with higher recidivism probabilities are also plausibly more likely to result in arrest. The bias arises because these case characteristics are observed by police officers, but not by the researcher. Our IV estimates reveal that OLS would lead to the mistaken conclusion that arrest has no effect on repeat DV calls, when in fact, arrest results in a halving of repeat calls.

These results provide compelling evidence that repeat emergency calls fall by a large amount after an arrest. But this arrest effect can only be interpreted as breaking the cycle of domestic violence if it is not driven by changes in reporting behavior. DV is notoriously underreported, and arrest might affect the probability of reporting. If arrest encourages DV reporting, our estimates represent a lower bound of the arrest effect on repeat DV abuse. However, it is also possible that arrest discourages victims from reporting future DV incidents, in which case, the arrest effect we estimate could reflect a change in reporting behavior rather than a reduction in incidence. Given the importance of figuring out whether arrest changes DV incidence versus reporting behavior, we defer further robustness checks, heterogeneity analyses, and tests of mechanisms until after the next section where we tackle this issue.

5 A Reduction in Incidence or Reporting?

Whether the drop in repeat DV calls after an arrest is good or bad from a policy perspective depends on whether it is driven by a reduction in incidence or merely a change in reporting behavior. If there is a drop in abusive incidents after an arrest, victims are better off. If instead there is a drop in reporting without a reduction in actual abuse, victims are worse off because they are not getting help from police to resolve emergency situations, which in turn could embolden men to commit even more abuse.

In this section, we provide two tests based on the composition of repeat DV calls for whether the reduction reflects a drop in incidence versus a change in reporting behavior. Both tests point to a reduction in incidence as the primary explanation.

that we can calculate the fraction of compliers in each subgroup. Finally, we reweight the OLS estimation sample so that the proportion of compliers in each subgroup matches the share of the estimation sample for that subgroup.

5.1 Test 1: Severity of Repeat DV Calls

Our first test is based on a simple threshold model for reporting behavior, with testable implications regarding the severity of repeat DV calls. The model builds on those used by Dahl and Knepper (2021) and Boone and Van Ours (2006) in the contexts of workplace sexual harassment and workplace safety, respectively.

In our threshold model, a woman will call the police if the immediate benefit of police intervention exceeds the retaliatory costs of reporting. Let θ be the retaliatory abuse a woman expects to occur if she calls the police. We model the benefit μ of calling the police to be a function of the level of abuse α perpetrated by a man against his partner, where we assume $\partial\mu(\alpha)/\partial\alpha > 0$. The idea is that in the heat of the moment, intervention by a police officer reduces harmful violence, and more so the more serious the abuse. Since benefits are increasing, we can define the threshold level of abuse above which a woman will report as the level of $\alpha = \bar{\alpha}$ that equates benefits to costs.

Now consider the impact of an arrest, ceteris paribus. On the one hand, an arrest could disempower victims from reporting DV in the future. This might happen if arrest causes backlash. In this case, the cost of calling the police in the future rises and hence the reporting threshold $\bar{\alpha}$ will be higher. The intuition is that women experiencing backlash after an arrest are more likely to remain silent, only reporting when abuse is so severe that the immediate benefit of police intervention outweighs the higher expected level of retaliation. On the other hand, an arrest could empower women to report by signaling to them that something is done about DV. In this scenario, the cost of calling the police decreases, and there will be a reduction in the reporting threshold, with women tolerating less abuse before calling the police. The testable implication of this model is that the composition of repeat DV calls after an arrest should be more severe on average if women are disempowered from reporting, whereas repeat calls should be less severe on average if empowerment dominates.

Extending the model to account for the optimal responses of men only reinforces these predictions. Men's optimal response in a world where an arrest disempowers victims from reporting is to increase the amount and severity of DV they commit. With empowerment, men will instead optimally decrease both the amount and severity of DV.[15]

[15] One model that would yield this prediction is as follows. Consider two partners who play a sequential game with perfect information. After an initial DV incident, the male partner moves first and decides whether to repeat DV and, if so, chooses the severity α of it. The female partner observes these choices, and if subjected to repeat DV,

We use the call handler′s grading of repeat DV calls to measure severity. We define severe repeat cases as those with a call grade of 1 (61% of repeat DV calls), i.e., those where there is a ″danger to life/use or immediate threat of use of violence/serious injury to a person.″ We define less severe repeat cases as those with a call grade of 2 (29% of repeat calls), i.e., those where there is a ″concern for someone′s safety″ plus call grades lower than 2 (10% of repeat calls), where the case is even less serious.

The first column of Table 4 reports our baseline estimate for comparison purposes. Columns (2) and (3) decompose the baseline estimate by severity. Each column is a separate regression which uses the entire sample and the same instrument, but column (2) uses repeat calls which are severe as the outcome, while column (3) uses repeat calls which are less severe. By construction, the estimates in columns (2) and (3) will sum to the estimate in column (1).

As column (3) shows, there is a 55.2 percentage point drop in severe DV calls after an arrest. Relative to the estimated control complier mean of 83.0%, this is a sizeable 67% reduction. In contrast, column (2) reveals a 6.4 percentage point *increase* in the number of less severe DV calls after an arrest. Relative to the control complier mean of 13.2%, this is a 48% rise. These compositional effects are consistent with a drop in victim′s reporting thresholds and an accompanying decline in the level of abuse. Backlash would have predicted a compositional shift to more severe cases being reported, which is exactly the opposite of what we find. A formal test rejects the null hypothesis of backlash versus the alternative of empowerment (p-value=0.036).[16] Since both the volume and severity of repeat calls fall after an arrest, we conclude that the true incidence of repeat DV falls after an arrest.

5.2 Test 2: Who Reports Repeat DV Calls

Our second test is based on the composition of who reports repeat DV. Using a similar threshold model, if victims are more subject to backlash than neighbors, then the composition of

decides whether to report according to the threshold model we have outlined. The utility function of the husband is of a piecewise form: $u = \phi$ if he decides not to repeat DV, where ϕ measure the utility he obtains from a violence-free relationship; $u = v(\alpha)$ if he repeats DV and the female partner does not report it ($\alpha < \bar{\alpha}$), where $v(\alpha)$ measures the utility he derives from it as a function of severity α; and $u = v(\alpha) - c$ if he repeats DV and the wife reports it ($\alpha \geq \bar{\alpha}$), where c is the cost to the husband of being reported to the police. Assuming that $v(\cdot)$ is a single-peaked function with a maximum at some α^*, comparative statics on the sub-game perfect equilibrium of this two-stage game give rise to the predictions described in the text.

[16] We conduct a one-sided test of the null hypothesis of backlash versus empowerment by restricting the sample to cases with a repeat call, defining the outcome to be a dummy variable for a severe case, and using our preferred IV specification. The estimated effect of an arrest is -0.452 (s.e.=0.251).

future calls after an arrest should be skewed towards those being reported by third parties rather than the victim. The intuition is that victims facing backlash will remain silent for fear of retaliation, while such concerns should be smaller for neighbors as they do not live with the abuser and can report anonymously. In contrast, if victims are empowered relative to their neighbors, the composition of future calls should tilt towards victims.

In columns (4) and (5) we test these predictions by decomposing the baseline estimate into who reports. Column (4) shows that repeat DV calls by a victim fall by 9.9 percentage points after an arrest, which is 36% drop relative to the estimated control complier mean. Column (5) reveals a proportionately larger decrease in repeat DV calls by a third party. There is a 39.0 percentage point reduction, which is a 57% drop relative to the control complier mean.

The fact that repeat calls by third parties fall by 57%, while victim calls fall by only 36%, is the opposite of what backlash would predict. Instead, the compositional shift is consistent with empowerment, with actual abuse falling and women reporting abuse at a lower threshold. Since the overall and relative volume of victim repeat calls fall after an arrest, this is consistent with a decline in true incidence after an arrest. However, we cannot reject the model of backlash at conventional significance levels (p-value = 0.354).[17]

6 Mechanisms

In this section, we investigate factors which could explain the results we document, finding that (i) arrest has a short-term incapacitation effect as well as a longer-term deterrent effect and (ii) arrest leads to consequential legal sanctions for the offender.

