Randomization inference for before-and-after studies with multiple units: an application to a criminal procedure reform in Uruguay

Matias D. Cattaneo[1] ⓘ **, Carlos Díaz[2] and Rocío Titiunik[3‡]** ⓘ

[1]Department of Operations Research and Financial Engineering, Princeton University, Sherrerd Hall, Charlton St, Princeton, NJ 08544, USA
[2]Millennium Nucleus on Criminal Complexity and School of Economics and Business, Universidad Alberto Hurtado, Erasmo Escala 1835, Santiago, Región Metropolitana, Chile
[3]Department of Politics, Princeton University, 001 Fisher Hall, Princeton, NJ 08544, USA

Address for correspondence: Rocío Titiunik, Department of Politics, Princeton University, 001 Fisher Hall, Princeton, NJ 08544, USA. Email: titiunik@princeton.edu

Abstract

Learning about the immediate causal effects of large-scale policy interventions poses a significant challenge for quasi-experimental methods that rely on long-term trends or parametric modelling assumptions. As an alternative, we develop a randomization inference framework for before-and-after studies with multiple units, designed specifically for short-term causal inference and allowing for general assignment mechanisms. The method provides finite-sample-valid statistical inferences without relying on parametric time series models or extrapolation. We demonstrate its utility by analyzing a major criminal justice reform in Uruguay that switched from an inquisitorial to an adversarial system in November 2017. Our method relies on the key assumption of no local time trends near the policy adoption time, which is supported by several falsification tests in our empirical study. We find a statistically significant short-term causal effect: an increase of approximately 25 daily police reports (an 8% rise) in the first week of the new justice system. Our randomization inference framework provides a robust and flexible methodology for evaluating policy adoptions in before-and-after studies with multiple units.

Keywords: before and after studies, crime, criminal procedure, event studies, interrupted time series studies, randomization inference

1 Introduction

On 1 November 2017, a new code of criminal procedure (CCP or *Código del Proceso Penal* in Spanish) became effective in Uruguay. The reform ended the old inquisitive and written tradition, and established an accusatory, adversarial, oral, and public criminal system in its place. Under the old CCP, the inquisitorial judge led the investigation and decided the appropriate punishment for a particular crime. Once the new criminal procedure came into effect, judges became neutral referees focussed on ensuring the correct procedure, while prosecutors became responsible for leading the investigation. Under the new regime, the investigation is the exclusive responsibility of prosecutors who, representing society, must present evidence to judges; judges then decide what evidence to admit into the record. This separation between the roles of prosecutor and judge is a key reason why adversarial systems are generally considered fairer and less susceptible to abuse than inquisitorial systems.

Despite its advantages, this type of reform can lead to unintended changes in the costs associated with offending. The new CCP may have changed both the severity and the certainty of

---

‡ Ringold Institution ID 6740

punishments through several channels, possibly affecting the propensity and ability of individuals to commit crimes. For instance, the new adversarial system introduced substantial changes to the adjudication process of criminal law such as plea bargain, alternatives to oral trials, and exceptional use of preventive detention, all of which might result in lighter sentences. In addition, prosecutors faced both a new role and a significant increase in their workload, while policing was affected by new rules and supervision of police investigations. These changes resulted in coordination problems between prosecutors and police officers during the first months of the adversarial system, possibly affecting the probability of detection and conviction.

We propose a methodology to evaluate the immediate impact of this procedural reform on the number of offences reported to the police in Montevideo, the capital and largest city of Uruguay. The reform was introduced nationwide simultaneously but, in contrast to a standard interrupted time series design (e.g. Cook et al., 2002), we observe multiple units in each time period because we collect crime reports (*denuncias*) for each of the 62 neighbourhoods in Montevideo. This leads to a before-and-after observational study with multiple units: a setting where several cross-sectional units are first observed for several periods, a policy is then adopted at the same time for all units, and finally, the same units are observed after the intervention for several more periods.

This type of observational study is often analyzed using linear panel data models where the outcome for unit i in period t is typically regressed on a unit's fixed effect, a time fixed effect, time trends, unit-level covariates, and time-indexed treatment indicators—see Freyaldenhoven et al. (2019, 2025) and Miller (2023) for reviews. This empirical strategy requires many units and many time periods, and relies on time homogeneity over a typically long time span. The number of parameters to be estimated is often large, and multicollinearity issues are common. A relatively small number of units can make estimation of the time fixed effects unstable, while a similar problem occurs for the estimation of unit fixed effects with few time periods. The linear panel event study approach also relies on a parametric specification to model the global trend of the outcome to separate it from the effect of the policy. As a consequence, this observational method is most useful when many units are observed over many time periods, and the time homogeneity of the flexible parametric regression model is plausible, allowing researchers to model the time series globally over a long time span. Furthermore, a crucial limitation of before-and-after studies is the lack of a control group (Miller, 2023), which suggests focussing on small windows of time around the intervention to isolate the treatment effect from potential time confounders.

We consider an alternative approach that is tailored to learning about the immediate effect of the new CCP on the number of police reports, while allowing for a small number of units and time periods, avoiding global time series specifications, and providing robustness to cross-sectional and time-series dependence. Building on the causal inference literature (Rosenbaum, 2002, 2010), we develop a Fisherian randomization inference approach for before-and-after studies with multiple units. Our approach is based on localizing around the time of the policy adoption rather than modelling the time series globally, which serves two purposes simultaneously: (i) explicitly targets the immediate causal effect of the policy, and (ii) removes the need for parametric time series modelling. In our framework, the units' potential outcomes are nonstochastic, and we assume that the policy adoption time can be approximated by a known distribution.

While in reality all units are untreated in the preintervention period and treated in the postintervention period, we consider two hypothetical assignment mechanisms: in the *Treatment Reversal* (TR) assignment mechanism, each unit could have been treated either in the preintervention or in the postintervention period, while in the *Adoption Timing* (AT) assignment mechanism, each unit is always untreated before adoption and treated afterwards, but the timing of the intervention could have been sooner or later than the actual adoption time. Because the potential outcomes are nonstochastic, both assignment mechanisms provide the distribution of any test statistic under suitable assumptions and null hypotheses, providing valid p-values to hypothesis testing. Moreover, our framework is general and allows for other assignment mechanisms in addition to the two we propose.

Our hypothesized randomization mechanisms are more easily justified for a small window around the time of adoption of the new policy. Moreover, the TR mechanism specifically relies on the assumption that the average potential outcomes in a window around adoption time exhibit no trends, which is implausible when the window is large. Because we observe multiple units for every time period, we can analyze the smallest possible window around the time of the

intervention, thereby reducing extrapolation to a minimum. To guide the choice of windows larger than the minimum, we propose a procedure based on the procedure by Cattaneo et al. (2015): choosing an artificial (or 'placebo') adoption time well before the adoption of the real intervention, we test the null hypothesis of no treatment effect for the main outcome for a sequence of nested windows of increasing length, stopping when the hypothesis is rejected. Because the treatment effect is known to be zero, this procedure guides the choice of window around the real adoption time under the assumption that time trends in the placebo windows are the same as in our real window. We show that, for various artificial adoption times, windows of length up to approximately 14 days within the cutoff have outcome trends that cannot generally be distinguished from zero. We further assess our assumptions with a falsification analysis that, using the chosen window length, replicates the analysis using data from the past.

