Referee Report and Replication — Microeconometrics II

Does Strengthening Self-Defense Law Deter Crime or Escalate Violence? Evidence from Expansions to Castle Doctrine

Authors: Cheng Cheng and Mark Hoekstra

Published in: Journal of Human Resources (2013)

Student: Vinícius de Almeida Nery Ferreira

Professor: Bruno Ferman

TA: Arthur Botinha

December 5, 2024

Contents

1 Summary and Parameter of Interest

Until the early 2000s, one foundational principle of US self-defense law was that one has the “duty to retreat” before using lethal force against a threatening individual. The exception is when the assault occurs in one household, as the home is one’s *castle*.

Starting in 2005, many US states started to pass laws that represented a somewhat opposite view: the *Castle Doctrine*, which expanded the places where lethal self-defense is justifiable – that is, one’s *castles* – and removes potential civil liabilities of lethal self-defense acts. From an economic point of view, these laws lower the expected cost of lethal self-defense and, as a consequence, increase the expected cost of violent crime from the perspective of the perpetrator.

Cheng and Hoekstra (2013)’s goal is to evaluate the effect of these policies on violent crime. Precisely, the parameter of interest is the average treatment effect of the *Castle Doctrine* on the treated states (ATT) – that is, on states that passed some law expanding self-defense rights.

Cheng and Hoekstra (2013) find that, contrary to economic theory, the law did not have any deterrence effect on violent crime such as burglary, robbery or assault. However, the *Castle Doctrine* is associated with an average increase of 8-10 p.p. in homicides in treated states, which is consistent with the view that the law reduced the expected cost of lethal violence. Furthermore, this increase cannot be accounted for murders that are justified by self-defense and holds using both traditional and permutation inference techniques.

By their calculations, this increase amounts to an additional 600 homicides per year in treated states, which highlights the unintended consequences of strengthening self-defense laws. This can be relevant not only for states and countries considering similar laws, but also for regions where these laws are already in effect and may warrant reform.

2 Empirical Setting

To accomplish these goals, Cheng and Hoekstra (2013) leverage the within-state variation in self-defense laws induced by the 21 states that passed such laws between 2005 and 2010. Thus, they use a difference-in-differences (DiD) framework with staggered treatment adoption.

Intuitively, the goal of the design is to measure whether violent crime changed more in states that adopted *Castle Doctrine* laws relative to those that did not. To answer this, they construct a panel dataset and estimate a Two Way Fixed Effects (TWFE) regression:

$$Y_{it} = \beta_0 CDL_{it} + \beta_1 X_{it} + c_i + u_t + \varepsilon_{it}, \tag{1}$$

where CDL_{it} indicates the proportion of the year t in which state i has a *Castle Doctrine* Law (CDL).[1] c_i and u_t are state and time fixed-effects, respectively. X_{it} is a vector with time-varying controls including region-by-year fixed effects, so comparisons are made only within a region.[2] Note that different states passed the set of laws at different times and no state revoked these laws, and so we have a staggered design with treatment being an "absorbing" condition.

Their outcomes can be divided into three categories. First, they conduct falsification tests using larcenies ("*furtos*") and motor vehicle theft, which they argue should not be affected by *Castle Doctrine* laws. Second, they investigate the deterrence effects of the law on violent crime such as burglary (home-breaking), robbery, and aggravated assault. Finally, they consider its effects on the homicide rates.

Their identifying assumptions is that parallel trends holds conditional on their covariates X_{it}. In their words:

> The identifying assumption is that in the absence of the Castle Doctrine laws, adopting states would have experienced changes in crime similar to nonadopting states in the same region of the country. (Cheng and Hoekstra 2013, p. 831)

Equation 1 is estimated with ordinary and weighted least squares (OLS and WLS, respectively) using the state's population as analytical weights. Standard-errors are clustered at the state-level, allowing for serial correlation. They also conduct inference with permutation tests.

[1] Note that, for the first year of the law, $CDL_{it} \in (0, 1)$. Thus, in principle, treatment is a continuous variable. However, as reported by the authors and showed in the replication, effects and significances are similar when considering CDL_{it} as a binary variable indicating if the law was valid during all of the year t in state i.

[2] The four regions are West, Midwest, Northeast and South. The low number of regions makes clustering standard errors at this level troublesome.

3 Referee Report

3.1 Major Points

3.1.1 TWFE and Staggered Treatment Adoption

First, some notation will be useful. Since no state revoked *Castle Doctrine* laws, treatment CDL_{it} is an absorbing state, and so we can partition units depending on *when* they first take treatment. Let $G_i = \min\{t : t \in \mathbb{N} \text{ and } CDL_{it} = 1\}$ denote such a variable, with the convention that $G_i = 0$ indicates that unit i was never treated. Potential outcomes can be written as

$$\begin{cases} Y_{it}(0) = c_i + u_t + \varepsilon_{it} \\ Y_{it}(G_i) = \tau_{it}(G_i) + Y_{it}(0), \end{cases}$$

with $\tau_{it}(G_i)$ being the treatment effect at time t for unit i first treated at G_i. In principle, $\tau_{it}(G_i)$ is a random variable across i, t and G_i.

By using a static TWFE model, the authors' identification strategy implicitly assumes that treatment effects are homogeneous across cohorts and periods, that is, $\tau_{it}(G_i) = \tau$ for all i and $t \geq G_i$, where $\tau \in \mathbb{R}$. This is very strong: all units have the same treatment effect and this effect is constant regardless of the length of exposure to treatment (Roth et al. 2023, p. 2224).

The authors do not present evidence in favor of this assumption, which, as argued by Roth et al. (2023), is not suitable for many economic contexts. If treatment effects are heterogeneous, then the TWFE estimand β_0 in Equation 1 is potentially a non-convex weighted average of treatment effects (De Chaisemartin and d'Haultfoeuille 2020; Goodman-Bacon 2021), which arise from "forbidden comparisons" between already-treated units. Crucially, weights can be negative, meaning that the TWFE estimand wouldn't satisfy the minimum requirement to be causally interpretable.[3]

Even if weights are positive, they are not intuitive and reflect statistical properties of the TWFE estimand rather than a meaningful economic quantity. The authors should either provide compelling evidence in favor of the implicit homogeneous treatment effect assumption or abandon TWFE and use a new estimator for staggered adoption (Roth et al. 2023, Section 3).

3.1.2 Evidence of Validity of Identifying Assumption and Dynamic Effects

Even ignoring the problems with TWFE, the authors should present more evidence toward the validity of the parallel trends assumption (PT). Although placebo tests using crimes not likely to be affected by *Castle Doctrine* are appreciated, the lack of dynamic specifications and pre-trends tests are concerning.

In that respect, the authors seem to have some confusion regarding the untestability of PT. They claim that, by including an indicator in Equation 1 for the two years prior to the laws, they are able to conduct a "formal statistical test" of PT (p. 831). This is not true, as it is a

[3] Mogstad and Torgovitsky (2024) call this minimum desirable property of an estimand being a convex weighted average of treatment effects as "weak causality".

test of pre-trends. Furthermore, the choice of a two-year indicator variable, instead of two or more separate indicators, is not justified: is it for statistical power?

The only indicative of fully testing pre-trends and of using a dynamic specification is given in Figure 2 (p. 838). However, this too is puzzling: the authors pool pairs of years together and only consider violations of (-4 and -3) and (-2 and -1) years prior to treatment, whereas the eleven-year panel data allows for much more.