6.1 Short-Term and Longer-Term Effects of Arrest

In the first three columns of Table 5, we explore the short-term effects of arrest. The time window we focus on in column (1) is the first 96 hours after an arrest. Two pieces of evidence highlight the importance of arrests in the short-run. First, the control complier mean for a repeat DV call within 96 hours is 25 percentage points. Second, and strikingly, if a suspected batterer is arrested on the scene, the probability of a repeat DV call within 96 hours falls by 20

[17] Using a similar test as in footnote 16, the estimated effect of an arrest on the outcome of a third party report is -0.105 (s.e.=0.280).

percentage points, which amounts to a 79% reduction.

One possibility is that this sharp drop is due to an incapacitation effect. In columns (2) and (3) we further explore this by splitting the 96-hour time window into hours 1-48 and hours 49-96, respectively. To get some insight into a potential incapacitation effect, hours 1-48 are of particular interest, since this is when offenders are typically under investigation and are potentially placed in temporary custody. Comparing the IV estimates in columns (2) and (3), we see that almost the entire short-term effect of arrest is explained by a sharp drop in repeat DV calls in hours 1-48. We interpret this as evidence that an import part of the arrest effect is driven by short-term incapacitation. Given that DV appears to be disproportionately common in the first few days after an initial emergency call, this suggests the ”cooling off” period provided by an arrest is highly effective.

Columns (4) to (6) of Table 5 investigate the longer-term effects of arrest. As the dependent variable, column (4) uses a dummy for whether there is at least one repeat DV case within 12 months, excluding the first 96 hours. An arrest leads to a statistically significant 45 percentage point decrease in repeat calls during this time period. Note that the arrest effects in columns (1) and (4) do not need to sum to our main estimate appearing in Table 3. The reason is that repeat victimization can occur in more than one time period. In columns (5) and (6), we focus instead on whether repeat cases are reported in months 1-6 (again omitting the first 96 hours) and months 7-12, respectively. For both time windows, we find that arrest reduces the probability of a repeat call by roughly 50% of the control complier mean, though the estimate for repeats in months 7-12 is not statistically different from zero. Thus, besides causing a strong short-term incapacitation effect, arrest appears to persistently reduce repeat calls both 1-6 and 7-12 months subsequent to the initial incident.

6.2 Criminal Sanctions

Sceptics of arrest argue that it will not have a deterrent effect due to its weak “dosage” of punishment, with offenders and victims learning that the consequences of an arrest are not very serious.[18] Advocates for a deterrence effect argue that arrest makes salient the seriousness of the crime, in part due to a higher probability of criminal sanctions.

[18] Related, one of the most commonly cited reasons for not reporting DV is the perception that the police do not take cases seriously enough or have no means of providing help. See National Research Council (1998) and Fleury et al. (1998).

If a DV case is formally investigated, the investigative officer coordinates with the prosecutor to determine whether the offender should be formally charged with a crime and summoned to court. Non-arrested compliers have an estimated 2% probability of being charged. As column (7) in Table 5 shows, an arrest increases this probability by a statistically significant 10.4 percentage points, which is a five-fold increase. Criminal charges can lead to a conviction with the possibility of jail time, community service, or parole, although we do not observe these outcomes in our data.[19] Based on the large increase in criminal charges, combined with the persistent reduction in repeat calls shown in columns (4)-(6) of Table 5, we conclude that arrest has a sizeable deterrence effect.

7 Measurement, Exclusion Restriction, Robustness, and Heterogeneity

We now return to the task of probing (i) the robustness of our outcome measure, (ii) the exclusion restriction, (iii) specification checks of our main results, and (iv) heterogeneous effects of arrests.

7.1 Using Geo-Coordinates to Measure Repeat Calls

We measure repeat DV emergency calls based on whether a new DV case with the same geo-coordinates is reported within 12 months of the initial call. The measure is potentially prone to misclassification errors. We might over-assign repeat cases if several households live at a given geo-location, or under-assign repeat cases if victims move away from it. In this section, we assess whether these two types of misclassification error are likely to bias our estimates.

To minimize the chances of over-assigning repeat cases, we exploit features of the built environment in the sample area. High-rise buildings with multiple apartments in the same geo-location can lead to DV cases being reported by different victims. This problem should be less of a concern if we focus on areas largely made up of detached houses, which is the most common type of accommodation in the region we study. In column (1) of Appendix Table A3, we restrict our sample to wards with at least 80% detached houses. Although the sample shrinks

[19] Having a criminal record is consequential in the UK, as employers are are allowed to ask about recent convictions on job applications. Punishment for future crimes is also influenced by prior criminal sanctions.

by a third, the arrest effect remains largely unchanged compared to our main estimate in Table 3. In column (2), following a similar argument, we exclude the city center of Birmingham—the largest city in the county—by excluding areas within a 3 km radius of the city center, where mostly larger buildings are located. The estimate for arrests remains close to the baseline estimate, consistent with over-assignment of repeat cases not confounding our results.

A second potential bias comes from under-assigning repeat cases. A particular concern is that victims of DV might be more likely to make residential moves after an arrest, but continue to be victimized in their new location. To assess this concern, we check whether the accuracy of our geo-coded repeat variable is affected by an arrest. To do so, we proceed in two steps. We first draw on the subsample of our data where we observe unique victim IDs (the 59% of cases where a criminal investigation occurs), and construct a dummy for re-victimization based on their victim ID. We then compare this variable to our geo-coded re-victimization measure for the same subsample. This allows us to work out how many victim-ID based repeat cases we miss with our geo-based definition. Overall, 91% of the no-repeats based on our geo-coded measure turn out to also be no-repeats based on the victim-ID measure, while 9% of our geo-coded non-repeats miss that the victim is actually re-victimized. Importantly, the false negative rate is not significantly different across cases with and without a prior arrest, implying that the misclassification error is not correlated with our treatment variable. Appendix Table A4 summarizes this sensitivity check.

7.2 Exclusion Restriction

The IV estimate requires the exclusion restriction that first response teams with higher arrest propensities only affect repeat DV calls by increasing the probability of an arrest. In this section, we provide a series of exclusion tests related to (i) other actions taken by responding officers as part of their on-scene police work and (ii) characteristics of responding officers.

The on-scene actions by first response teams are multidimensional in nature. The first priority of responding officers is to make the victim safe, which may mean arresting a suspect if it is considered a necessary and proportionate response. In addition, responding officers are also tasked with taking steps to build a case for a potential evidence-led prosecution. This includes, *inter alia*, collecting evidence, convincing victims to cooperate, and communicating with other involved parties (e.g., witnesses). In this setting, a violation of the exclusion restriction would arise if officers with higher arrest propensities are better at building evidence-led cases. This issue can be addressed by augmenting our IV model to include a measure for the quality of a team's on-scene work as an additional endogenous regressor plus an extra instrument for it. We proxy for the quality of a team's on-scene work using whether the case ends up being formally investigated. Indeed, how well responding officers carry out their on-scene tasks is a critical input into whether an investigative officer opens a formal investigation and enters the case in the national crime database (College of Policing, 2022).

We construct an instrument for formal investigations that is similar to our instrument for arrest: the propensity of a first response team's cases to be formally investigated. There is independent variation in formal investigations and arrests. Formal investigations happen much more often (59% of the time versus 3% for arrest), and an arrest is only followed by a formal investigation 67% of the time. As column 2 in Appendix Table A5 shows, the instrument for formal investigation strongly predicts whether an investigation will happen. Table 6 reports IV estimates, with column 1 repeating our main specification for arrests and column 2 adding in formal investigations. Formal investigations reduce repeat DV calls by a modest, and statistically insignificant, 5 percentage points. More importantly, our main arrest effect is virtually unchanged. Thus, our arrest finding does not appear to be driven by how effective police are at doing their job.

One directly observable part of the work that first response officers do is how much time they spend on the scene. We use this variable for a second test of the exclusion restriction. A plausible interpretation of extra minutes spent is that officers are more diligent, take the needed time to finish their tasks, and spend more time with victims and offenders. As above, the concern is that officers with higher arrest propensities could influence recividism not only through arrest but also by how much time they spend on the sence. In column 3 we include time on the scene as an additional endogenous variable, instrumenting it with a team's average time spent on the scene in other cases.[20] While the instrument is strongly predictive (Appendix Table A5, column 4), time spent on the scene does not affect repeat DV calls, nor does it change the estimated effect of arrest (Table 6, column 3).