We find that the implementation of Uruguay's new criminal procedure resulted in a local increase in the number of crimes reported to the police in Montevideo of about 8.2 percent in the week after the reform. These estimates are consistent with the view that legal codes are far from innocuous (Acemoglu & Johnson, 2005). Our randomization-based approach indicates that the effect of the policy is distinguishable from zero within 7 days of the intervention with the TR mechanism (p-value ≤ 0.02) and within 14 days with both the TR and the AT mechanism (p-values ≤ 0.03). Our inferences also reject the null hypothesis of no effect when multiple windows within 14 days are considered jointly.

Our focus on immediate or short-run effects is well suited to before-and-after designs, which are typically employed to study wide-ranging policy reforms that are adopted everywhere in a country or region. The immediate effects of such policies are often critical for the reform's trajectory, as they have the potential to shape public opinion and influence its long-term success o failure (Jain & Mukand, 2005; Stokes, 1996, 2001). In our case study, the documented increase in crime reports appears to have affected the perceived legitimacy of the new adversarial system during its initial phase, which led to future changes and impacted the ultimate features of the new adversarial regime. More generally, when a major legal reform is implemented during a time of heightened public concern, short-term results may have long term consequences. Unrealistic expectations can quickly turn into frustration when early indicators fail to show progress, raising the risk of backsliding or even reversal.

1.1 Related literature

Methodologically, our approach contributes to a rich literature on randomization inference methods for program evaluation and causal inference, including methods for the analysis of natural experiments (Ho & Imai, 2006), instrumental variables (Imbens & Rosenbaum, 2005; Kang et al., 2018), and regression discontinuity designs (Cattaneo et al., 2015, 2017). Our approach is also connected to the so-called regression discontinuity (RD) design in time (Hausman & Rapson, 2018), where there are multiple cross-sectional units and the running variable is defined as time to the event—a setting that is not a standard RD design (Cattaneo & Titiunik, 2022). Hausman and Rapson (2018) identify a dilemma in applying standard RD methods to before-and-after studies: using a large region of data around the policy change can lead to extrapolation, while using a local region often leaves too few observations for valid asymptotic approximations. Our method circumvents these limitations by localizing near the treatment time, avoiding the extrapolation inherent in standard parametric approaches. See Cattaneo et al. (2024, Section 2) for an introduction to local randomization methods in the context of standard RD designs.

Substantively, our research adds to the economics of crime literature by showing how procedural law impacts delinquency. More lenient punishments—often associated with plea bargaining (Bushway & Redlich, 2012)—and a lower probability of being caught and convicted have been argued to increased crime (Becker, 1968). Prior empirical studies have shown that changes in evidentiary rules, trial length, court efficiency, and enforcement incentives can all affect criminal behaviour (Atkins & Rubin, 2003; Dalla Pellegrina, 2008; Dušek, 2015; Soares & Sviatschi, 2010). In Latin America, the shift to adversarial systems has produced mixed results (Langer, 2007). For example, (Zorro Medina et al., 2023) documents increases in crime due to reduced detection rates in Colombia, while Kronick (2024) finds no effects on crime despite higher arrest rates.

2 Criminal procedure reform in uruguay

Uruguay was one of the last countries in Latin America to reform its code of criminal procedure. Beginning in the early 1990s, procedural reforms spread across the region and most countries transitioned from inquisitorial to adversarial systems—with the notable exceptions of Brazil and Cuba, which still preserve their traditional frameworks (Fandiño & González Postigo, 2020). This shift constituted a major transformation of the region's criminal justice system (Langer, 2007).

Historically, Latin America's criminal procedures were governed by inquisitorial, written systems introduced in the nineteenth and early twentieth centuries. As described by Langer (2007), these systems typically shared two core features, both of which also characterized the Uruguayan CCP. First, proceedings were divided into two written phases: a pretrial investigation phase and a verdict-and-sentencing phase, with the latter largely based on a dossier compiled by judges and police. Second, the judge played a dual role—leading the investigation and later adjudicating the case. The defendant was excluded from the investigation, which remained secret, and pretrial detention was the norm. Over time, however, the rising prominence of human rights (in the 1970s) and the region's democratization processes (in the 1980s and 1990s) led to mounting concerns that such systems failed to uphold basic due process guarantees (Langer, 2007; Maier & Struensee, 2000).

In response, countries began to adopt adversarial, accusatory, oral, and public systems, featuring three well-defined stages: the formalization of the investigation, a preliminary hearing, and a trial phase (Fandiño & González Postigo, 2020). Under this model, the preliminary investigation is conducted by the police and led by a prosecutor, and the defence and prosecution present their arguments in public hearings. Uruguay's reformed system closely follows this structure. During the first stage, the General Prosecution Office decides whether to pursue criminal charges (i.e. the judicialization of the case). If the investigation is formalized, the second stage consists of a preliminary hearing to review investigative outcomes. Finally, an oral trial is held for cases not resolved through alternative outcomes. Uruguay has a classic adversarial procedure where defence and prosecution arguments are presented at a hearing.

The adoption of Uruguay's adversarial CCP marked a complex and far-reaching institutional reform. Although the law was passed on 19 December 2014, the new system did not take effect until 1 November 2017. The scale and scope of the changes required nearly 3 years of preparation. The reform redefined the roles of key actors in the criminal justice system—particularly prosecutors and police officers. Under the new criminal adjudication law, prosecutors became responsible for directing investigations, representing a fundamental change in both their function and workload. The transition was so demanding that the prosecutors' union publicly opposed the reform, citing overwhelming workloads. Some prosecutors were assigned several hundred cases simultaneously, leading many to take mental health leave during the first year of implementation (Solomita, 2019). Police officers, for their part, were required to adapt to new rules and supervision, with a prosecutor overseeing the case instead of a judge. In sum, the reform affected many aspects of the criminal system. For this reason, when we refer to the effect of the CCP reform, we mean the combined effect of the bundle of treatments that were changed as part of the reform, not simply the adoption of a more due-process-oriented system.

Important for the interpretation of our results, the public was intentionally informed about the reform, as the government engaged in efforts to disseminate information and bring public awareness. The Office of the Attorney General launched a public information campaign, as a result of which the reform received considerable media coverage. In Section S4 in the online supplementary material, we show a billboard that was placed in Montevideo bus stops highlighting the transparency and guarantees of the new procedural law.