Most concerning is the lack of confidence intervals: in all cases, pre-trends coefficients are positive, but neither the authors nor the reader can assess their statistical significance. Simply stating that pre-trends coefficients are small relative to post-treatment effects is not enough, as they can be biased precisely due to the violation of PT. Furthermore, these figures are only shown for homicides: what about the other outcomes analyzed?

Finally, the authors try to provide further evidence in favor of PT in Figure 1 (p. 837), which shows, for each G_i, the average of states that passed the law in that year versus the average of never treated states. This is appreciated, but there are some problems: the violation of pre-trends is pretty clear in the 2007, 2008 and 2009 cohorts and, without confidence intervals, we can't say much about 2005 and 2006. Finally, TWFE does not make the comparison between treated and never-treated units, but rather between units whose treatment status did not change in that period (being them treated or not).

In this sense, the authors should include dynamic specifications, plotting event studies with confidence intervals for all outcomes. It is preferable that these intervals be simultaneous, rather than pointwise (Freyaldenhoven et al. 2021). Furthermore, if they opt to continue with TWFE, they should be cautious with "cross-lag contamination" and negative weights when there is heterogeneity in treatment effects across cohorts G_i (Sun and Abraham 2021), or, preferably, use a different estimator that allows for arbitrary heterogeneity.

In addition to allowing pre-trends tests, dynamic specifications answer policy-relevant questions, such as if the effects are increasing over time or peak after the laws were passed. In case of a pre-trends violation, they can use the sensitivity analysis of Rambachan and Roth (2023) to bound the PT violation by pre-trends deviations.

3.2 Minor Points

3.2.1 Mechanisms

The main result of the paper is that *Castle Doctrine* laws "caused" an increase in homicides, while having no deterrence effects on other types of crime. However, there is little discussion on *why* homicides increased other than the argument of economic theory, that is, that the expected cost of lethal violence decreased.

The only phrase in this respect is written in p. 825: "Collectively, these laws lower the cost of using lethal force to protect oneself, though they also lower the cost of escalating violence in other conflicts". The authors could make it clearer why this is the case.

3.2.2 Time-Varying and Bad Controls

The authors include controls in Equation 1 to improve the credibility of PT. However, covariates being time-varying allows for the possibility of bad controls. For example, under their hypothesis that *Castle Doctrine* increases lethal violence by decreasing its costs, it would too have an increase in police operation and in incarceration: two variables they control for.

To avoid such issues, they should restrict their attention to controls measured at baseline, which can be difficult to do with TWFE (Huntington-Klein 2023; Caetano et al. 2024). This is another reason to consider alternative estimators, such as Callaway and Sant'Anna (2021), which better account for pre-treatment covariates and for the staggered DiD design.

3.2.3 Control States

In footnote 11 (p. 827), it is implicit that states that passed *Castle Doctrine* laws before 2000 are included as controls, which is confirmed in our replication. Moreover, these states are coded as never treated by not passing the set of laws between 2000 and 2010.

Depending on the number of states treated before 2000 and on the persistence of the laws' effects, this invalidates any comparison, be it made with TWFE or other estimator. The authors should present which and how many states passed such laws before 2000 and, preferably, remove them from the control sample.

3.2.4 Inference

One of the strongest points of the paper in my view is their inference methods. Instead of relying solely on standard errors clustered at the state level,[4] they also use permutation tests in the spirit of Bertrand, Duflo, and Mullainathan (2004). They use an "out of sample" version to compute placebo distributions, randomly selecting 11-year panels from 1960 to 2004 and randomly assigning treatment dates found it the actual 2000-2010 data. This procedure is done without replacement, avoiding some implementation concerns raised in Ferman (2022, A.3., A.7.2.) in the context of bootstraps. They use the placebo estimates as the test statistic.

Additionally, they run simulations to see if homicide rates of treated states ever diverged like the deviations seen after *Castle Doctrine.* Starting from 1960 and until 2000, they construct 40 datasets with an iterative procedure and repeat treatment dates and states according to the structure of the original dataset.

Both of these design-based exercises reveal that the result of the original sample is "extreme". In my opinion, this is the most convincing piece of evidence they present in favor of their results (p. 842-844). However, as discussed in Section 3.1, we can't ignore the fact that their identifying assumption of PT may not hold in the original panel or that their estimator does not fit their staggered design.

[4] There are 21 treated clusters, so the numbers of cluster is probably not "too small" to warrant methods such as Conley and Taber (2011) and Ferman and Pinto (2019). Nonetheless, we run some simple simulations with iid standard normal outcomes in Section 5.3 to assess if cluster-robust standard errors are valid (Ferman 2022) – something that the authors should also consider.

4 Replication of Main Results

The paper does not have a replication package, but the data was made available by Cunningham (2021). In this short replication, we focus on the results using Cheng and Hoekstra (2013)'s preferred specification (Equation 1), which has state, year and region-year fixed effects, as well as a set of time-varying controls.[5]

We use the `feols` function from the `fixest` **R** package to estimate TWFE regressions. Results are shown in Table 1. All standard errors are clustered at the state level. In most cases, time-varying controls are not statistically significant at the 5% level.

Table 1: Main Results of Cheng and Hoekstra (2013) – Preferred Specification

	OLS						Weighted by State Population					
Dependent Variable	Homicide	Larceny	Motor Theft	Burglary	Robbery	Assault	Homicide	Larceny	Motor Theft	Burglary	Robbery	Assault
CDL	0.0600 (0.0693)	-0.0019 (0.0213)	0.0081 (0.0413)	0.0066 (0.0272)	0.0084 (0.0393)	0.0343 (0.0439)	0.0937*** (0.0294)	-0.0091 (0.0141)	-0.0252 (0.0401)	0.0223 (0.0226)	0.0262 (0.0232)	0.0372 (0.0324)
Observations	550	550	550	550	550	550	550	550	550	550	550	550

Notes: ***: 0.01, **: 0.05, *: 0.1. Standard errors clustered at the state level in parentheses.
All dependent variables are measured at the annual level and enter regressions in log.
All regressions include a set of time-varying controls and state, year and region by year fixed effects.

Point-estimates are the same as those reported by Cheng and Hoekstra (2013) (Tables 3, 4 and 5). Standard errors are slightly different, but only at the third or fourth decimal column. Note that the main result of the paper only occurs when using WLS.

As mentioned in the summary of the empirical strategy, CDL_{it} is a continuous variable in the first year of treatment. To implement the new methods of the staggered DiD literature, we transform it to a binary variable in two different ways.

We first follow Cheng and Hoekstra (2013) and Cunningham (2021) and consider an indicator variable if the state had *Castle Doctrine* laws in effect for the whole year, which differs from the previous treatment definition only on the first year. For robustness, we also consider the case when the state has laws in effect for over half of the year.

Comparative results of these treatment definitions are shown in Table 2, where we only show the WLS estimates for the homicide rate for brevity. Results remain largely the same, and so we continue with treatment starting when the state has laws in effect for the full year.

Table 2: Continuous versus Binary Treatments – WLS Estimates on log(Homicide)

Treatment	Continuous	Full Year	Majority of the Year
CDL	0.0937*** (0.0294)	0.0821*** (0.0272)	0.0808*** (0.0279)
Observations	550	550	550

Notes: ***: 0.01, **: 0.05, *: 0.1. Standard errors clustered at the state level in parentheses.

Regressions include a set of time-varying controls and state, year and region by year fixed effects.

[5] This includes the shares of white and black population, the log of government expenditure on subsidies and public welfare per capita, the log of number of policemen per capita, the unemployment and poverty rates, the log of median income and the log of the lagged and current number of incarcerations per capita.