In our dataset, response teams are recorded as taking one of four actions (in order of decreasing

[20] The time on scene variable is missing in 30.8% of cases. It appears to be missing at random, as the team's leave-out average time spent on the scene regressed on whether the variable is missing yields a precise zero (controlling for call grade, time, and ward fixed effects). We therefore replace missings with the mean time spent in non-missing cases, and add an indicator for missing in the regression.

severity): arrest, recommend an investigation, provide advice, or do nothing. So far, we have been using arrest as our key independent variable and instrumenting for it. As a third test of the exclusion restriction, in column 4 of Table 6 we include variables for these other actions, instrumenting for them using officer team propensities along each of these margins. While the data only records the most severe intervention, it typically includes the less severe actions as well. Therefore we code the action of arrest as also including recommend investigation and provide advice, and the action of recommend investigation as also including provide advice.[21] The arrest estimate remains virtually unchanged, while the coefficients for recommend investigation and advice are statistically insignificant. In column 5, we add in all of the endogenous regressors found in columns 2-4 simultaneously, and the results remain unchanged.

As a fourth, and final, test of the exclusion restriction, we probe whether team characteristics could cause a violation of the exclusion restriction. A prime example is the gender composition of response teams, as female officers are thought to increase the quality of DV-related police work (Miller and Segal, 2019). Experience could also matter. This source of bias can be eliminated by controlling for team characteristics. We observe both the gender and age of officers. Thus, in column 6 of Table 6, we control for the fraction of females on the team and the average age of officers. This does little to change the estimated effect of arrest. As an alternative to this test, in Appendix Table A6 we instead control for the additional team propensities used as instruments in Table 6. This also does not appreciably affect our arrest estimate.

7.3 Alternative Specifications

In Appendix Table A7, we probe the robustness of our result to alternative specifications. Column (1) changes the definition of the outcome variable, focusing now on the intensive instead of the extensive margin. The new outcome is the *number* of repeat DV calls within 12 months, and we use the same set of controls as our baseline specification in column (4) of Table 3. The point estimate indicates a statistically significant drop of 1.5 calls after an arrest. Relative to the estimated control complier mean of 2.9 calls, this is roughly a 50% reduction. This intensive margin result closely mirrors the magnitude of the effect for the extensive margin.

The next two columns explore robustness to changing how the instrument is constructed. Our estimation sample is limited to current cases with a call grade of 1 or 2. The rationale is that

[21] If we alternatively code up arrest, recommend investigation, and provide advice as three non-overlapping actions, the results are similar.

call grades of 1 or 2 must be dealt with immediately, with the closest available officers sent to the scene and hence leaving little room for nonrandom officer assignment. However, to construct the instrument, we do not require a randomized set of cases for validity. Therefore in our baseline specification, we use all call grades to construct the instrument. In column (2), we report estimates where we limit the construction of the instrument to calls with a priority grade of 1 or 2. The estimate is virtually identical to our baseline estimate. In column (3), we define the instrument using the traditional leave-one-out measure which only excludes the current case (instead of excluding all calls from the same geo-location, whether it is the current case or any other case). This does not appreciably change the estimate.

In our main specification, we include any DV emergency call for which the assigned response team has handled at least 400 other DV cases. Columns (4) and (5) investigate what happens if we instead require response teams to have handled at least 300 cases or at least 500 cases, respectively. When applying these alternative thresholds, the results remain qualitatively similar.

7.4 Heterogeneous Effects

In Appendix Table A8 we explore whether there are heterogeneous effects by case characteristics. In column (1), we present IV results that distinguish between emergency DV calls with and without a prior DV investigation. We construct a dummy variable for whether a call has been preceded by a formally investigated DV case in the past 12 months and interact it with our arrest variable. There is some evidence that the effect is larger for calls with a prior investigated case, but the difference is not statistically significant. Column (2) similarly divides the sample into two groups, but now based on the predicted probability of an arrest estimated using the case characteristics appearing in Table 1. The arrest effect is fairly similar across calls with high versus low arrest propensities.

The next two columns show IV estimates broken down by the characteristics of response teams. In column (3), we distinguish between all-male response teams and teams with at least one female officer. The arrest effect is similar for both types of response teams. Column (4) breaks the sample down into calls handled by response teams whose average age (a proxy for experience) is above or below the mean. Although the statistically significant point estimate for younger teams is larger than the non-significant estimate for older teams, the two are not statistically different from each other.

8 Conclusion

Our findings provide compelling evidence that arrest helps break the cycle of domestic violence. Using a rich dataset and quasi-random variation, we find that an arrest reduces future DV calls in the following year by 51%. We provide empirical evidence that the reduction in calls is not driven by a change in reporting behavior due to a fear of retaliation, but rather a decline in repeat victimization. In terms of mechanisms, we find that arrest virtually eliminates the large spike in re-victimization which occurs in the 48 hours after an emergency call, suggesting arrest facilitates a cooling off period during a volatile, at-risk time. Arrests also result in a 5-fold increase in the probability an offender will by charged with a crime, consistent with the longer-run deterrence effect we document.

These findings bear on recent proposals to decriminalize DV. In our setting where the arrest rate is low, the optimal policy response would be to arrest more suspected batterers if the objective is to reduce future abuse. We caution, however, that our results do not necessarily imply that arrest should occur in all cases and in all settings; for example, in countries where the arrest rate is high, the pendulum could well have swung too far in the other direction. Future research for other countries and in other contexts can help shed light on this issue.

References

Aizer, A. (2010). The gender wage gap and domestic violence. *American Economic Review*, 100(4):1847–59.

Aizer, A. (2011). Poverty, violence, and health the impact of domestic violence during pregnancy on newborn health. *Journal of Human Resources*, 46(3):518–538.

Aizer, A. and Dal Bo, P. (2009). Love, hate and murder: Commitment devices in violent relationships. *Journal of Public Economics*, 93(3-4):412–428.

Aizer, A. and Doyle Jr, J. J. (2015). Juvenile incarceration, human capital, and future crime: Evidence from randomly assigned judges. *Quarterly Journal of Economics*, 130(2):759–803.

Alesina, A., Brioschi, B., and La Ferrara, E. (2021). Violence against women: A cross-cultural analysis for africa. *Economica*, 88(349):70–104.

Angrist, J. D. (2006). Instrumental variables methods in experimental criminological research: What, why and how. *Journal of Experimental Criminology*, 2(1):23–44.

Autor, D., Kostol, A., Mogstad, M., and Setzler, B. (2019). Disability benefits, consumption insurance, and household labor supply. *American Economic Review*, 109(7):2613–54.

Berk, R. A. (1993). What the scientific evidence shows: On average, we can do no better than arrest. In Gelles, R. J. and Loseke, D. R., editors, *Current Controversies on Familiy Violence*, pages 323–336. Sage Publications.

Berk, R. A., Campbell, A., Klap, R., and Western, B. (1992). A Bayesian analysis of the Colorado Springs spouse abuse experiment. *Journal of Criminal Law and Criminology*, 83(1):170–200.

Bhalotra, S., GC Britto, D., Pinotti, P., and Sampaio, B. (2021). Job displacement, unemployment benefits and domestic violence. *CEPR Discussion Paper No. DP16350*.

Bhuller, M., Dahl, G. B., Løken, K. V., and Mogstad, M. (2020). Incarceration, recidivism, and employment. *Journal of Political Economy*, 128(4):1269–1324.

Black, D. A., Grogger, J., Sanders, K., and Kirchmaier, T. (2022). Criminal charges, risk assessment, and violent recidivism in cases of domestic abuse. *Working Paper*.

Boone, J. and Van Ours, J. C. (2006). Are recessions good for workplace safety? *Journal of Health Economics*, 25(6):1069–1093.

Chin, Y.-M. and Cunningham, S. (2019). Revisiting the effect of warrantless domestic violence arrest laws on intimate partner homicides. *Journal of Public Economics*, 179:104072.

College of Policing (2022). First response: Authorised professional practice. `https://www.college.police.uk/app/major-investigation-and-public-protection/domestic-abuse/first-response`. Accessed: 2022-11-03.