Uruguay experienced an unprecedented increase in the number of police reports following the implementation of the new CCP. While there was broad consensus around the normative desirability of transitioning to an adversarial model, the reform was blamed by public officials for the observed rise in insecurity. In particular, both the President and the Interior Minister referred to the sharp increase in police reports as the 'November Effect', invoking the date of adoption of the reform (Palumbo, 2018). Although the short-term impacts led to eventual modifications of the policy such as clarifying guidelines from the Attorney General and even legislative amendments, these changes occurred well outside the time periods we analyze. Our approach does rely on the

Table 1. Average daily crimes reported to police in Montevideo

	Two-Year Window 11/01/16 to 10/31/18	Old CCP (a)	New CCP (b)	Δ (b)-(a)
Theft	152.19	130.05	174.33	44.28
	46%	(0.942)	(1.221)	(1.543)
Robbery	51.81	41.60	62.02	20.42
	16%	(0.426)	(0.614)	(0.747)
Domestic Violence	33.66	32.79	34.53	1.75
	10%	(0.379)	(0.382)	(0.539)
Other Crimes	92.98	78.98	106.99	28.02
	28%	(0.662)	(0.869)	(1.092)
Total Police Reports	330.63	283.41	377.87	94.47
	100%	(1.524)	(2.013)	(2.525)
Days	730	365	365	

Note. Percentage of total number of reports in *italic*; standard errors in parentheses.
Old CCP: inquisitorial system (1 November 2016 to 31 October 2017).
New CCP: adversarial system (1 November 2017 to 31 October 2018).
Source: Ministry of Interior of Uruguay.

assumption that the nature of the reform stayed constant in the window around the reform's adoption time that we use for analysis; we are not aware of any modifications to the implementation of the reform during this period.

2.1 Data and overall patterns

We collected data on the offences reported to the police in Montevideo from the Ministry of Interior of Uruguay. Montevideo is the capital and largest city of Uruguay, home to 40% of Uruguay's total population. We first present a descriptive analysis of the years before and after the adoption of the reform. Since the new CCP came into effect on 1 November 2017, we start by building a two-year symmetric window from 1 November 2016 through 31 October 2018. Table 1 reports the average number of crimes reported daily to the police during this overall period, as well as for the years before and after the switch to the new CCP. An average of 331 crimes were reported to police every day during this two-year period, with three categories accounting for more than 7 of every 10 reports processed in Montevideo: theft ($\approx$ 46% of all offences reported to police), robbery ($\approx$ 16%) and domestic violence ($\approx$ 10%).

Table 1 shows a considerable increase in the number of police reports in Montevideo during the first year of the new CCP. When we compare the last year of the old system to the first year of the new system, the total number of daily reported offences increases by approximately 95 incidents (from roughly 283 to 378 per day), a statistically significant difference that represents a 33% increase. This upward trend is also reflected in the three most frequent crimes reported in Montevideo. Regarding property crimes, the number of thefts and robberies reported to police every day increased by 34% and 49%, respectively. Reports of domestic violence, the most frequent crime against the person, exhibited a much smaller increase of 5%.

An initial visual inspection of the data suggests a strong association between the increase in police reports and the timing of the CCP reform. Figure 1 plots the daily number of offences reported to the police in Montevideo from 600 days before the new CCP came into effect (early 2016) until 400 days after (late 2018).

Figure 1 suggests an increase in the number of police reports shortly after the new CCP came into effect, consistent with the previously mentioned 'November Effect'. To identify a potential immediate causal effect, all our analyses will focus on time windows contained within 21 days before and 21 days after the date of implementation of the new CCP, as marked by the dashed red lines in Figure 1.
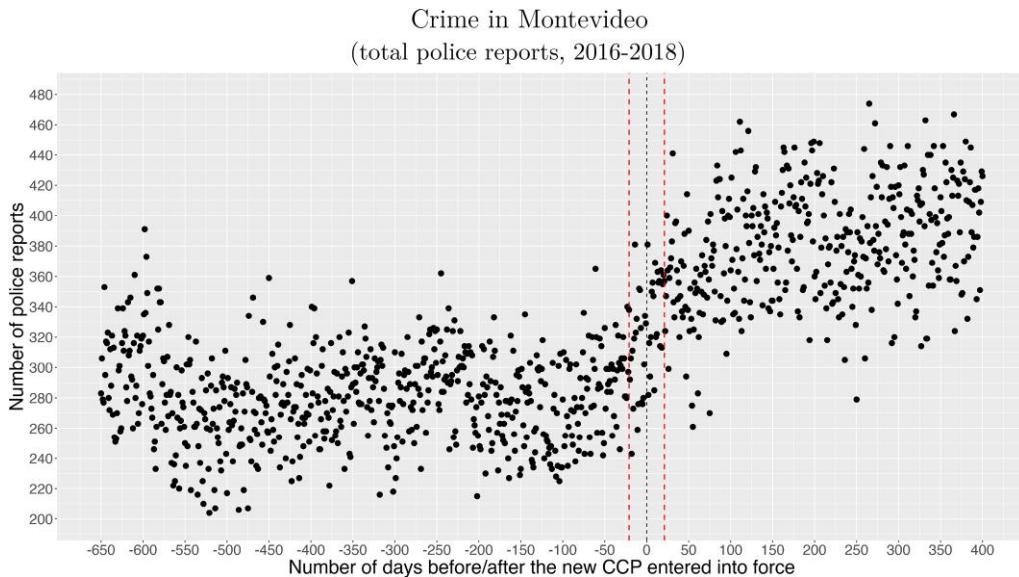
Figure 1. Crime in Montevideo (total police reports, 2016–2018).

3 Randomization inference framework

Given the possibility of time trends and other confounders, the patterns reported in the prior section cannot be taken as conclusive evidence that Montevideo's crime wave was caused by the changes to Uruguay's CCP. We develop a randomization inference framework to provide further empirical evidence that may address these (and other) methodological concerns. Our framework is designed to analyze a before-and-after study with multiple units, that is, a setting where multiple cross-sectional units are first observed for some periods, a reform affecting all units is introduced at the same time for all units, and then the same units are observed after the reform for some more periods.

We adopt a Fisherian framework with nonrandom potential outcomes. Each cross-sectional unit $i = 1, 2, \ldots, n$ is observed over $t = 1, 2, \ldots, T$ time periods, and an intervention is adopted at time period $t = a_0 \in \{1, 2, \ldots, T\}$. The fixed treated and untreated potential outcomes at time t for unit i are denoted, respectively, by $y_{i,t}(1)$ and $y_{i,t}(0)$. This notation implicitly assumes that the specific time of adoption a_0 can only affect the potential outcomes indirectly via the triggering of the intervention, but not directly—an exclusion restriction that is most plausible within small time windows around a_0. In our sample, $n = 62$ neighbourhoods, and the new CCP began precisely at 12:00 AM on 1 November 2017, becoming active at the same specific moment for all units; thus, in the actual policy assignment, all units are untreated in all periods $t < a_0$ and treated in all periods $t \geq a_0$.

Assuming enough time periods are available, we define a window including τ periods before and τ periods after the time of policy adoption,

$$\mathcal{W}_\tau = \{a_0 - \tau, \ldots, a_0 - 1, a_0, a_0 + 1, \ldots, a_0 + \tau - 1\},$$

where we consider symmetric windows only for simplicity. The length of $\mathcal{W}_\tau$ will depend on the unit of time. Since each unit appears in every period, the total sample size for $\mathcal{W}_\tau$ is $2\tau n$ observations. For example, in our application t measures days, $\mathcal{W}_1 = \{a_0 - 1, a_0\}$ includes the day immediately before adoption and the adoption day, and hence there are 124 observations within $\mathcal{W}_1$. We discuss a data-driven procedure for selecting τ in Section 3.4.