5 Further Analysis and Extensions

In this analysis, we will focus on the major points of Section 3: accounting for heterogeneity treatment effects in the staggered DiD and providing evidence in favor of pre-trends (Section 5.1); and conducting sensitivity analysis (Section 5.2). Furthermore, we will do some simple simulations in Section 5.3 to partially assess the validity of cluster-robust standard errors in our setting. Section 6 provides some final remarks.

5.1 Accounting for Arbitrary Heterogeneity in Treatment Effects

5.1.1 Estimation Theory and Details

In order to account for treatment effect heterogeneity in the staggered DiD setting, we will use the estimator of Callaway and Sant'Anna (2021). Let $G_g = \mathbb{1}\{G_i = g\}$ be a binary variable indicating that unit i was first treated at time g, with $G_g = 0$ for never-treated units. Our parameters of interest will be the *group-time average treatment effects*, which are the average treatment effect for units of group g at period t:

$$ATT(g,t) = \mathbb{E}\left[Y_{it}(g) - Y_{it}(0) \mid G_g = 1\right]$$

To avoid the pitfalls of "forbidden comparisons", we will use as controls the states that did not pass *Castle Doctrine* laws in 2000-2010 ("never-treated").[6] Callaway and Sant'Anna (2021, Assumption 4, p. 204) make the following assumption:

(Conditional) Parallel Trends Based on a "Never-Treated" Group: For each group g and all $t \in \{2, ..., T\}$ such that $t \geq g$,

$$\mathbb{E}\left[Y_{it}(0) - Y_{i(t-1)}(0) \mid X, G_i = g\right] = \mathbb{E}\left[Y_{it}(0) - Y_{i(t-1)}(0) \mid X, G_i = 0\right]$$

Under this "generalized" parallel trends assumption, they propose a doubly robust estimand that identifies $ATT(g,t)$ without imposing any restriction on treatment effect heterogeneity. The estimator is constructed using sample analogs and inference is done using an asymptotically valid multiplier bootstrap which allows for construction of simultaneous confidence intervals.

Note that Callaway and Sant'Anna (2021) allow for parallel trends to hold conditional on covariates. These are incorporated on the propensity score of the doubly robust estimand using the values of the last pre-treatment period.

Additionally, instead of looking at all $ATT(g,t)$, we may be interested in more aggregate measures of treatment effects. To that end, Callaway and Sant'Anna (2021) propose weighted means in either the t or the g dimension:

[6] As said in Section 3.2, some states that passed laws before 2000 are coded as never treated. Without additional data gathering, we can't know which are these states and proceed with original data of Cheng and Hoekstra (2013). With this in mind, there are 29 never-treated states, which is a sizable number and allows this to be a valid comparison group. Also, keep in mind that some states might have passed such laws after 2010.

$$\theta_{es}(e) = \sum_g \mathbb{1}\{g + e \leq T\}\mathbb{P}(G_i = g \mid G_i + e \leq T)ATT(g, g+e)$$

$$\theta_{sel}(\tilde{g}) = \frac{1}{T - \tilde{g} + 1}\sum_{t=\tilde{g}}^{T} ATT(\tilde{g}, t)$$

The first aggregation corresponds to a typical event-study, while the second is the average treatment effect for units of group $\tilde{g}$ across all their post-treatment periods. We can further aggregate this measure as

$$\theta_{sel}^{O} = \sum_g \theta_{sel}(g)\mathbb{P}(G_i = g \mid G_i \leq T)$$

which is a scalar that represents a weighted average of treatment effects across all treated units, with weights being the share of group g among all treated.

Over the course of 2000 through 2010, 21 states passed laws that altered self-defense rights, while 29 did not. Thirteen of them were treated[7] in 2007, four in 2008 and two in 2009. Florida was the first state to be treated (2006) and Montana the last (2010).

Thus, one may worry about inference validity, given our relatively small sample size of 50 states (21 treated) across 11 periods. Callaway and Sant'Anna (2021, Supp. Appendix C) show in simulations that, in a setting very similar to ours[8], their method tends to overreject slightly. This distortion is small, however, not surpassing 7.2% in tests of size 5% even for longer horizons of $\theta_{es}(e)$. Crucially, these results depends on correctly specifying the DGP and on the overlap condition being valid when conditioning on covariates.

The biggest issues arises when there are few units per group, which is true in our case except for 2007. Callaway and Sant'Anna (2021) argue that, in these cases, analysis of more aggregate parameters such as θ_{sel}^{O} suffer less from finite sample problems. Finally, note that this is of course more of an issue if we find significant effects, and not so much so if our confidence bands cross zero.

5.1.2 Falsification Tests and Deterrence of Crime

In what follows, we assume unconditional parallel trends. This is mainly because there is no obvious reason that the baseline values of the controls used by Cheng and Hoekstra (2013) influence the *trend* of the outcome, which is what is relevant for the DiD. For example, it is plausible that the baseline unemployment rate influence the level of crime, but it influencing its trend is not an immediate conclusion. Nonetheless, we conduct robustness checks using their region-by-year fixed effects in some cases, which allows for the four different regions of the country to have different criminality trends.

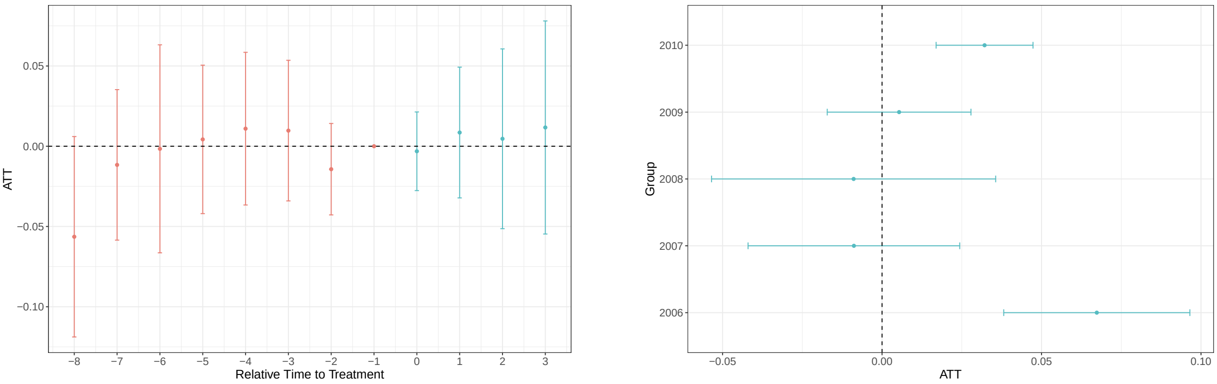
Following the somewhat unusual ordering of Cheng and Hoekstra (2013), we will start with falsification tests, that is, with the impacts of the *Castle Doctrine* laws on crimes that should

[7] Using our binary treatment definition. Laws were passed during the previous year.

[8] In particular, they run simulations with $n = 50$, $T = 4, 20$ and four groups. The difference is that we have $T = 11$ and five groups. Notably, two groups only have one treated state: Florida (2006) and Montana (2010).

not be affected. For that, Cheng and Hoekstra (2013) use larcenies and motor vehicle theft. All dependent variables are logs of the annual per capita value and estimations use the state's populations as weights.