Currie, J., Mueller-Smith, M., and Rossin-Slater, M. (2022). Violence while in utero: The impact of assaults during pregnancy on birth outcomes. *Review of Economics and Statistics*, 104(3):525–540.

Dahl, G. B. and Knepper, M. M. (2021). Why is workplace sexual harassment underreported? The value of outside options amid the threat of retaliation. Working Paper 29248, National Bureau of Economic Research.

Dahl, G. B., Kostøl, A. R., and Mogstad, M. (2014). Family welfare cultures. *Quarterly Journal of Economics*, 129(4):1711–1752.

Di Tella, R. and Schargrodsky, E. (2013). Criminal recidivism after prison and electronic monitoring. *Journal of Political Economy*, 121(1):28–73.

Dobbie, W., Goldin, J., and Yang, C. S. (2018). The effects of pretrial detention on conviction, future crime, and employment: Evidence from randomly assigned judges. *American Economic Review*, 108(2):201–40.

Dobbie, W. and Song, J. (2015). Debt relief and debtor outcomes: Measuring the effects of consumer bankruptcy protection. *American Economic Review*, 105(3):1272–1311.

Doyle Jr, J. J. (2007). Child protection and child outcomes: Measuring the effects of foster care. *American Economic Review*, 97(5):1583–1610.

Doyle Jr, J. J. (2008). Child protection and adult crime: Using investigator assignment to estimate causal effects of foster care. *Journal of Political Economy*, 116(4):746–770.

Dunford, F. W., Huizinga, D., and Elliott, D. S. (1990). The role of arrest in domestic assault: The Omaha police experiment. *Criminology*, 28(2):183–206.

Erten, B. and Keskin, P. (2021). Trade-offs? the impact of WTO accession on intimate partner violence in Cambodia. *The Review of Economics and Statistics*, 12:1–40.

Fleury, R. E., Sullivan, C. M., Bybee, D. I., and Davidson II, W. S. (1998). Why don't they just call the cops?: Reasons for differential police contact among women with abusive partners. *Violence and Victims*, 13(4):333–346.

French, E. and Song, J. (2014). The effect of disability insurance receipt on labor supply. *American Economic Journal: Economic Policy*, 6(2):291–337.

Golestani, A., Owens, E., and Raissian, K. (2021). Specialization in criminal courts: Decision making, recidivism, and re-victimization in domestic violence courts in Tennessee. *Unpublished manuscript*.

González, L. and Rodríguez-Planas, N. (2020). Gender norms and intimate partner violence. *Journal of Economic Behavior & Organization*, 178:223–248.

Goodmark, L. (2018). *Decriminalizing domestic violence*. University of California Press.

Grogger, J., Gupta, S., Ivandic, R., and Kirchmaier, T. (2021). Comparing conventional and machine-learning approaches to risk assessment in domestic abuse cases. *Journal of Empirical Legal Studies*, 18(1):90–130.

Hirschel, J. D. and Hutchison III, I. W. (1992). Female spouse abuse and the police response: The Charlotte, North Carolina experiment. *Journal of Criminal Law and Criminology*, 83(1):73–119.

HM Inspectorate of Constabulary (2014). Everyone's business: Improving the police response to domestic abuse. `https://www.justiceinspectorates.gov.uk/hmicfrs/wp-content/uploads/2014/04/improving-the-police-response-to-domestic-abuse.pdf`. Accessed: 2022-05-17.

Iyengar, R. (2009). Does the certainty of arrest reduce domestic violence? Evidence from mandatory and recommended arrest laws. *Journal of public Economics*, 93(1-2):85–98.

Katz, L. F., Kling, J. R., and Liebman, J. B. (2001). Moving to opportunity in boston: Early results of a randomized mobility experiment. *Quarterly Journal of Economics*, 116(2):607–654.

Kling, J. R. (2006). Incarceration length, employment, and earnings. *American Economic Review*, 96(3):863–876.

Maestas, N., Mullen, K. J., and Strand, A. (2013). Does disability insurance receipt discourage work? Using examiner assignment to estimate causal effects of ssdi receipt. *American Economic Review*, 103(5):1797–1829.

Marbach, M. and Hangartner, D. (2020). Profiling compliers and noncompliers for instrumental-variable analysis. *Political Analysis*, 28(3):435–444.

Miller, A. R. and Segal, C. (2019). Do female officers improve law enforcement quality? Effects on crime reporting and domestic violence. *The Review of Economic Studies*, 86(5):2220–2247.

National Research Council (1998). *Violence in families: Assessing prevention and treatment programs*. National Academies Press.

Office for National Statistics (2013). Homicide. `https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/compendium/focusonviolentcrimeandsexualoffences/yearendingmarch2016/homicide#how-are-victims-and-suspects-related`. Accessed: 2021-09-29.

Office for National Statistics (2020). England and Wales 2011 Census. `https://www.ethnicity-facts-figures.service.gov.uk/uk-population-by-ethnicity/national-and-regional-populations/regional-ethnic-diversity/latest#ethnic-groups-by-area`. Accessed: 2022-06-24.

Pate, A. M. and Hamilton, E. E. (1992). Formal and informal deterrents to domestic violence: The Dade County spouse assault experiment. *American Sociological Review*, 57(5):691–697.

Pinotti, P. (2020). The credibility revolution in the empirical analysis of crime. *Italian Economic Journal*, 6(2):207–220.

Schmidt, J. D. and Sherman, L. W. (1993). Does arrest deter domestic violence? *American Behavioral Scientist*, 36(5):601–609.

Sherman, L. W. and Berk, R. A. (1984). The specific deterrent effects of arrest for domestic assault. *American Sociological Review*, pages 261–272.

Sherman, L. W., Schmidt, J. D., Rogan, D. P., and Smith, D. A. (1992). The variable effects of arrest on criminal careers: The Milwaukee domestic violence experiment. *Journal of Criminal Law and Criminology*, 83(1):137–169.

Tjaden, P. G. and Thoennes, N. (2000). *Full report of the prevalence, incidence, and consequences of violence against women: Findings from the National Violence Against Women Survey*. US Department of Justice, Office of Justice Programs, National Institute of Justice.

Tur-Prats, A. (2019). Family types and intimate partner violence: A historical perspective. *Review of Economics and Statistics*, 101(5):878–891.

Tur-Prats, A. (2021). Unemployment and intimate partner violence: A cultural approach. *Journal of Economic Behavior & Organization*, 185:27–49.

Walby, S. and Allen, J. (2004). *Domestic violence, sexual assault and stalking: Findings from the British Crime Survey*. Home Office.

WHO (2002). World report on violence and health.

WHO (2021). Violence against women prevalence estimates, 2018: Global, regional and national prevalence estimates for intimate partner violence against women and global and regional prevalence estimates for non-partner sexual violence against women.