We denote the treatment indicator by $D_{i,t}$, which is equal to one if unit i is treated at time t and zero otherwise. This is the random variable that determines the assignment of the intervention in each period for each unit. The (random) observed outcome is

$$Y_{i,t} = D_{i,t} \cdot y_{i,t}(1) + (1 - D_{i,t}) \cdot y_{i,t}(0).$$

In a before-and-after design, all units are untreated during some periods and treated during others. For each unit, we define two average observed outcomes within $\mathcal{W}_\tau$:

$$\bar{Y}_{i,\tau,1} = \frac{1}{N_{i,\tau,1}} \sum_{t \in \mathcal{W}_\tau} D_{i,t} Y_{i,t}, \quad N_{i,\tau,1} = \sum_{t \in \mathcal{W}_\tau} D_{i,t},$$

and

$$\bar{Y}_{i,\tau,0} = \frac{1}{N_{i,\tau,0}} \sum_{t \in \mathcal{W}_\tau} (1 - D_{i,t}) Y_{i,t}, \quad N_{i,\tau,0} = \sum_{t \in \mathcal{W}_\tau} (1 - D_{i,t}),$$

which correspond to the average outcome across all periods $t \in \mathcal{W}_\tau$ during which the unit is treated, and the average outcome across all periods $t \in \mathcal{W}_\tau$ during which the unit is untreated, respectively.

According to the actual intervention, $D_{i,t} = \mathbb{1}(t \geq a_0)$ for all $i = 1, 2, \ldots, n$. Our approach involves hypothesizing different possible assignment mechanisms governing D_{it}. For a time of adoption a, we define four average potential outcomes within the window $\mathcal{W}_\tau$ for every unit, corresponding to treated and untreated potential outcomes averaged over the preintervention $(t < a)$ or postintervention $(t \geq a)$ periods:

$$\bar{y}_{i,\tau,-}(0, a) = \frac{1}{n_-(a)} \sum_{t \in \mathcal{W}_\tau} \mathbb{1}(t < a) y_{i,t}(0), \quad \bar{y}_{i,\tau,-}(1, a) = \frac{1}{n_-(a)} \sum_{t \in \mathcal{W}_\tau} \mathbb{1}(t < a) y_{i,t}(1),$$

with $n_-(a) = \sum_{t \in \mathcal{W}_\tau} \mathbb{1}(t < a)$, and

$$\bar{y}_{i,\tau,+}(0, a) = \frac{1}{n_+(a)} \sum_{t \in \mathcal{W}_\tau} \mathbb{1}(t \geq a) y_{i,t}(0), \quad \bar{y}_{i,\tau,+}(1, a) = \frac{1}{n_+(a)} \sum_{t \in \mathcal{W}_\tau} \mathbb{1}(t \geq a) y_{i,t}(1),$$

with $n_+(a) = \sum_{t \in \mathcal{W}_\tau} \mathbb{1}(t \geq a)$, and where $n_-(a) + n_+(a) = 2\tau$ for all $a \in \mathcal{W}_\tau$. Of these average potential outcomes, we only observe $\bar{y}_{i,\tau,-}(0, a_0)$ and $\bar{y}_{i,\tau,+}(1, a_0)$ for $i = 1, \ldots, n$; all the others are counterfactual and thus unobserved.

Our notation explicitly distinguishes between observed and potential averages: the subindeces 0 and 1 used in $\bar{Y}_{i,\tau,1}$ and $\bar{Y}_{i,\tau,0}$ denote, respectively, untreated and treated observed average outcomes whenever they may occur, while the $-$ and $+$ subindeces in the nonrandom averages $\bar{y}_{i,\tau,+}(\cdot, a)$ and $\bar{y}_{i,\tau,-}(\cdot, a)$ denote, respectively, preintervention and postintervention periods relative to adoption time $t = a$, regardless of whether the periods entering the averages were untreated or treated according to the assignment D_{it}.

Finally, to describe different assignment mechanisms, we define the vector $\mathbf{D}_t = (D_{1,t}, D_{2,t}, \ldots, D_{n,t})'$ for each period $t \in \mathcal{W}_\tau$ and collect them into the $n \times 2\tau$ matrix

$$\mathbf{D} = [\mathbf{D}_{a_0-\tau}, \ldots, \mathbf{D}_{a_0-1}, \mathbf{D}_{a_0}, \mathbf{D}_{a_0+1}, \ldots, \mathbf{D}_{a_0+\tau-1}].$$

In an unrestricted setting where each unit could be assigned to control or treatment in any period, we could observe zeros or ones in any of the entries of $\mathbf{D}$ regardless of the values taken by the other entries, resulting in $2^{2n\tau}$ different possible assignment matrices. A before-and-after design, however, places known restrictions on the columns of $\mathbf{D}$. Our approach to inference considers two different assignment mechanisms that restrict the matrix $\mathbf{D}$ in different ways, and lead to two different randomization distribution of test statistics that can be used to test null hypotheses. Other assignment mechanisms in addition to the ones we describe can be accommodated with different restrictions $\mathbf{D}$.

3.1 Treatment reversal (TR) assignment mechanism

The first assignment mechanism keeps the adoption time at $t = a_0$ for all units and hypothesizes a possible reversal of the treatment condition between the period before and the period after a_0. For each unit, this mechanism independently assigns the treatment to either all preintervention periods or all postintervention periods in $\mathcal{W}_\tau$. For example, if the window is $\mathcal{W}_2 = \{a_0 - 2, a_0 - 1, a_0, a_0 + 1\}$, the TR mechanism assigns all the postintervention periods ($t \in \{a_0, a_0 + 1\}$) to treatment—and therefore assigns all the preintervention periods ($t \in \{a_0 - 2, a_0 - 1\}$) to control—with some known probability. Although in the realized assignment all neighbourhoods in Montevideo started the new CCP on 1 November 2017, this mechanism imagines that the assignment of neighbourhoods to treatment could have been reversed, with some neighbourhoods instead assigned to the old CCP on 1 November 2017, and the new CCP before this date.

Every unit thus has two possible assignments, $\{t : t < a_0, t \in \mathcal{W}_\tau\}$ treated and $\{t : t \geq a_0, t \in \mathcal{W}_\tau\}$ control, or vice-versa. To formalize, we define a binary random variable Z_i that is equal to one when unit i is assigned the treatment in the postintervention period, and is equal to zero otherwise. The treatment indicator becomes

$$D_{i,t} = (1 - Z_i)\mathbb{1}(t < a_0) + Z_i\mathbb{1}(t \geq a_0), \quad t \in \mathcal{W}_\tau,$$

for all units $i = 1, \ldots, n$.