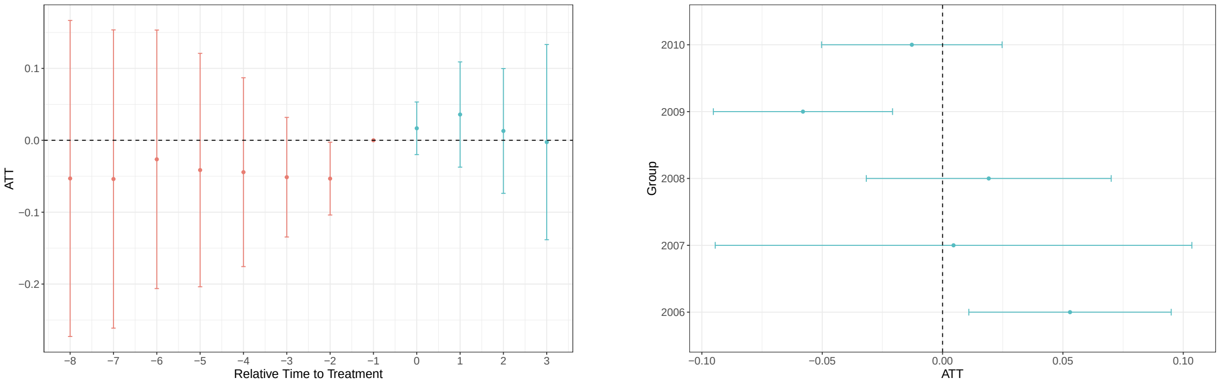
Results are shown in Figures 1 and 2 for larceny and motor vehicle theft, respectively. For all event studies plots in the left-side of the panels, we restrict the relative time horizon to be between -8 and 3, as only Florida has four post-treatment periods. As the $(g, t) = (2006, 2010)$ group is small – unitary in fact –, confidence bands are most likely not valid. For transparency, we plot $\theta_{sel}(\tilde{g})$ for all groups in the right-side of the panels.[9]



(a) Event Study

(b) Effects by Group

Figure 1: Treatment Effects on Larceny



(a) Event Study

(b) Effects by Group

Figure 2: Treatment Effects on Motor Vehicle Theft

As opposed to Cheng and Hoekstra (2013), we now can fully analyze pre-trends, which seems to hold for larceny, but not so much so in vehicle theft. In both event-studies plots in Figures 1a and 2a, we can't reject the null of no treatment effect at any length of exposure using simultaneous confidence bands.

In the group plots in Figures 1b and 2b, we see some effects positive effects for both categories for Florida (2006) and negative effects for motor vehicle theft for the 2009 group, which only has two treated. As mentioned in the previous subsection, inference for these small groups should be done cautiously, as Callaway and Sant'Anna (2021) overrejects in these cases.

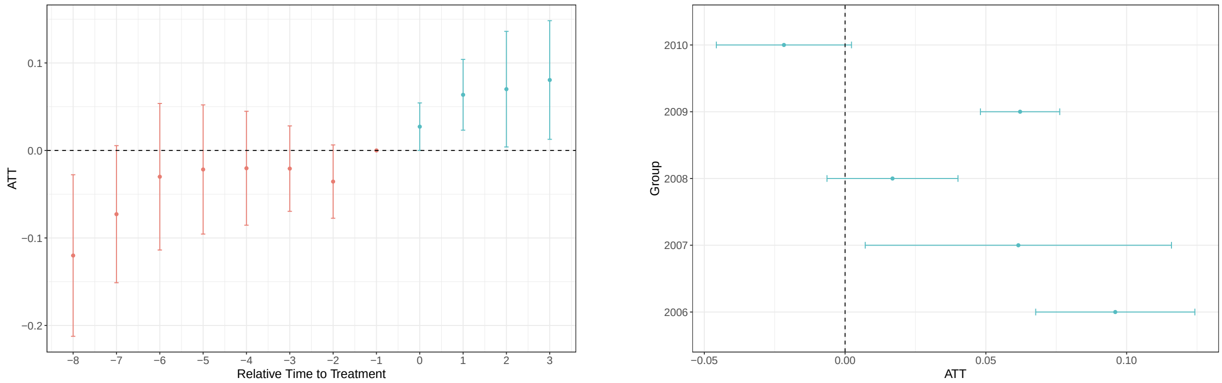
[9] The $\theta_{sel}(\tilde{g})$ for $g = 2006$ (Florida) is now the average of its five post-treatment periods (2006–2010). However, we have a unitary group for Montana (2010), and so its results are likely not precise.

As argued by Callaway and Sant'Anna (2021), it is best to look at the aggregate measure of treatment effects across groups θ^O_{sel}, which gives higher weight to largest groups. We show these measures in Table 3. In both cases, the confidence interval includes zero and we can't reject the null of $H_0 : \theta^O_{sel} = 0$, and so the falsification tests seem to hold.

Table 3: Aggregate Measures of Treatment Effects – Falsification Tests

	ATT θ^O_{sel}	Std. Error	95% Conf. Int.
Larceny	0.0035	0.0152	[-0.0263, 0.0333]
Motor Vehicle Theft	0.0087	0.0241	[-0.0387, 0.0560]

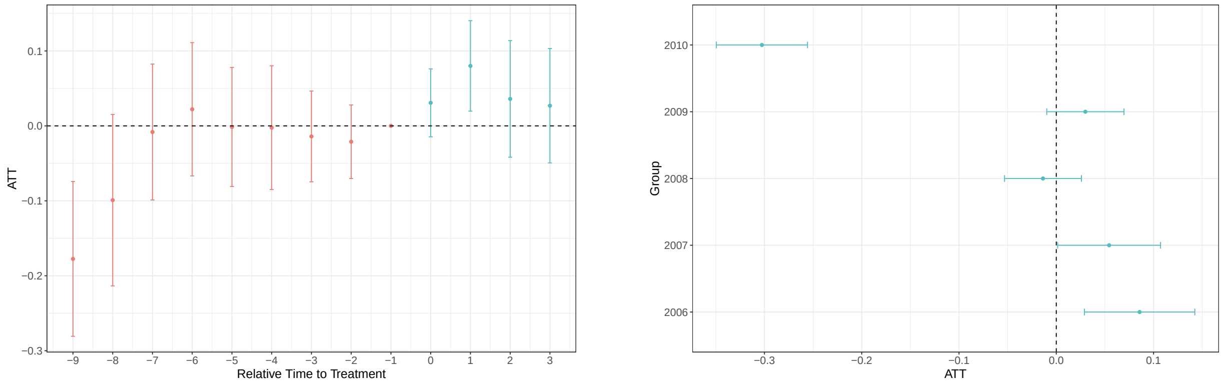
We now proceed to evaluate crimes that may have been deterred by *Castle Doctrine* laws: burglaries, robberies and assaults. Results are shown in Figures 3, 4 and 5, respectively.



(a) Event Study

(b) Effects by Group

Figure 3: Treatment Effects on Burglary



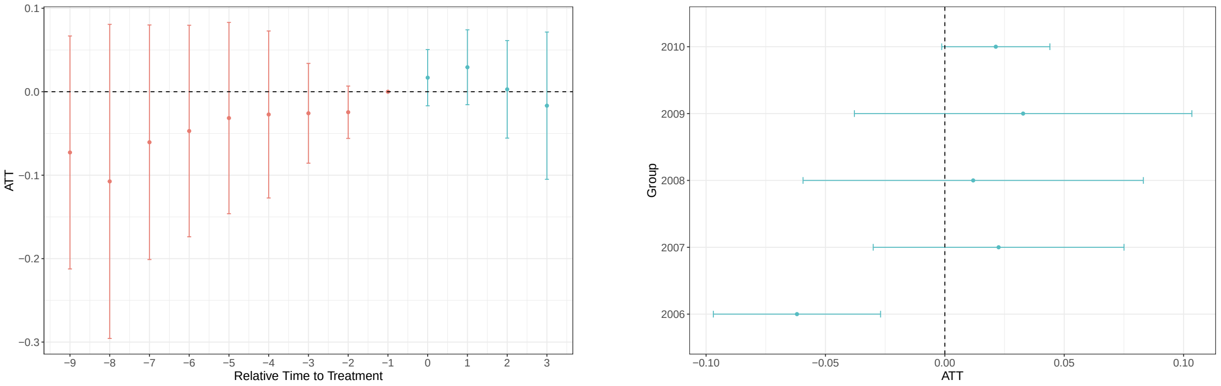
(a) Event Study

(b) Effects by Group

Figure 4: Treatment Effects on Robbery

We have some relatively large pre-trends violations, but only in the eighth pre-treatment periods. We seem to have the opposite of deterrence on burglary rates, which increase in the post-treatment period. We don't have any evidence of deterrence on robbery and on aggravated assaults. The exceptions are robberies on Montana and assaults on Florida, which have a big negative effect which, again, should be inferred with caution.