Tables and Figures

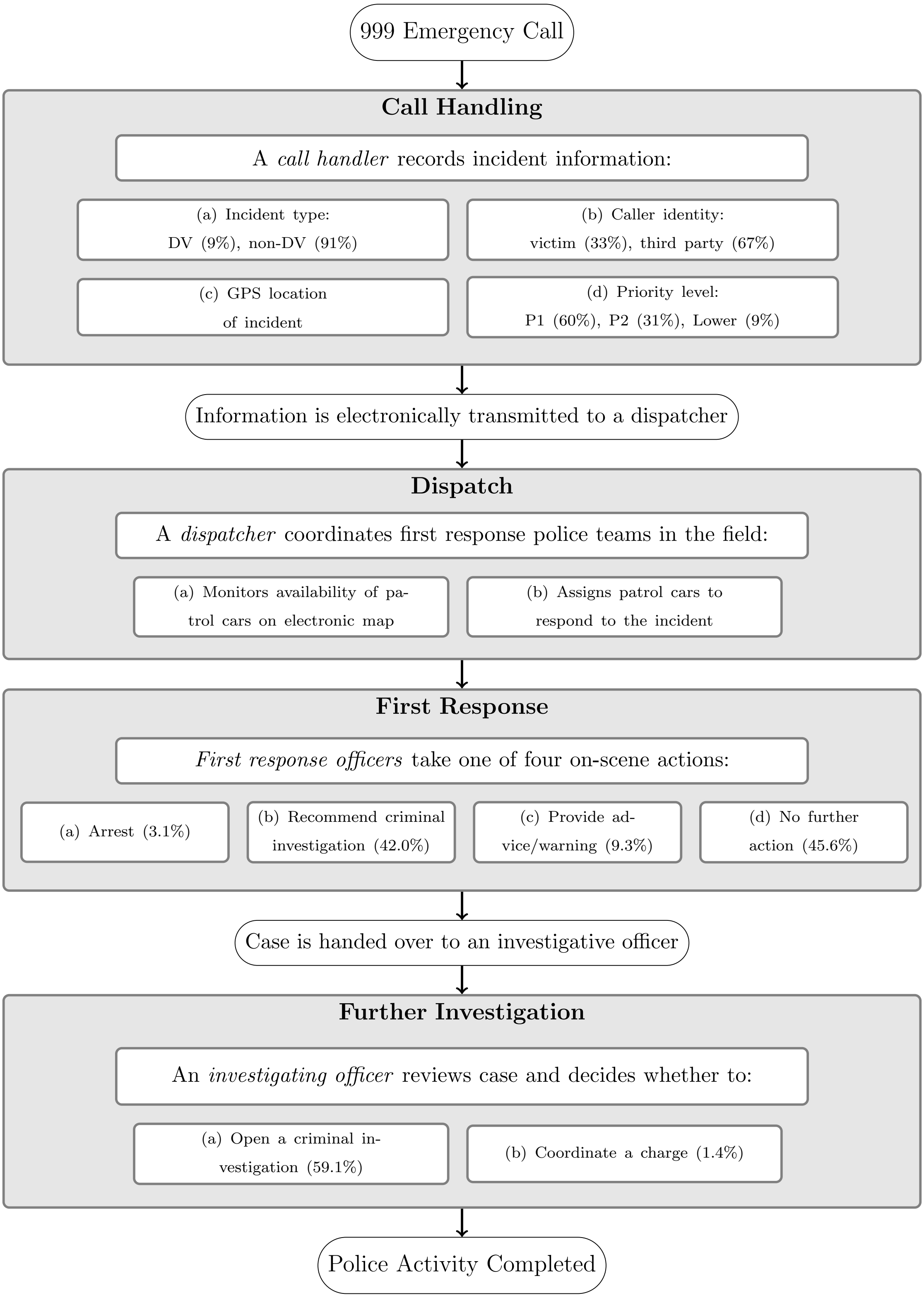
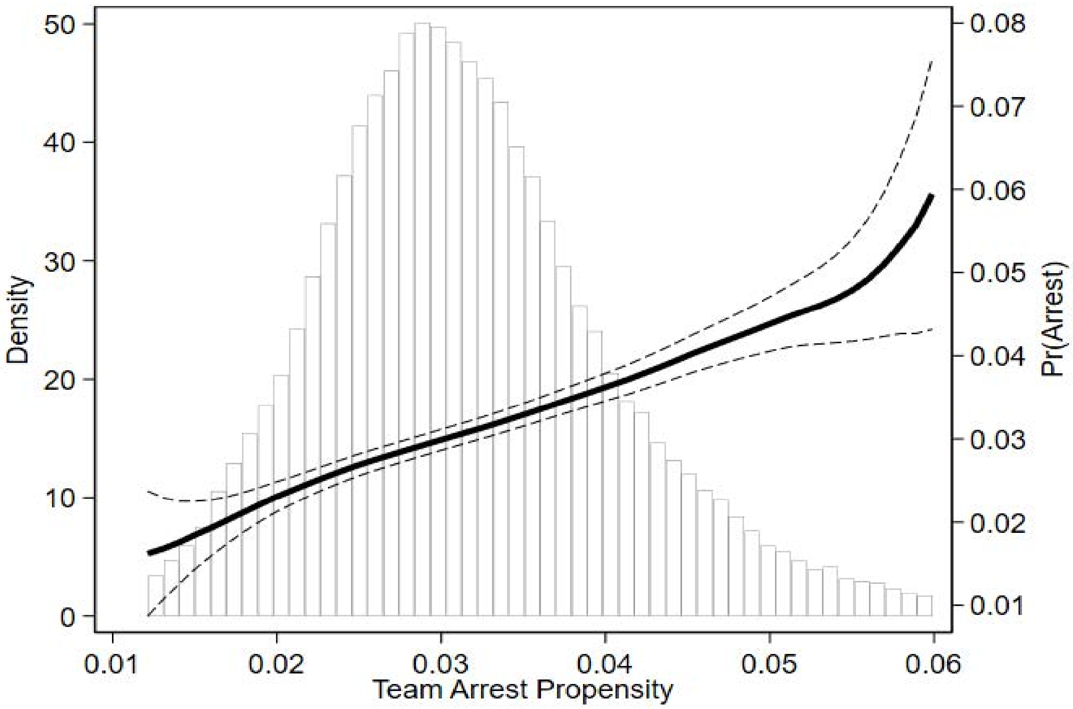

999 Emergency Call
Call Handling
A call handler records incident information:
(a) Incident type: DV (9%), non-DV (91%)
(b) Caller identity: victim (33%), third party (67%)
(c) GPS location of incident
(d) Priority level: P1 (60%), P2 (31%), Lower (9%)
Information is electronically transmitted to a dispatcher
Dispatch
A dispatcher coordinates first response police teams in the field:
(a) Monitors availability of patrol cars on electronic map
(b) Assigns patrol cars to respond to the incident
First Response
First response officers take one of four on-scene actions:
(a) Arrest (3.1%)
(b) Recommend criminal investigation (42.0%)
(c) Provide advice/warning (9.3%)
(d) No further action (45.6%)
Case is handed over to an investigative officer
Further Investigation
An investigating officer reviews case and decides whether to:
(a) Open a criminal investigation (59.1%)
(b) Coordinate a charge (1.4%)
Police Activity Completed

Figure 1: Police Handling of DV Emergency Calls



Notes: The probability of an arrest is plotted on the right-hand axis against team arrest propensity on the x-axis. Plotted values are based on mean-standardized residuals from a regression of arrest on the baseline set of call grade, time, and geography controls used in Table 3 column (4). The solid line shows a local linear regression of arrest on team arrest propensity, with dashed lines indicating 95% confidence intervals. The histogram shows the density of team arrest propensities along the left-hand axis (top and bottom 1% excluded).

Figure 2: First Stage Graph of Arrest on Team Arrest Propensity

Table 1: Testing Random Assignment of First Response Teams

	Dependent variable:	
	Arrest x 100	**Team arrest propensity x 100**
	(1)	(2)
Past DV history:		
Case in past 12 months	0.484**	-0.015
	(0.172)	(0.011)
Arrest in past 12 months	1.777***	0.016
	(0.325)	(0.014)
Formal investigation in past 12 months	0.163	0.015
	(0.184)	(0.011)
Criminal charge in past 12 months	2.043***	-0.014
	(0.305)	(0.013)
Case characteristics:		
Caller identity (=1 victim)	0.161	0.013
	(0.110)	(0.008)
Gender of call handler (=1 female)	-0.126	0.006
	(0.126)	(0.007)
Call handler experience (years)	0.008	0.000
	(0.008)	(0.001)
Mean of dep. var.	3.120	3.166
Joint F-statistic	22.936	0.946
[p-value]	[0.000]	[0.478]
Observations	124,216	124,216

Notes: OLS regressions controlling for the baseline set of call grade, time, and geography variables used in Table 3 column (4). Standard errors are reported in parentheses and are clustered at the level of the dispatched officer on a team with the most domestic violence cases.
*p<.10, **p<.05, ***p<.01

Table 2: Testing the Monotonicity Assumption

	Dependent variable: **Arrest**					
	Prior DV call		DV hotspot		Time of day	
	Yes	No	Yes	No	Day	Night
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Baseline instrument						
Team arrest propensity	0.656***	0.799***	0.799***	0.656***	0.543***	0.853***
	(0.103)	(0.083)	(0.090)	(0.085)	(0.097)	(0.083)
Mean of dep. var.	0.037	0.026	0.030	0.032	0.024	0.037
Observations	58,139	66,049	52,922	71,294	53,164	71,036
Panel B: Reverse sample instrument						
Reverse team arrest propensity	0.570***	0.674***	0.682***	0.441***	0.424***	0.319***
	(0.105)	(0.082)	(0.088)	(0.083)	(0.083)	(0.078)
Mean of dep. var.	0.038	0.027	0.031	0.032	0.024	0.037
Observations	52,865	59,680	47,554	65,640	48,907	65,697