In this assignment mechanism, if the first postintervention period is treated (untreated), we know all subsequent periods are treated (untreated) and all preintervention periods are untreated (treated). Thus, $D_{i,t} = D_{i,a_0-1}$ if $t < a_0$, and $D_{i,t} = D_{i,a_0}$ if $t \geq a_0$, for $t \in \mathcal{W}_\tau$. It follows that $D_{i,-1} + D_{i,0} = 1$, that is, exactly one of the two periods before and after adoption time is assigned to treatment. These restrictions reduce the number of possible assignment matrices $\mathbf{D}$ from $2^{2\tau n}$ to 2^n; we collect these matrices in the set $\mathcal{D}_{TR}$. Because we assume that each matrix assignment is equally likely, we have $\mathbb{P}[\mathbf{D} = \mathbf{d}] = 2^{-n}$ for the allowable $\mathbf{d} \in \mathcal{D}_{TR}$.

The conditions for using a randomization inference approach under the TR assignment mechanism are summarized in the following assumption.

> **Assumption 1**   ((TR Assignment Mechanism)). There exists a window $\mathcal{W}_\tau$ of τ periods before and after the adoption time a_0 of the intervention such that the following conditions hold for all $i = 1, 2, \ldots, n$.
>
> (i) $\bar{y}_{i,\tau,-}(0, a_0), \bar{y}_{i,\tau,-}(1, a_0), \bar{y}_{i,\tau,+}(0, a_0)$ and $\bar{y}_{i,\tau,+}(1, a_0)$ are nonstochastic.
> (ii) $\bar{y}_{i,\tau,-}(0, a_0) = \bar{y}_{i,\tau,+}(0, a_0)$ and $\bar{y}_{i,\tau,-}(1, a_0) = \bar{y}_{i,\tau,+}(1, a_0)$.
> (iii) $\mathbb{P}[\mathbf{D} = \mathbf{d}] = 2^{-n}$ for all $\mathbf{d} \in \mathcal{D}_{TR}$.

Assumption 1(i) allows for the deployment of Fisherian inference methods because it fixes the potential outcomes. Assumption 1(ii) rules out time trends in the average potential outcomes, which allows us to distinguish the effect of the treatment from over-time changes in the outcomes and thus impute all missing potential outcomes under our maintained null hypothesis. This assumption will be most plausible for relatively small windows $\mathcal{W}_\tau$. The choice of the window $\mathcal{W}_\tau$ is therefore crucial; we discuss it in Section 3.4. Finally, Assumption 1(iii) sets the assignment mechanism to the TR mechanism.

We consider the null hypothesis that the average treated potential outcome is equal to the average untreated potential outcome in the τ-length postintervention period for every unit. Formally,

$$H_{TR} : \bar{y}_{i,\tau,+}(0, a_0) = \bar{y}_{i,\tau,+}(1, a_0) \quad \text{for all } i = 1, 2, \ldots, n.$$

Under Assumption 1, this hypothesis is sharp, and the average potential outcomes can be imputed for all realizations of the matrix $\mathbf{D}$. It follows that for any test statistic based on the observed average outcomes, $\mathbf{D}$ will be the only source of randomness.

Following standard randomization inference ideas, we define a test statistic $S(\mathbf{D}, \bar{\mathbf{Y}}_0, \bar{\mathbf{Y}}_1)$, where $\bar{\mathbf{Y}}_s = (\bar{Y}_{1,\tau,s}, \ldots, \bar{Y}_{n,\tau,s})$ for $s = 0, 1$ are the vectors that collect the observed average untreated and treated outcomes. Under Assumption 1 and H_{TR}, $S(\mathbf{D}, \bar{\mathbf{Y}}_0, \bar{\mathbf{Y}}_1) = S(\mathbf{D}, \bar{\mathbf{y}}(a_0))$, where $\bar{\mathbf{y}}(a_0) = (\bar{y}_{1,\tau,-}(0, a_0), \ldots, \bar{y}_{n,\tau,-}(0, a_0), \bar{y}_{1,\tau,+}(1, a_0), \ldots, \bar{y}_{n,\tau,+}(1, a_0))'$,   and   therefore   the

distribution of $S(\mathbf{D}, \bar{\mathbf{Y}}_0, \bar{\mathbf{Y}}_1)$ is fully determined by the randomization distribution of $\mathbf{D}$. The test statistic S is a function of the average observed outcomes because Assumption 1 places restrictions on the average potential outcomes, which is a simple way to think about (local to $t = a_0$) time trends. However, one could change the assumption to make it on the individual (nonaverage) outcomes, in which case the test statistic could be more general.

We denote the observed, realized value of the matrix $\mathbf{D}$ as $\mathbf{d}_{\mathrm{obs}}$, and the observed value of test statistic as $s_{\mathrm{obs}} = S(\mathbf{d}_{\mathrm{obs}}, \bar{\mathbf{Y}}_0, \bar{\mathbf{Y}}_1)$. The exact two-sided p-value is given by

$$p_{\mathrm{TR}} = \mathbb{P}\big[|S(\mathbf{D}, \bar{\mathbf{Y}}_0, \bar{\mathbf{Y}}_1)| \geq |s_{\mathrm{obs}}|\big] = \frac{1}{2^n} \sum_{\mathbf{d} \in \mathcal{D}_{TR}} \mathbb{1}(|S(\mathbf{d}, \bar{\mathbf{y}}(a_0))| \geq |s_{\mathrm{obs}}|),$$

which can be calculated for every realization of $\mathbf{D}$ based on observed data. For even moderate values of n, the total number of possible treatment assignment vectors is too large, so complete enumeration is typically unfeasible. In our application, $n = 62$, and hence we compute the p-value by simulation. In each simulation, we construct one of the possible treatment assignment vectors using the postulated assignment mechanism, and computing the value of the test statistic. We repeat this simulation 10,000 times, and then calculate the share of the 10,000 test statistics that have absolute value equal to or higher than $|s_{\mathrm{obs}}|$.

3.2 Adoption timing (AT) assignment mechanism

The TR assignment mechanism allows the treated period to come before the untreated period, a feature that does not respect the chronological ordering of the policy in the real assignment. We consider an alternative mechanism, the Adoption Timing (AT) assignment mechanism, where we retain the time ordering of the event study, and instead hypothesize that the adoption time could have happened sooner or later than a_0 for each unit. The AT assignment mechanism is based on a random time of adoption, which is represented by a discrete uniform random variable A_i with support $\mathcal{A} \subseteq \mathcal{W}_\tau$, allowing for up to $\tau - 1$ back-dating periods and up to τ forward-dating periods from the actual adoption time $t = a_0$. For example, $\mathcal{A} = \{-1, 0, 1\}$ would allow backdating one period, staying at the actual adoption time, or forward-dating one period, while $\mathcal{A} = \{-2, -1\}$ would only allow back-dating up to two periods. Letting J denote the number of periods in $\mathcal{A}$, we have $\mathbb{P}[A_i = a] = \frac{1}{J}$ for all $a \in \mathcal{A}$. Allowing the generality of back-dating and forward-dating may be important in some applications when anticipation or delayed effects are a concern.