These results are confirmed when looking at the aggregate ATT measures in Table 4, with burglary being the only significant coefficient.



(a) Event Study (b) Effects by Group

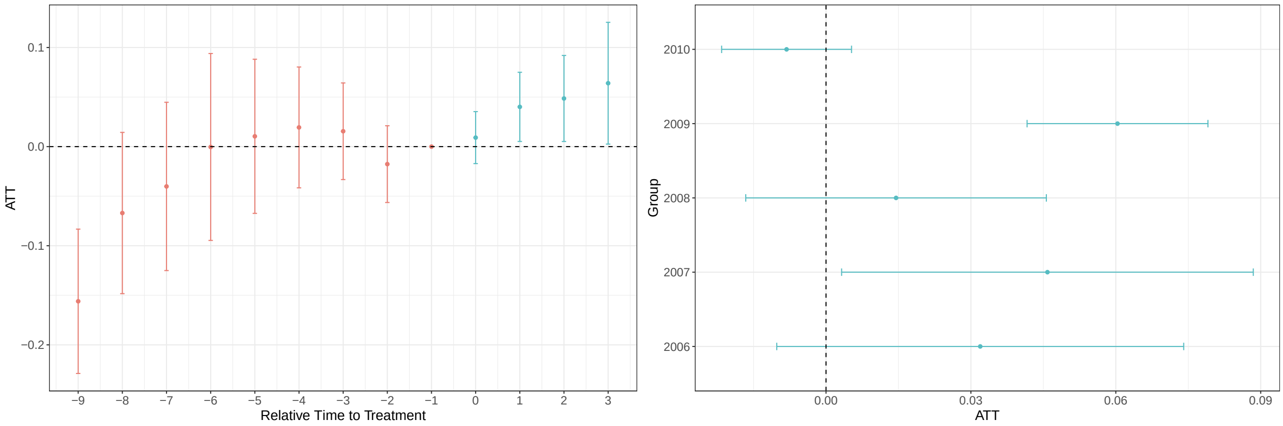
Figure 5: Treatment Effects on Aggravated Assault

Table 4: Aggregate Measures of Treatment Effects – Deterrence

	ATT θ^O_{sel}	Std. Error	95% Conf. Int.
Burglary	0.0533	0.0140	[0.0259, 0.0807]
Robbery	0.0345	0.0643	[-0.0916, 0.1606]
Aggravated Assault	0.0089	0.0187	[-0.0277, 0.0455]

Cheng and Hoekstra (2013) also find a positive effect on burglaries on some specifications, but not on their preferred one. To assess robustness of our results, we include region-by-year fixed effects. Results are shown in Figure 6.

Coefficients are overall smaller, but we still seem to have significant effect. Indeed, the estimate of θ^O_{sel} is 0.0363 with a confidence interval of [0.01, 0.63], which is marginally significant. We view this positive result more as an indicative of no deterrence effects rather than an actual increase on robberies.[10]



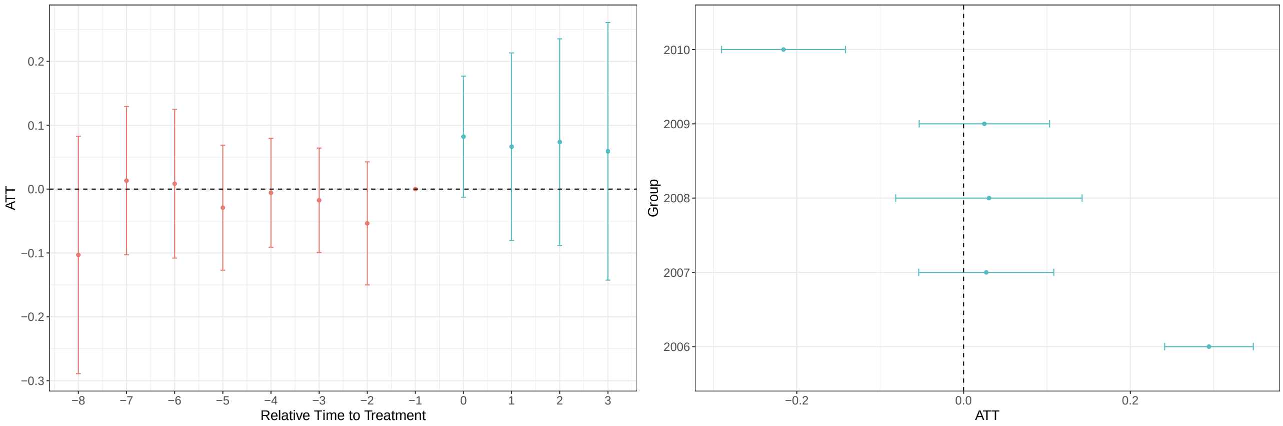
(a) Event Study (b) Effects by Group

Figure 6: Treatment Effects on Burglary with Region-Year Fixed Effects

[10] This is confirmed by the slight overrejection, which we show for iid normal outcomes in Section 5.3.

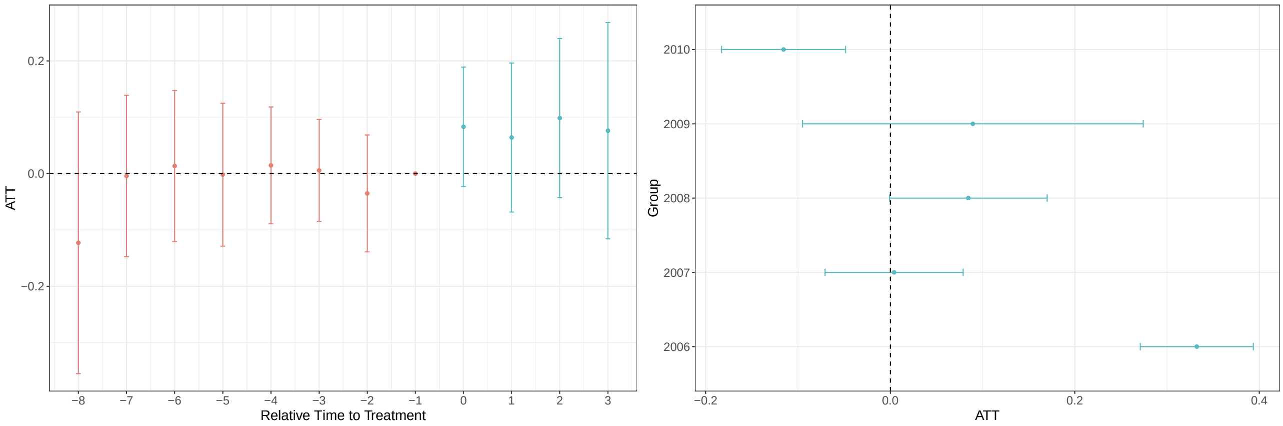
5.1.3 Homicides

We now reanalyze the main results of the paper: the effect of *Castle Doctrine* laws on homicides. We do so both using no controls as well as including region-by-year fixed effects, which forces comparisons to be made within-regions. Results are presented in Figures 7 and 8, respectively. Aggregate measures are shown in Table 5.