Notes: First stage estimates regressing arrest on team arrest propensity/reverse team arrest propensity for different subsamples, controlling for the baseline set of call grade, time, and geography variables used in Table 3 column (4). Prior DV call refers to whether an emergency DV call was made in the previous 12 months. DV hotspot is defined as wards where the fraction of DV calls relative to the population is above the 75th percentile. For time of day, day is defined as 6 am to 6 pm and night as 6 pm to 6 am. Panel A uses the instrument constructed using the entire baseline sample. Panel B uses reverse sample instruments constructed using cases in the opposite subsample. For example, the reverse sample instrument in column (1) is based on cases with no prior DV call. Note that sample sizes are smaller in panel B, as we require at least 400 cases per team in each subsample. Standard errors are reported in parentheses and are clustered at the level of the dispatched officer on a team with the most domestic violence cases.
*p<.10, **p<.05, ***p<.01

Table 3: The Effect of Arrest on Repeat Emergency Calls for Domestic Violence

	Dependent variable: **Repeat call for DV**			
	OLS	IV		
	(1)	(2)	(3)	(4)
Arrest	0.001	-0.517***	-0.488***	-0.488***
	(0.008)	(0.170)	(0.187)	(0.187)
Call grade, time, ward F.E.'s	yes	yes	yes	yes
Ward x time F.E.'s	yes	no	yes	yes
Ward x call grade F.E.'s	yes	no	no	yes
Mean of dep. var.	0.492			
Control complier mean	0.962			
First stage		0.772***	0.723***	0.722***
		(0.068)	(0.070)	(0.070)
Reduced Form		-0.400***	-0.353***	-0.352***
		(0.131)	(0.136)	(0.136)
Kleibergen-Paap Wald F statistic		128	108	107
Observations	124,216	124,216	124,216	124,216

Notes: Call grade fixed effects are for call grade 1 versus 2. Time fixed effects include year, calendar month, day of week, time of day (6 hour intervals), and bank holidays. Ward fixed effects correspond to 146 geographical regions. The baseline specification in column (4) adds in complete interactions of the ward and time fixed effects as well as interactions between ward and call grade fixed effects. The first stage is a regression of arrest on the team arrest propensity instrument and the reduced form is a regression of repeat call for DV on the team arrest propensity instrument, controlling for the relevant fixed effects. Standard errors are reported in parentheses and are clustered at the level of the dispatched officer on a team with the most domestic violence cases.
*p<.10, **p<.05, ***p<.01

Table 4: Testing for a Reduction in Incidence versus Reporting

	Dependent variable:				
	Repeat call for DV	**Low severity repeat call**	**High severity repeat call**	**Victim-initiated repeat call**	**Third party-initiated repeat call**
	(1)	(2)	(3)	(4)	(5)
Arrest	-0.488***	0.064	-0.552***	-0.099	-0.390**
	(0.187)	(0.150)	(0.173)	(0.157)	(0.190)
Mean of dep. var.	0.492	0.192	0.300	0.182	0.311
Control complier mean	0.962	0.132	0.830	0.275	0.687
Observations	124,216	124,216	124,216	124,216	124,216

Notes: Column (1) repeats the baseline specification from Table 3 column (4). Columns (2) and (3) decompose the baseline estimate by the severity of repeat calls. Low severity repeat calls are defined as those with a call priority grade of 2 or lower. High severity repeat calls are those with a call priority grade of 1. Columns (4) and (5) decompose the baseline estimate by the identity of the caller. All specifications include the baseline set of call grade, time, and geography variables used in Table 3 column (4). Standard errors are reported in parentheses and are clustered at the level of the dispatched officer on a team with the most domestic violence cases.

*p<.10, **p<.05, ***p<.01

Table 5: Mechanisms

	Dependent variable: **Repeat call for DV in the specified time frame**						Dependent variable: **Criminal Charge**
	within 96 hours	in hours 1-48	in hours 49-96	within 12 months (excl. hours 1-96)	in months 1-6 (excl. hours 1-96)	in months 6-12	filed by a prosecutor
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Arrest	-0.198*	-0.197**	-0.025	-0.450**	-0.323*	-0.244	0.104**
	(0.104)	(0.099)	(0.055)	(0.184)	(0.185)	(0.178)	(0.053)
Mean of dep. var.	0.069	0.054	0.018	0.469	0.364	0.273	0.014
Control complier mean	0.251	0.233	0.043	0.909	0.684	0.507	0.020
Observations	124,216	124,216	124,216	124,216	124,216	124,216	124,216

Notes: All specifications include the baseline set of call grade, time, and geography variables used in Table 3 column (4). Standard errors are reported in parentheses and are clustered at the level of the dispatched officer on a team with the most domestic violence cases.
*p<.10, **p<.05, ***p<.01

Table 6: Testing the Exclusion Restriction

	Dependent variable: **Repeat call for DV**					
	(1)	(2)	(3)	(4)	(5)	(6)
Arrest	-0.488***	-0.487***	-0.487***	-0.486***	-0.480***	-0.518***
	(0.187)	(0.186)	(0.183)	(0.187)	(0.182)	(0.191)
Formal investigation		-0.049			-0.040	
		(0.034)			(0.035)	
Time on scene			0.002		0.009	
			(0.022)		(0.022)	
Recommend investigation				-0.044	-0.031	
				(0.061)	(0.061)	
Advice				-0.001	-0.000	
				(0.053)	(0.053)	
Instrument: Team arrest propensity	yes	yes	yes	yes	yes	yes
Instrument: FI propensity	no	yes	no	no	yes	no
Instrument: Time on scene propensity	no	no	yes	no	yes	no
Instrument: Recommend FI propensity	no	no	no	yes	yes	no
Instrument: Advice propensity	no	no	no	yes	yes	no
Control: Team characteristics	no	no	no	no	no	yes
Mean of dep. var.	0.492					
Control complier mean	0.962					
Observations	124,216	124,216	124,216	124,216	124,216	124,216

Notes: The first columns add in additional endogenous variables and instruments. The corresponding first stages are found in Appendix Table A5. Formal investigation is a dummy for whether a case is formally investigated. Time on the scene is measured in hundreds of minutes, and is based on the difference between when the first officer arrives and the last officer leaves. Recommend investigation and advice are dummy variables for whether the response team's actions include recommending a criminal investigation and providing advice, respectively. Instruments for the variables in column (2) to (4) are constructed analogously to how we construct the team arrest propensity instrument for arrests. The team characteristics in column (6) include the fraction of females on a team (mean=0.20) and the average age of the team (mean=36.9). All specifications include the baseline set of call grade, time, and geography variables used in Table 3 column (4). Standard errors are reported in parentheses and are clustered at the level of the dispatched officer on a team with the most domestic violence cases.

*p<.10, **p<.05, ***p<.01

Online Appendix

”Deterrence or Backlash? Arrests and the Dynamics of Domestic Violence”

by Sofia Amaral, Gordon B. Dahl, Victoria Endl-Geyer, Timo Hener, and Helmut Rainer

Online Appendix A: Additional Tables

Table A1: Sample sizes

Estimation sample	
Domestic violence cases classified by call handlers (2011-2016)	184,468
Nonmissing dispatch time	174,130
At least 400 DV cases in dispatched team	136,649
Call grade is 1 or 2 (baseline estimation sample)	124,216
Instrument construction sample	
Officer-case level observations in call handler defined DV cases (2010-2019)	631,834

Table A2: Characterization of Compliers

	Case in past 12 months	Arrest in past 12 months	Formal investigation in past 12 months	Criminal charge in past 12 months
	(1)	(2)	(3)	(4)
Population mean	0.468	0.062	0.376	0.058
Complier mean	0.602	0.045	0.404	0.142
Bootstrap std. err.	[0.117]	[0.062]	[0.114]	[0.052]
Observations	124,216	124,216	124,216	124,216

Notes: For details on these calculations see Appendix B.