The treatment indicator for each unit and time period is

$$D_{i,t}(A_i) = \mathbb{1}(t \geq A_i),$$

which respects the temporal sequence of the real assignment: given an adoption time a, a treated period always occurs for $t \geq a$, while an untreated period always occurs for $t < a$. This mechanism imposes different restrictions on the matrix $\mathbf{D}$. For each unit i, there are now J possible adoption times $a \in \mathcal{A}$ where $D_{i,t} = 0$ for $t < a$ and $D_{i,t} = 1$ for $t \geq a$. Since each row of $\mathbf{D}$ corresponds to the assignment of each unit, and each unit is assigned an adoption time independently, there are J^n possible values of $\mathbf{D}$. Denoting by $\mathcal{D}_{\mathrm{AT}}$ the set of all allowed assignment matrices, we have $\mathbb{P}[\mathbf{D} = \mathbf{d}] = J^{-n}$ for $\mathbf{d} \in \mathcal{D}_{\mathrm{AT}}$.

The conditions to apply randomization inference under this mechanism are summarized in the following assumption.

> **Assumption 2**   ((AT Assignment Mechanism)). There exists a window $\mathcal{W}_\tau$ of τ periods before and after the time of the intervention such that the following conditions hold for all $i = 1, 2, \ldots, n$.
>
> (i) $\bar{y}_{i,\tau,-}(0, a)$, $\bar{y}_{i,\tau,-}(1, a)$, $\bar{y}_{i,\tau,+}(0, a)$ and $\bar{y}_{i,\tau,+}(1, a)$ are nonstochastic, for all $a \in \mathcal{A}$.
> (ii) $\mathbb{P}[\mathbf{D} = \mathbf{d}] = J^{-n}$ for all $\mathbf{d} \in \mathcal{D}_{\mathrm{AT}}$.

Assumption 2(i) is analogous to Assumption 1(i), while Assumption 2(ii) is analogous to Assumption 1(iii). Under this assignment mechanism, a no-trends assumption is not needed.

We consider the following null hypothesis,

$$H_{\text{AT}}: \sum_{t \in \mathcal{W}_\tau} \mathbb{1}(a \leq t < a_0)[y_{i,t}(0) - y_{i,t}(1)] = 0,$$

$$\sum_{t \in \mathcal{W}_\tau} \mathbb{1}(a_0 \leq t < a)[y_{i,t}(0) - y_{i,t}(1)] = 0,$$

$$\text{for all } i = 1, 2, \ldots, n \text{ and } a \in \mathcal{A}.$$

In contrast to the null hypothesis under the TR assignment mechanism, H_{AT} is stated for the sum of individual potential outcomes for periods in $\mathcal{A}$ rather than average outcomes. This is necessary to impute the average potential outcomes $\bar{y}_{i,\tau,+}(1, a)$ and $\bar{y}_{i,\tau,+}(0, a)$ for all possible values of $a \in \mathcal{A}$. To see this, note that the average observed treated outcome is $\bar{Y}_{i,\tau,1} = \frac{1}{N_{i,\tau,1}} \sum_{t \in \mathcal{W}_\tau} D_{it} Y_{it} = \frac{1}{\sum_{t \in \mathcal{W}_\tau} \mathbb{1}(t \geq a_0)} \sum_{t \in \mathcal{W}_\tau} \mathbb{1}(t \geq a_0)y_{i,t}(1) = \bar{y}_{i,\tau,+}(1, a_0)$. Whenever the assignment mechanism moves the adoption time from a_0 to another value $a > a_0$, the observations for periods $a_0 \leq t < a$ must be assigned to the control group, even though they are treated according to the actual mechanism. The second condition in H_{AT} allows for the imputation of the untreated average potential outcome $\bar{y}_{i,\tau,+}(0, a)$ in this case. A similar argument shows that the first condition allows the imputation of $\bar{y}_{i,\tau,+}(1, a)$ when $a < a_0$. If the set $\mathcal{A}$ only includes back-dating (forward-dating), only the first (second) condition in H_{AT} will be applicable.

If $J = 1$, H_{AT} is equivalent to $y_{i,a_0-1}(0) - y_{i,a_0-1}(1) = 0$ and $y_{i,a_0}(0) - y_{i,a_0}(1) = 0$ for all $i = 1, 2, \ldots, n$ and $a \in \mathcal{A}$, which is a sharp null hypothesis for individual potential outcomes. For $J > 1$, H_{AT} is weaker than assuming no treatment effects for every unit in every period. The number of forward-dating periods included in $\mathcal{A}$ can be as large as τ, while the number of back-dating periods can be as large as $\tau - 1$ to ensure that at least one period is untreated for every unit. The choice of J can also affect the power of the testing procedure.

Under Assumption 2 and H_{AT}, once again for any test statistic $S(\mathbf{D}, \bar{\mathbf{Y}}_0, \bar{\mathbf{Y}}_1)$ we have $S(\mathbf{D}, \bar{\mathbf{Y}}_0, \bar{\mathbf{Y}}_1) = S(\mathbf{D}, \bar{\mathbf{y}}(a_0))$, and therefore its distribution is fully determined by the randomization distribution of $\mathbf{D}$. The randomization-based p-value can be obtained analogously to p_{TR} as

$$p_{\text{AT}} = \mathbb{P}\big[|S(\mathbf{D}, \bar{\mathbf{Y}}_0, \bar{\mathbf{Y}}_1)| \geq |s_{\text{obs}}|\big] = \frac{1}{J^n} \sum_{\mathbf{d} \in \mathcal{D}_{\text{AT}}} \mathbb{1}(|S(\mathbf{d}, \bar{\mathbf{y}}(a_0))| \geq |s_{\text{obs}}|),$$

3.3 Other inference procedures

The Fisherian framework can also be used to derive randomization-based confidence intervals by inversion of hypothesis tests under a model of treatment effects. We follow Rosenbaum (2002) and assume a constant treatment effect model. In the TR assignment mechanism, this modifies our null hypothesis to $\tilde{H}_{\text{TR}}: y_{i,\tau,+}(1) = y_{i,\tau,+}(0) + \theta_0$ for all $i = 1, 2, \ldots, n$. Confidence intervals of level $1 - \alpha$ are obtained by collecting all the values of θ_0 that are not rejected by a randomization-based test of $\tilde{H}_{\text{TR}}$ at level α. The calculation of our Fisherian p-value is the same as before, except that we use the adjusted observed outcomes instead of the raw observed outcomes, where the adjustment is conducted using the constant treatment effect model, i.e. $Y_{i,\tau,k,\text{adj}} = Y_{i,\tau,k} - \theta_0 D_{i,t}$ for $k = 0, 1$. Similar ideas can be used for inversion of H_{AT}.