(a) Event Study (b) Effects by Group

Figure 7: Treatment Effects on Homicides



(a) Event Study (b) Effects by Group

Figure 8: Treatment Effects on Homicides with Region-Year Fixed Effects

Table 5: Aggregate Measures of Treatment Effects – Homicides

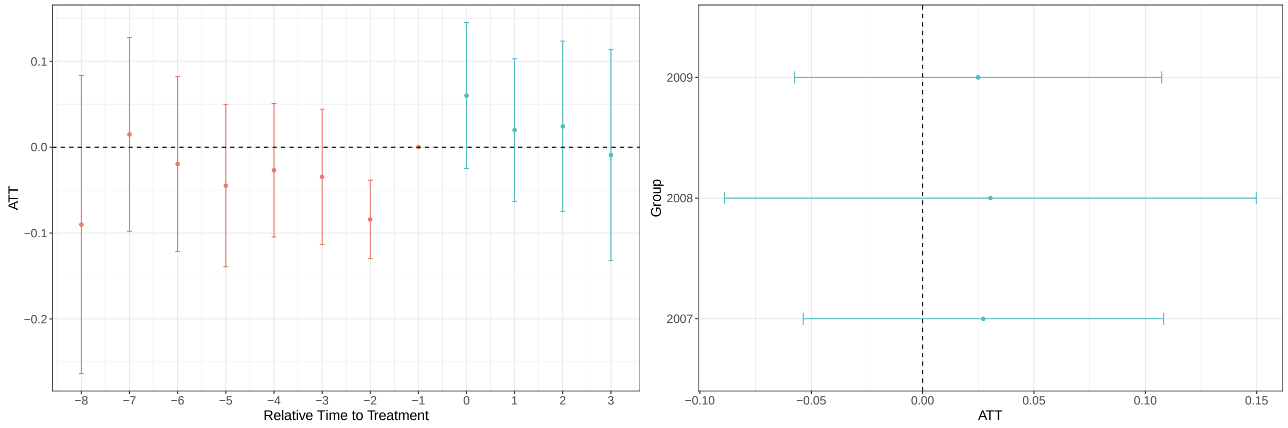
	ATT θ^O_{sel}	Std. Error	95% Conf. Int.
Unconditional	0.0632	0.0368	[-0.0089, 0.1353]
Region-by-Year Fixed Effects	0.0803	0.0236	[0.0341, 0.1265]

We can see that pre-trends holds until eight periods pre-treatment. Furthermore, the coefficients are somewhat constant starting from the seventh pre-treatment period. Together with the falsification tests in Section 5.1.2, it seems plausible that PT holds in our context.

Interestingly, we don't find significant effects in the event studies, although they are positive and somewhat constant over length of exposure to *Castle Doctrine* laws. An advantage of

Callaway and Sant'Anna (2021) is that it allows to see which groups are driving the treatment effect. In this case, 2006 and 2010 (Florida and Montana) are clear outliers.

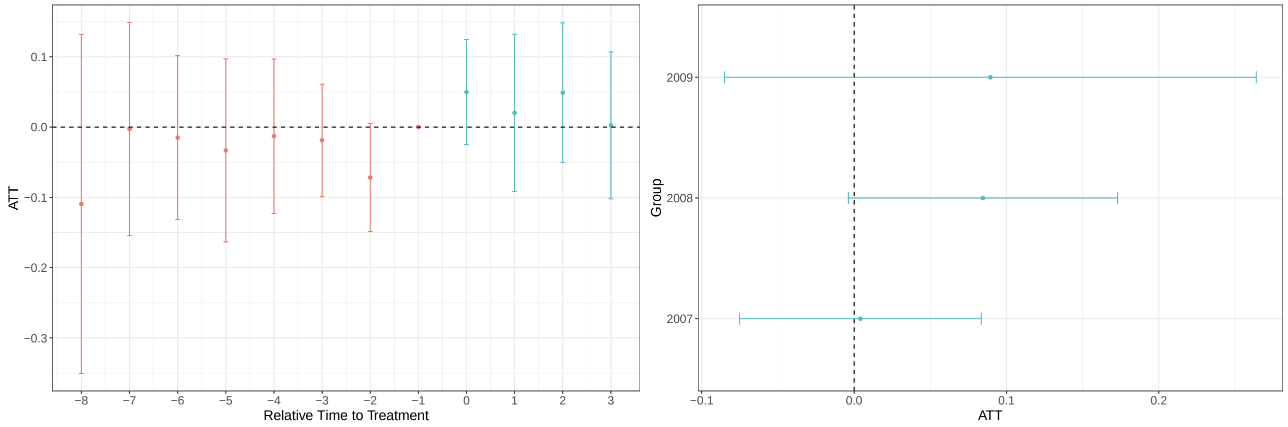
We conduct a robustness test in the spirit of the "leave-one-out" procedure of Abadie, Diamond, and Hainmueller (2010). To that end, we remove Florida and Montana from the sample. Results are shown in Figure 9, Figure 10 and Table 6.



(a) Event Study

(b) Effects by Group

Figure 9: Effects on Homicides Without Florida and Montana



(a) Event Study

(b) Effects by Group

Figure 10: Effects on Homicides with Region-Year Fixed Effects and Without Florida and Montana

Table 6: Aggregate Measures of Treatment Effects – Homicides Without Florida and Montana

	ATT θ^{O}_{sel}	Std. Error	95% Conf. Int.
Unconditional	0.0280	0.0304	[-0.0316, 0.0876]
Region-by-Year Fixed Effects	0.0409	0.0278	[-0.0135, 0.0954]

Without region-year fixed effects, there is evidence of a violation in pre-trends during the second-to-last pre-treatment period. Furthermore, there is a clear trend in the pre-treatment coefficients, which suggests PT does not hold. It is interesting that the evidence supporting our identifying assumption was directly dependent on two outlier states in the case without region-year dummies.

We then focus on the results with region by year fixed effects, which, although too exhibit a trend in the pre-treatment coefficients, do not violate pre-trends, and so we have some more evidence in favor of PT. Without the two outlier states, results are insignificant for all groups and time periods, but still positive.

Together, out results point to *Castle Doctrine* laws having no effect on any sort of crime. Although we find a significant effect on homicides when including region-by-year fixed effects, the results are largely driven by outlier groups with only one treated state. Once these are removed, we have no significant evidence that *Castle Doctrine* increased homicides in the states that adopted it, although our point estimates are still positive.

5.2 Sensitivity Analysis and Bounds Using Pre-Trends

When including Florida and Montana, one may wonder how sensitive the positive homicide results are with respect to the validity of the PT assumption. To that end, we will employ the framework of Rambachan and Roth (2023), which has been applied to the Callaway and Sant'Anna (2021) by Pedro Sant'Anna.

Rambachan and Roth (2023) formalize the intuition that motivates pre-trends testing: the violation of PT cannot be "too different" from the deviations in pre-trends. In particular, they study the case where the PT violations is bounded by $\bar{M}$ times the largest pre-trend coefficient, constructing confidence sets for the target parameter that are uniformly valid.[11]

In the Callaway and Sant'Anna (2021) implementation, this target parameter is a bit different from θ^O_{sel} we have used. Instead of taking group effects and averaging them, we take the event study coefficients and average them across post-treatment periods:

$$\theta^O_{es} = \frac{1}{T-1} \sum_{e=0}^{T-2} \theta_{es}(e)$$

Callaway and Sant'Anna (2021) prefer the group-based aggregation we have been using because it is not sensitive to compositional effects across values of e. Furthermore, note that θ^O_{sel} and θ^O_{es} need not be equal; in fact, they will be the same only if $ATT(g,t)$ is constant for all groups and periods, which provides an informal test on heterogeneous treatment effects.