Table A3: Testing for Bias due to Misclassification Errors

	Dependent Variable: **Repeat call for DV**	
	Only areas with min. 80% of HH's in detached houses	Excluding city center of Birmingham (3km radius)
	(1)	(2)
Arrest	-0.468*	-0.441**
	(0.263)	(0.194)
Mean of dep. var.	0.479	0.491
Control complier mean	0.931	0.914
Observations	81,953	118,559

Notes: In column (1), the sample is restricted to areas (wards) where at least 80% of the households live in detached houses. Data on the dwelling types by ward in West Midlands come from NOMIS labour market statistics. In column (2), the city center of Birmingham (defined as the 3km radius around St. Philips Cathedral) is excluded. All specifications include the baseline set of call grade, time, and geography variables used in Table 3 column (4). Standard errors are reported in parentheses and are clustered at the level of the dispatched officer on a team with the most domestic violence cases.
*p<.10, **p<.05, ***p<.01

Table A4: Testing for Differential Accuracy of Geo-Coded Location as a Function of Arrest

	Based on official victim ID		
	No repeat identified	Repeat identified	Total
Based on geo-coordinates			
Full Estimation Sample			
No repeat identified %	90.93	9.07	100.00
Arrest Subsample			
No repeat identified %	91.61	8.39	100.00
No Arrest Subsample			
No repeat identified %	90.90	9.10	100.00
p-value for mean difference	0.37		

Notes: In the paper, we match victims over time based on the geo-coordinates of the incident location. For the subsample of the data where the investigative officer opens an investigation, we can use official victim IDs to track individuals over time. This table constructs an alternative measure for a repeat DV call using official victim ID and compares it to our measure using geo-coordinates for the same subsample. While the table reveals that we miss 9% of repeat cases in this subsample, whether a repeat case is missing is not significantly related to whether there was an arrest.

Table A5: First Stages for the Extended IV Models

	Arrest	**Formal investigation**	**Arrest**	**Time on scene**	**Arrest**	**Recommend investigation**	**Advice**
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Team arrest propensity	0.728*** (0.070)	0.359** (0.142)	0.737*** (0.069)	-0.049 (0.147)	0.728*** (0.072)	1.237*** (0.134)	1.174*** (0.134)
FI propensity	0.018* (0.010)	0.942*** (0.031)					
Time on scene propensity			0.025*** (0.008)	1.048*** (0.025)			
Recommend FI propensity					0.004 (0.010)	1.086*** (0.031)	1.124*** (0.029)
Advice propensity					-0.005 (0.017)	0.286*** (0.047)	1.280*** (0.050)
Kleibergen-Paap Wald F statistic	52		58			48	
Observations	124,216	124,216	124,216	124,216	124,216	124,216	124,216

	Arrest	**Formal investigation**	**Recommend investigation**	**Advice**	**Time on scene**
	(8)	(9)	(10)	(11)	(12)
Team arrest propensity	0.733*** (0.073)	0.294** (0.146)	-0.133 (0.154)	1.258*** (0.132)	1.184*** (0.134)
FI propensity	0.014 (0.011)	0.984*** (0.034)	-0.082** (0.034)	-0.116*** (0.032)	-0.041 (0.030)
Time on scene propensity	0.024*** (0.008)	-0.061** (0.024)	1.087*** (0.026)	-0.004 (0.024)	0.011 (0.023)
Recommend FI propensity	-0.006 (0.011)	-0.008 (0.036)	-0.009 (0.035)	1.134*** (0.033)	1.139*** (0.031)
Advice propensity	0.009 (0.017)	0.226*** (0.056)	0.251*** (0.056)	0.267*** (0.048)	1.279*** (0.051)
Kleibergen-Paap Wald F statistic			42		
Observations	124,216	124,216	124,216	124,216	124,216

Notes: This table displays first stages corresponding to Table 6. All specifications include the baseline set of call grade, time, and geography variables used in Table 3 column (4). Standard errors are reported in parentheses and are clustered at the level of the dispatched officer on a team with the most domestic violence cases.
*p<.10, **p<.05, ***p<.01

Table A6: Controlling for Propensities of Actions other than Arrest

	Dependent Variable: **Repeat call for DV**				
	(1)	(2)	(3)	(4)	(5)
Arrest	-0.488***	-0.511***	-0.487***	-0.561***	-0.551***
	(0.187)	(0.187)	(0.183)	(0.190)	(0.188)
Instrumented: Team arrest propensity	yes	yes	yes	yes	yes
Control: FI propensity	no	yes	no	no	yes
Control: Time on scene propensity	no	no	yes	no	yes
Control: Recommend FI propensity	no	no	no	yes	yes
Control: Advice propensity	no	no	no	yes	yes
Mean of dep. var.			0.492		
Control complier mean			0.962		
Observations	124,216	124,216	124,216	124,216	124,216

Notes: Formal investigation (FI) propensity is the propensity of a team's cases to be formally investigated. Time on scene propensity is a team's average time spend on the scene in other cases. Recommend investigation propensity is the team's average propensity in other DV cases to recommend a criminal investigation, and advice propensity is similarly defined. All specifications include the baseline set of call grade, time, and geography variables used in Table 3 column (4). Standard errors are reported in parentheses and are clustered at the level of the dispatched officer on a team with the most domestic violence cases.
*p<.10, **p<.05, ***p<.01

Table A7: Testing Robustness to Alternative Specifications

	Dependent Variable: **Repeat call for DV**				
	Intensive margin	IV: call grade 1 and 2	IV: Traditional l-o-o	Team DV cases: 300	Team DV cases: 500
	(1)	(2)	(3)	(4)	(5)
Arrest	-1.516*	-0.514***	-0.474**	-0.406**	-0.441**
	(0.813)	(0.198)	(0.187)	(0.184)	(0.192)
Mean of dep. var.	1.337	0.493	0.492	0.491	0.494
Control complier mean	2.867	0.921	0.948	0.880	0.915
Observations	123,063	118,788	124,229	134,534	111,437

Notes: Column (1) uses the intensive margin of the number of repeat DV cases in the following 12 months as the outcome variable. For this column, we trim the data to exclude the top 1% of the outcome variable. The remaining columns all use the baseline outcome. Our main specification uses high priority calls (call grade 1 or 2) to define the sample, but uses calls of any grade to construct the team arrest propensity instrument; in column (2) we construct the instrument using only high priority calls. Our main specification only uses DV cases from other geo-locations (whether the current case or any other case) to construct the instrument; in column (3) we use a standard leave-one-out instrument that only excludes the current case. Our main specification requires the first response team to handle at least 400 DV cases; in columns (4) and (5) we require the teams to handle at least 300 or at least 500 DV cases, respectively. All specifications include the baseline set of call grade, time, and geography variables used in Table 3 column (4). Standard errors are reported in parentheses and are clustered at the level of the dispatched officer on a team with the most domestic violence cases.
*p<.10, **p<.05, ***p<.01

Table A8: Heterogeneous Effects of Arrest

	Dependent Variable: **Repeat call for DV**			
	Formally investigated DV case in past 12 months	Pr(Arrest|X) is high	At least one female officer on response team	Response team with above mean age
	(1)	(2)	(3)	(4)
Yes	-0.744**	-0.626**	-0.552**	-0.322
	(0.308)	(0.307)	(0.242)	(0.230)
No	-0.431**	-0.513***	-0.474**	-0.677***
	(0.185)	(0.190)	(0.222)	(0.230)
Observations	124,216	124,216	124,216	124,211

Notes: This table shows heterogeneity analyses based on IV regressions which interact both the arrest variable and the team arrest propensity instrument with identifiers for the specified groups. In column (2), we differentiate between cases with arrest predictions above or below the 75th percentile, where the predictions are based on the variables appearing in Table 1. All specifications include the baseline set of call grade, time, and geography variables used in Table 3 column (4). Standard errors are reported in parentheses and are clustered at the level of the dispatched officer on a team with the most domestic violence cases.
*p<.10, **p<.05, ***p<.01

Online Appendix B: Estimating Control Complier Means and Complier Characteristics

Estimating the Shares of Compliance Types in the Sample

Our estimation of the fraction of compliers (π_c), always-takers (π_a) and never-takers (π_n) follows the methodology outlined in Dahl et al. (2014). We refer the reader to the Technical Appendix in Dahl et al. (2014) and repeat only the central estimation steps here. Let $\underline{z}$ and $\overline{z}$ denote the minimum (least-likely arresting officers) and maximum (most-likely arresting officers) values of the instrument, assumed to be the values for the bottom 1 percentile and top 1 percentile of arrest propensity. Let $\hat{\alpha}_0$ and $\hat{\alpha}_1$ denote the estimated coefficients from our first stage regression in equation (2). We calculate π_c from $\hat{\alpha}_1(\overline{z} - \underline{z})$, π_a from $\hat{\alpha}_0 + \hat{\alpha}_1\underline{z}$, and π_n from $1 - \hat{\alpha}_0 - \hat{\alpha}_1\overline{z}$. The estimated shares of the three compliance types in our sample are $\pi_c = 0.040$, $\pi_a = 0.017$, and $\pi_n = 0.943$.