Our approach can also be generalized to a joint test of multiple hypotheses, which allows us to test hypotheses in many windows simultaneously around adoption time, rather than performing one test per window. The null hypothesis can be made for L different windows $\mathcal{W}_{\tau_l}$, $l = 1, 2, \ldots, L$, under either the TR or AT mechanisms. Then, a test statistic can be calculated for each hypothesis in each window following the steps above. Letting s_l be the test statistic associated with window τ_l, we can combine the statistics for L windows, $s_1, s_2, \ldots, s_L$ into a single joint statistic such as the mean, maximum, or Hotelling's T^2, and calculate a p-value by repeating the procedure for different realizations of the treatment vector. This procedure can be used to test the hypothesis that the effect is jointly zero in all windows $\mathcal{W}_1, \mathcal{W}_2, \ldots, \mathcal{W}_L$, as we illustrate in the following section.

3.4 Window selection

Our hypothesized randomization mechanisms assume that the treatment assignment could have been different in a small window around adoption time. Moreover, the TR mechanism assumes that, within this window, there is no difference in average potential outcomes between the preadoption and postadoption periods. The implementation of our approach thus requires the careful choice of τ, the length of the window around adoption time.

Our approach allows us to reduce extrapolation to a minimum by considering the smallest window $\mathcal{W}_\tau$ around adoption time. By choosing $\tau = 1$, researchers can compare the average outcome in the period immediately before adoption of the intervention to the average outcome in the exact period when the intervention goes into effect. Although this window minimizes extrapolation as much as it is possible given the data, it may be too small for causal effects to appear.

For windows $\mathcal{W}_\tau$ larger than the minimum, two types of strategies can guide the choice of τ. The first strategy is qualitative, and it involves considering the mechanisms by which the intervention is likely to affect the outcome, and the period of time that is plausibly required for those mechanisms to become active and cause visible changes. In our application, it is in principle possible for the change in incentives brought about by the new CCP to cause individuals to change their criminal behaviour the day of adoption. Thus, a window of one day before to one day after the intervention is reasonable in our context to capture immediate effects. However, in other applications, this window length might not be practical. For example, for scientists who wish to consider the effect of emission policies on climate change, a window of one day before and after the policy will be uninformative, as we know that the mechanisms by which reducing emissions affect climate take years to mature.

The second strategy is quantitative, which we develop following the method by Cattaneo et al. (2015) for window selection in local randomization RD designs. In the original method, preintervention covariates (for which the treatment effect is known to be zero) are used to choose the window, sequentially decreasing its size until covariates are balanced and the null hypothesis of no treatment effect fails to be rejected.

In our context, all preintervention covariates are balanced exactly in any window around adoption time, because we observe the same exact units before and after the intervention. Thus, a window selector procedure cannot be based on preintervention covariates. Instead of covariates, we propose a method based on the time-varying outcome before the intervention occurs. Because all units are untreated before the intervention, the null hypothesis of no treatment effect should fail to be rejected unless there are time trends that cause an incorrect rejection. The suggested procedure is therefore as follows: (i) pick a date before the actual intervention to be used as an artificial adoption time; (ii) choose a sequence of window lengths, for example from 1 until the end of the pretreatment period, and obtain the randomization-based p-value for each window in this sequence; and (iii) select the window length such that the p-value is (approximately) larger than 0.15 in that window and in all windows of smaller length.

The procedure will work under the assumption that the time trends that characterize the outcome in the artificial windows are similar to the time trends that characterize the window around the real intervention. In the absence of this assumption, any balance in the outcome in artificial windows will be uninformative regarding the plausibility of the no trends assumption. This is similar to the role of covariate balance tests in randomized controlled experiments, where observing balance in covariates is taken as indirect evidence towards the plausibility of the independence between the treatment assignment and the potential outcomes. We illustrate how to implement this procedure in Section 4.

As in the RD case, the selection of the window based on the preintervention outcome is in itself a type of falsification analysis: if the window length is chosen following the outlined procedure and researchers assume that time trends are similar within the artificial and the real windows, the window selector suggests values of τ where the assumption of no trends might be plausible. In addition to the window selection procedure, our framework allows us to design other falsification tests based on prior periods. In our application, the intervention occurred on 1 November 2017, but we have data on our outcome variable from several years prior. For our chosen window lengths using the window selection procedure, we consider additional falsification analyses that study the effect of the intervention in the years before and after the true intervention date, setting the cutoff to artificial values. We discuss this procedure in Section 4.

3.5 Choice of test statistic

The Fisherian framework provides exact inferences for any suitable test statistic whose distribution is known under the null. We use the average difference between the (average) outcome in the treated postadoption period and the (average) outcome in the untreated preadoption period, $S(\mathbf{D}, \bar{\mathbf{Y}}_0, \bar{\mathbf{Y}}_1) = \frac{1}{n}\sum_{i=1}^{n} \bar{Y}_{i,\tau,1} - \frac{1}{n}\sum_{i=1}^{n} \bar{Y}_{i,\tau,0} \equiv \widehat{\theta}_\tau$, which is justified by our assumptions and null hypotheses.

However, the framework is general and allows for the use of other test statistics. In particular, a test statistic that adjusts for time trends could be useful in our context, as it could allow researchers to analyze larger windows by using the adjusted outcome instead of the raw outcome—see, e.g. Cattaneo et al. (2017). For this, researchers would need to modify Assumptions 1 and 2 to refer to the adjusted outcomes. In Section S1 in the online supplementary material, we present results using a time-adjusted statistic.

4 Results

The new adversarial system came into effect nationwide at 12 a.m. on 1 November 2017. The cornerstone of our research design is hypothesizing different assignment mechanisms for the reform affecting each of Montevideo's 62 neighbourhoods in a small window around adoption time. We include in our analysis the smallest window, $\mathcal{W}_1$, which includes just 2 days and compares the offences that occurred on 31 October 2017, to the offences that occurred on 1 November 2017. The smallest possible window allows researchers to investigate whether treatment effects appear almost instantaneously. If the intervention requires more than one unit of time to affect the outcome, the effects in the smallest window will be null.

To guide the choice of additional windows, we implement the window selection procedure outlined in Section 3. We choose an artificial or 'placebo' adoption time before the actual intervention, and use the TR mechanism to test the null hypothesis H_{TR} that the placebo intervention had no effect on total crime for many windows around this artificial adoption time. Our placebo adoption time is the first Wednesday in October, 2017, about 1 month prior to the adoption of the real intervention, which occurred on the first Wednesday in November.

Figure 2 plots the randomization-based p-value against the half length τ of each window. We add a horizontal line at 0.15, following the recommendation of Cattaneo et al. (2015); since we are concerned with incorrectly failing to reject the null hypothesis (rather than incorrectly rejecting it), it is advisable to consider windows where failure to reject occurs for significance levels higher than conventional ones. The pattern shows that in all windows within 14 days of the selected placebo cutoff, the p-value associated with the null hypothesis H_{TR} for total crime is 0.15 or higher, validating choices of $\tau \leq 14$. For larger windows, the p-value decreases rapidly—it is higher than 0.30 for $\tau = 14$, less than 0.06 for $\tau = 15$, and less than 0.01 for $\tau = 21$. In other words, after day 15, the trend in the outcome is such that our randomization-based approach detects an effect even when there is none. This reinforces that our procedure is best suited for evaluating short-term effects.