For transparency, we show θ^O_{sel} and θ^O_{es} for homicides in Table 7. θ^O_{es} is bigger than θ^O_{sel} because we omitted the fourth post-treatment period in Figure 7; this effect is based only on Florida and is estimated to be very large, with a point-estimate of 0.255 and 0.332 in the specification without and with region-by-year fixed effects, respectively.

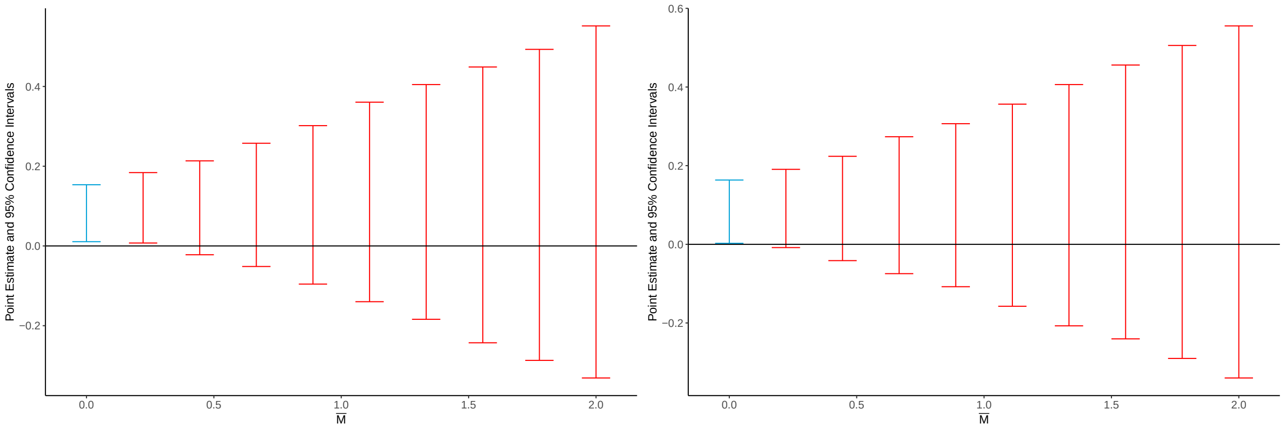
Table 7: Aggregate Measures of Treatment Effects – Homicides

	Group Aggregation			Time Aggregation		
	ATT θ^O_{sel}	Std. Error	95% Conf. Int.	ATT θ^O_{es}	Std. Error	95% Conf. Int.
Unconditional	0.0632	0.0368	[-0.0089, 0.1353]	0.1074	0.0508	[0.0077, 0.2070]
Region-by-Year Fixed Effects	0.0803	0.0236	[0.0341, 0.1265]	0.1321	0.0499	[0.0342, 0.2299]

[11] A very nice feature is that these sets also account for the uncertainty in estimating pre-trends coefficients.

The sensitivity analysis of Rambachan and Roth (2023) is done based on θ^O_{es} considering up to eight pre-treatment periods annd all four possible post-treatment periods. We show these results in Figure 11.

Even with deviations from full PT of less than half the biggest pre-trend deviation and including the huge effects from Florida, results on homicides become insignificant at 5%. This is further evidence toward the null effect of *Castle Doctrine* laws on homicides.



(a) Unconditional (b) Region-by-Year Fixed Effects

Figure 11: Sensitivity Analysis on Homicide Results (Rambachan and Roth 2023)

5.3 Inference and Simulations

As mentioned in Table 1, Cheng and Hoekstra (2013) main results on *Castle Doctrine* laws only hold when using WLS, and not OLS. As argued by Ferman (2022, A.7.1.), this may be due to size distortions when conducting inference using weights.

To assess if cluster-robust estimators are suited to our setting with 21 treated and 29 control states, we run simple simulations following Ferman (2022). We substitute outcomes with iid standard normals. In this case, the null of no treatment effect is true, and so, over $B \to \infty$ simulations of the dataset, we expect the rejection rate of $H_0 : \beta_0 = 0$ to be close to the chosen significance level α if our inference method is valid.

Note that this is a "minimum requirement test", in the sense that passing it does not imply valid inference, but having size distortions even in the best case of outcomes being iid normals should raise red flags. Importantly, this assessment does not detect problems associated with spatial correlation.

To this end, we run $B = 5,000$ simulations substituting the log of homicides per capita by iid normal variables. We use Cheng and Hoekstra (2013)'s preferred TWFE specification and, in each simulation, cluster standard errors at the state level. We do so for both the weighted and unweighted specifications considering nominal sizes of $\alpha = 5\%$.[12]

Over the 5,000 simulations, we get a rejection rate of 6.22% for the unweighted and 8.98% for

[12] In each simulation, we sample different iid normal variables for the weighted and unweighted specifications to avoid any dependence between results.

the weighted specifications. Thus, the weighted specification has a higher degree of overrejection, as in Ferman (2022), and this size distortion may be responsible for the contrasting results in Table 1. Interestingly, the table shows that the difference in standard errors between the weighted and unweighted is indeed larger than the difference in their coefficients.

We plot the empirical CDF of the simulated p-values in Figure 12. Under valid inference, this distribution should be closed to an Uniform$(0, 1)$, which we show as the dashed black line. We see that this is the case for the unweighted specifications, but no so much so for the weighted.

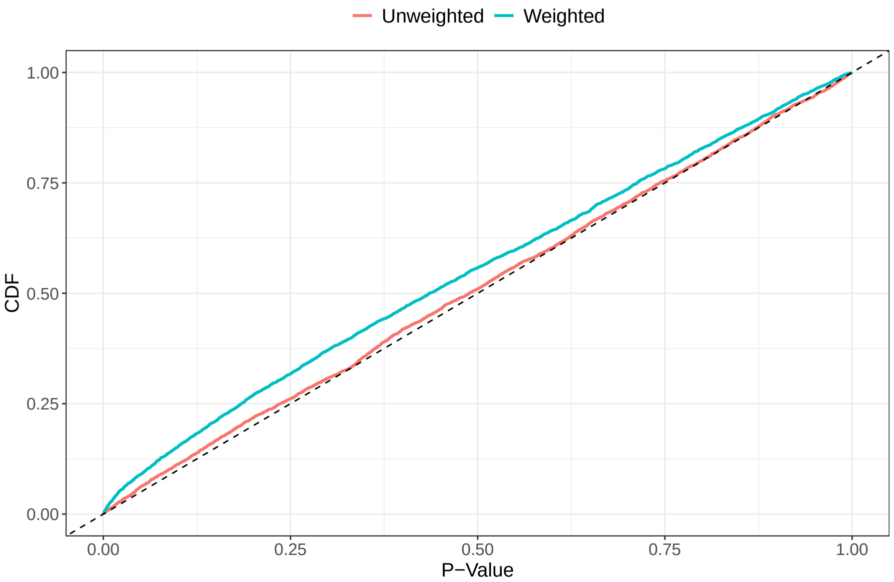


Figure 12: eCDF of Simulated P-Values

Combining the fact that Cheng and Hoekstra (2013) only find positive results in the weighted specification and that we have larger overrjection in this case, the results on homicides may be false positives. It is worth noting that the placebo exercises of Cheng and Hoekstra (2013) mitigate such concerns, but they cannot be ignored.[13]

We run similar simulations for our Callaway and Sant'Anna (2021) specifications. Although unlikely, it could be that we get null results due to underrejection, as their multiplier bootstrap procedure is only valid when the number of clusters goes to infinity. As argued by the authors (Remark 13), finite sample concerns are less pronounced when looking at θ^O_{sel} and θ^O_{es}.