Estimating Control Complier Means

The control complier mean (CCM) informs us about how likely repeat victimization would be if a call had not resulted in an arrest. We therefore need to estimate the sample analog of $E(DV_{i,t+1}(0)|A_{i,t}(\overline{z}) > A_{i,t}(\underline{z}))$. To do so, consider DV calls that do not result in an arrest ($A_{i,t} = 0$). Victims whose calls are handled by response teams with $Z_{j(i)} = \overline{z}$ are never-takers, and their conditional expectation of a repeat call is:

$$E(DV_{i,t+1}|A_{i,t} = 0, Z_{j(i)} = \overline{z}) = E(DV_{i,t+1}|A_{i,t}(\overline{z}) = A_{i,t}(\underline{z}) = 0) \tag{3}$$

Those victims whose calls are handled by response teams with $Z_{j(i)} = \underline{z}$ are a mixture of never-takers and compliers:

$$\begin{aligned} E(DV_{i,t+1}|A_{i,t} = 0, Z_{j(i)} = \underline{z}) = &\frac{\pi_n}{\pi_n + \pi_c} E(DV_{i,t+1}|A_{i,t}(\overline{z}) = A_{i,t}(\underline{z}) = 0) \\ &+ \frac{\pi_c}{\pi_n + \pi_c} E(DV_{i,t+1}(0)|A_{i,t}(\overline{z}) > A_{i,t}(\underline{z})). \end{aligned} \tag{4}$$

By combining these two equations, we obtain:

$$\begin{aligned} E(DV_{i,t+1}(0)|A_{i,t}(\overline{z}) > A_{i,t}(\underline{z})) = &\frac{\pi_n + \pi_c}{\pi_c} E(DV_{i,t+1}|A_{i,t} = 0, Z_{j(i)} = \underline{z}) \\ &- \frac{\pi_n}{\pi_c} E(DV_{i,t+1}|A_{i,t} = 0, Z_{j(i)} = \overline{z}). \end{aligned} \tag{5}$$

Above, we have already laid out how to estimate π_c and π_n. To obtain the two remaining quantities in equation (5), we need an estimate of the relationship between $DV_{i,t+1}$ and $Z_{j(i)}$ conditional on $A_{i,t} = 0$. We use a linear probability model to estimate

$$DV_{i,t+1} = \gamma_0 + \gamma_1 Z_{j(i)} + \gamma_2 A_{i,t} + X'_{i,t} + \epsilon_{i,t}. \tag{6}$$

We then calculate $E(DV_{i,t+1}|A_{i,t} = 0, Z_{j(i)} = \underline{z})$ from $\hat{\gamma}_0 + \hat{\gamma}_1\underline{z}$ and $E(DV_{i,t+1}|A_{i,t} = 0, Z_{j(i)} = \overline{z})$ from $\hat{\gamma}_0 + \hat{\gamma}_1\overline{z}$.

An alternative way of obtaining an estimate for the control complier mean is to calculate the difference between the treated complier mean (TCM) and our IV estimate for the arrest effect (Katz et al., 2001). For this, we need to estimate the sample analog of $E(DV_{i,t+1}(1)|A_{i,t}(\overline{z}) > A_{i,t}(\underline{z}))$, which by similar arguments as used above is given by:

$$\begin{aligned} E(DV_{i,t+1}(1)|A_{i,t}(\overline{z}) > A_{i,t}(\underline{z})) = \frac{\pi_a + \pi_c}{\pi_c} E(DV_{i,t+1}|A_{i,t} = 1, Z_{j(i)} = \overline{z}) \\ - \frac{\pi_a}{\pi_c} E(DV_{i,t+1}|A_{i,t} = 1, Z_{j(i)} = \underline{z}), \end{aligned} \tag{7}$$

where π_a and π_c are the shares of always-takers and compliers, respectively. The control complier mean is then

$$E(DV_{i,t+1}(0)|A_{i,t}(\overline{z}) > A_{i,t}(\underline{z})) = E(DV_{i,t+1}(1)|A_{i,t}(\overline{z}) > A_{i,t}(\underline{z})) - \hat{\beta}_1, \tag{8}$$

where $\hat{\beta}_1$ is our IV estimate for the arrest effect. Estimating $E(DV_{i,t+1}|A_{i,t} = 1, Z_{j(i)} = \overline{z})$ and $E(DV_{i,t+1}|A_{i,t} = 1, Z_{j(i)} = \underline{z})$ in equation (7) with a linear probability model and substituting the estimatess into equation (8), we obtain control complier means that are virtually identical to the ones reported in the paper. Our conclusion of a high value of $E(DV_{i,t+1}(0)|A_{i,t}(\overline{z}) > A_{i,t}(\underline{z}))$ is therefore robust to the choice of model.

Estimating Characteristics of Compliers

Our characterization of compliers adapts the binary-instrument methodology proposed by Marbach and Hangartner (2020) to a setting with a continuous instrument. Let $X_{i,t}$ be a covariate. By arguments we have established above, the covariate means for never-takers and always-takers can be estimated and are respectively given by:

$$E(X_{i,t}|A_{i,t} = 0, Z_{j(i)} = \overline{z}) = E(X_{i,t}|A_{i,t}(\overline{z}) = A_{i,t}(\underline{z}) = 0) \tag{9}$$

and

$$E(X_{i,t}|A_{i,t} = 1, Z_{j(i)} = \underline{z}) = E(X_{i,t}|A_{i,t}(\overline{z}) = A_{i,t}(\underline{z}) = 1). \tag{10}$$

We calculate $E(X_{i,t}|A_{i,t} = 0, Z_{j(i)} = \overline{z})$ from $\hat{\zeta}_0 + \hat{\zeta}_1\overline{z}$, where $\hat{\zeta}_0$ and $\hat{\zeta}_1$ are OLS estimates of the relationship between $X_{i,t}$ and $Z_{j(i)}$ conditional on $A_{i,t} = 0$. Similarly, we calculate $E(X_{i,t}|A_{i,t} = 1, Z_{j(i)} = \underline{z})$ from $\hat{\eta}_0 + \hat{\eta}_1\underline{z}$, where $\hat{\eta}_0$ and $\hat{\eta}_1$ are OLS estimates of the relationship between $X_{i,t}$ and $Z_{j(i)}$ conditional on $A_{i,t} = 1$.

Turning to compliers, we note that by the law of intereated expectations, the population mean of $X_{i,t}$ can be decomposed into the never-taker, always-taker, and complier means, weighted by the share of each group:

$$\begin{aligned} E(X_{i,t}) = \pi_n E(X_{i,t}|A_{i,t}(\overline{z}) = A_{i,t}(\underline{z}) = 0) + \pi_a E(X_{i,t}|A_{i,t}(\overline{z}) = A_{i,t}(\underline{z}) = 1) \\ + \pi_c E(X_{i,t}|A_{i,t}(\overline{z}) > A_{i,t}(\underline{z})) \end{aligned} \tag{11}$$

Combining equations (9) to (11), we calculate covariate means for compliers according to:

$$\begin{aligned} E(X_{i,t}|A_{i,t}(\overline{z}) > A_{i,t}(\underline{z})) = \frac{1}{\pi_c} E(X_{i,t}) - \frac{\pi_n}{\pi_c} E(X_{i,t}|A_{i,t} = 0, Z_{j(i)} = \overline{z}) \\ - \frac{\pi_a}{\pi_c} E(X_{i,t}|A_{i,t} = 1, Z_{j(i)} = \underline{z}). \end{aligned} \tag{12}$$