Note that our target parameter is the ATT of *Castle Doctrine* laws, which is better reflected in the aggregate measures θ^O_{sel} and θ^O_{es}. However, if we were interested in the ATT for a particular group or group-time, Ferman and Pinto (2019) would be a natural alternative to conduct inference with few treated if we were willing to make parametric assumptions on the functional form of errors, specially since our dependent variables are aggregated and weighted by population size.

We show the results for $B = 1,000$ simulations in Table 8.[14] Looking at the first two rows, we see that the inference of Callaway and Sant'Anna (2021) tends to overreject in our

[13] Unfortunately, we don't have access to the data spanning 1960-2000 to replicate their placebo tests.

[14] We decrease the number of simulations relative to the previous exercise due to computational constraints.

simulations, with this problem being far more pronounced in θ^O_{es} and when using weights, further corroborating the insights in Ferman (2022).

Curiously, removing the outlier groups of 2006 (Florida) and Montana (2010) improves the validity of the test, leading to correct sizes in θ^O_{sel} in the weighted specifications. This can be either due to them being actual outliers in terms of treatment effects or due to the poor performance of the estimator of Callaway and Sant'Anna (2021) in groups with one treated unit.

Table 8: Rejection Rates of IID Normal Simulations (%) – Callaway and Sant'Anna (2021)

	Dynamic θ^O_{es}		Group θ^O_{sel}	
	Unconditional	Region-Year Fixed Effects	Unconditional	Region-Year Fixed Effects
Unweighted	20.2	18.0	7.7	7.6
Weighted	21.3	19.0	11.4	10.6
Weighted (NoFM)	6.2	5.9	5.3	5.0

Note: "NoFM" = no Florida nor Montana (groups with only one treated state).

6 Conclusion

We have extended Cheng and Hoekstra (2013)'s results considering some of the new methods in the staggered DiD literature and accounting for arbitrary treatment effect heterogeneity. Our results align with theirs in terms of *Castle Doctrine* laws having no deterrence effects on violent crime such as burglary, robbery and assault, as well as on larceny and motor vehicle theft.

However, we reach different conclusions regarding the impact of *Castle Doctrine* on homicide rates. Although all of our point estimates are positive, they lack statistical significance at the conventional 5% level, specially once outlier states – Florida and Montana, the only ones treated in 2006 and 2010, respectively – are removed from the sample. It is worth remembering that the estimator of Callaway and Sant'Anna (2021) performs poorly in groups with only one treated.

Furthermore, even when considering the larger treatment measure θ^O_{es}, the sensitivity analysis of Rambachan and Roth (2023) using the estimator of Callaway and Sant'Anna (2021) shows that the significance of homicide results is not robust to deviations of PT as small as half of the biggest pre-trends violation.

Finally, we provide evidence that the fact that Cheng and Hoekstra (2013) only find significant effects on homicides in the weighted specification may arise due to invalid inference and relevant size distortions, even in the best case scenario of outcomes being iid normal variables (Ferman 2022). The estimator of Callaway and Sant'Anna (2021) also tends to overrejects in these simulations, specially considering the bigger measure of θ^O_{es}. The site distortions disappear when excluding the groups with only one treated state.

Overall, the results of *Castle Doctrine* laws increasing homicide rates are more fragile than what Cheng and Hoekstra (2013) find. However, it has to be said that their randomization inference is perhaps their most convincing evidence: never before have treated states experienced increases in homicides as large as those measured after the passage of laws. Due to the lack of data, we are not able to analyze this aspect.

References

Abadie, A., A. Diamond, and J. Hainmueller. 2010. “Synthetic control methods for comparative case studies: Estimating the effect of California’s tobacco control program.” *Journal of the American Statistical Association* 105 (490): 493–505. (Cited on page 14).

Bertrand, M., E. Duflo, and S. Mullainathan. 2004. “How much should we trust differences-in-differences estimates?” *Quarterly Journal of Economics* 119 (1): 249–275. (Cited on page 6).

Caetano, C., B. Callaway, S. Payne, and H. S. Rodrigues. 2024. *Difference in Differences with Time-Varying Covariates.* arXiv: 2202.02903 [econ.EM]. (Cited on page 6).

Callaway, B., and P. H. Sant’Anna. 2021. “Difference-in-differences with multiple time periods.” *Journal of Econometrics* 225 (2): 200–230. (Cited on pages 6, 8, 9, 10, 11, 14, 15, 17, 18).

Cheng, C., and M. Hoekstra. 2013. “Does Strengthening Self-Defense Law Deter Crime or Escalate Violence?: Evidence From Expansions to Castle Doctrine.” *Journal of Human Resources* 48 (3): 821–854. Available at: https://jhr.uwpress.org/content/48/3/821. (Cited on pages 2, 3, 7, 8, 9, 10, 12, 16, 17, 18).

Conley, T. G., and C. R. Taber. 2011. “Inference with ’difference in differences’ with a small number of policy changes.” *Review of Economics and Statistics* 93 (1): 113–125. (Cited on page 6).

Cunningham, S. 2021. *Causal Inference: The Mixtape.* Yale University Press. (Cited on page 7).

De Chaisemartin, C., and X. d’Haultfoeuille. 2020. “Two-way fixed effects estimators with heterogeneous treatment effects.” *American Economic Review* 110 (9): 2964–2996. (Cited on page 4).

Ferman, B. 2022. *Assessing Inference Methods.* arXiv: 1912.08772 [econ.EM]. (Cited on pages 6, 16, 17, 18).

Ferman, B., and C. Pinto. 2019. “Inference in differences-in-differences with few treated groups and heteroskedasticity.” *Review of Economics and Statistics* 101 (3): 452–467. (Cited on pages 6, 17).

Freyaldenhoven, S., C. Hansen, J. P. Pérez, and J. M. Shapiro. 2021. *Visualization, identification, and estimation in the linear panel event-study design.* Working Paper 29170. National Bureau of Economic Research. Available at: http://www.nber.org/papers/w29170. (Cited on page 5).

Goodman-Bacon, A. 2021. “Difference-in-differences with variation in treatment timing.” *Journal of Econometrics* 225 (2): 254–277. (Cited on page 4).

Huntington-Klein, N. 2023. *Controls in Difference-in-Differences Don't Just Work.* Available at: https://nickchk.substack.com/p/controls-in-difference-in-differences. (Cited on page 6).

Mogstad, M., and A. Torgovitsky. 2024. *Instrumental Variables with Unobserved Heterogeneity in Treatment Effects.* Working Paper 32927. National Bureau of Economic Research. Available at: http://www.nber.org/papers/w32927. (Cited on page 4).

Rambachan, A., and J. Roth. 2023. "A more credible approach to parallel trends." *Review of Economic Studies* 90 (5): 2555–2591. (Cited on pages 5, 15, 16, 18).

Roth, J., P. H. Sant'Anna, A. Bilinski, and J. Poe. 2023. "What's trending in difference-in-differences? A synthesis of the recent econometrics literature." *Journal of Econometrics* 235 (2): 2218–2244. (Cited on page 4).

Sun, L., and S. Abraham. 2021. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects." *Journal of Econometrics* 225 (2): 175–199. (Cited on page 5).