In Section S2 of the online supplementary material, we repeat the exercise for two additional placebo adoption times (21 and 35 days before the actual adoption, $a = -21, -35$); the overall pattern is roughly similar, although some p-values fall below 0.15 for smaller τ, particularly for $a = -21$.

Table 2 reports the results for our main outcome of interest, total crime reports, now using the real adoption time. The analysis is performed in three windows in the set of windows validated above: 1 day, 7 days, and 14 days before/after the adoption of the new code, that is, $\mathcal{W}_\tau$ for $\tau = 1, 7, 14$. For the implementation of the AT mechanism, we consider backdating up to $\tau - 1$ and include the actual adoption time: we set $\mathcal{A} = \{a_0 - 6, a_0 - 5, \ldots, a_0 - 1, a_0\}$ for $\tau = 7$, and $\mathcal{A} = \{a_0 - 13, a_0 - 12, \ldots, a_0 - 1, a_0\}$ for $\tau = 14$. The table reports the estimated average difference, $\widehat{\theta}_\tau$, and the randomization-based p-value corresponding to the two-sided test of the null hypotheses H_{TR} and H_{AT}, calculated using the TR and AT assignment mechanisms, respectively.

The first row reports the results in the smallest possible window, $\mathcal{W}_1$. On the day before adoption, the average number of crime reports per neighbourhood was 5.306; this average increased by 0.839 to 6.145 the day after the intervention, but the test of H_{TR} based on the TR assignment mechanism does not reject the null at conventional levels (p-value is 0.131, 95% confidence
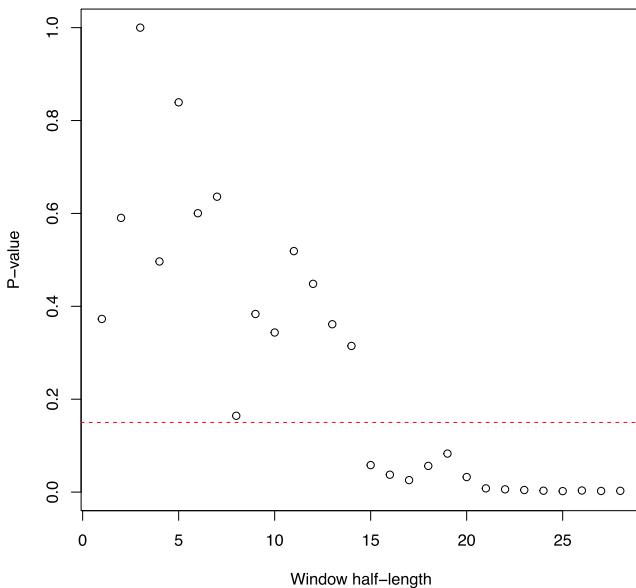
Figure 2. Window Selector Around *Placebo* Adoption Times *Placebo adoption times*: Up to 28 days prior to 1 November 2017 *Outcome*: Daily number of crimes reported to police.

interval ranges from -0.23 through 1.9). Note that the AT mechanism cannot be used in $\mathcal{W}_1$ because it is not possible to vary the adoption time and still keep observations on both sides.

The second row shows that, when comparing the 7 days before to the 7 days after the intervention, the reform seems to have resulted in a statistically significant increase in the total number of crimes reported when using the TR mechanism (null hypothesis rejected with p-value 0.013, 95% confidence interval of [0.09, 0.72] under a constant treatment effect model). The average number of daily crimes reported per neighbourhood during the 7 days preceding the intervention was 4.892, increasing to 5.295 in the week following the intervention—an increase of 0.403 police reports per day per neighbourhood. Since there are 62 neighbourhoods, this corresponds to an increase from approximately $4.892 \times 62 \times 7 \approx 2{,}123$ reports in the 7 days before the intervention to $(4.892 + 0.403) \times 62 \times 7 \approx 2{,}298$ after—an increase of about 8.2 percent. The pattern is similar for a window within 14 days of the start of the intervention—TR null H_{TR} rejected with p-value 0.020, confidence interval [0.05, 0.55]. Using the AT mechanism, H_{AT} is not rejected in $\mathcal{W}_7$ but is rejected in $\mathcal{W}_{14}$.

We also investigate the treatment effects in many windows simultaneously. Table 3 shows the results of joint randomization-based tests for all windows between day 1 and day 7, $\mathcal{W}_1, \ldots, \mathcal{W}_7$, and all windows between day 1 and day 14, $\mathcal{W}_1, \ldots, \mathcal{W}_{14}$. The p-values are calculated using three test statistics: the maximum of the average difference in each window, Hotelling's T^2 statistic using the vector of average differences in each window with their respective covariance matrices, and the mean of the average difference in each window. When jointly considering the treatment effect in the first seven or the first fourteen windows under the TR mechanism, we reject H_{TR} at 10% or lower level in all cases but one. For example, according to the maximum average difference across all windows, the hypothesis of no effect has p-value 0.064 for the first seven windows, and according to Hotelling's T^2, this p-value is 0.074. The p-values in the first 14 windows are smaller. The pattern is similar with the AT mechanism. Together, these joint hypothesis tests suggest a significant effect within 7 to 14 days after adoption of the new CCP.

4.1 Assessing assumptions

We present additional results from a falsification analysis that analyzes our main outcome at artificial adoption times. Our approach is similar to the strategy of estimating effects at placebo cutoffs in the RD design (see, e.g. Cattaneo et al., 2020, Section 5), where the presence of significant

Table 2. Short-term effects of CCP reform in Montevideo *outcome*: daily number of crimes reported to police *actual adoption time*: 1 November 2017

	Estimates		TR Mechanism		AT Mechanism	
τ	$\widehat{\theta}_\tau$	$\frac{1}{n}\sum_{i=1}^{n} \bar{Y}_{i,\tau,0}$	*p*-value	95% CI	*p*-value	95% CI
1 day	0.839	5.306	0.131	[−0.23, 1.90]	–	–
7 days	0.403	4.892	0.013	[0.09, 0.72]	0.430	[−0.69, 0.87]
14 days	0.298	5.065	0.020	[0.05, 0.55]	0.021	[0.04, 1.81]

Note. Sample is 62 neighbourhoods in Montevideo, each observed before and after the adoption of the CCP reform. The *p*-values are randomization-based for $\theta_\tau = \frac{1}{n}\sum_{i=1}^{n} \bar{Y}_{i,\tau,1} - \frac{1}{n}\sum_{i=1}^{n} \bar{Y}_{i,\tau,0}$ with 10,000 simulations. τ denotes the half-length of a symmetric window around adoption time. CI refers to confidence interval, calculated by inversion based on a constant treatment effect model. TR and AT refer to Treatment Reversal and Adoption Timing mechanisms, respectively. In the AT mechanism, $\mathcal{A} = \{-6, -5, \ldots, 0\}$ for $\tau = 7$ and $\mathcal{A} = \{-13, -12, \ldots, 0\}$ for $\tau = 14$.

Table 3. Joint inference for short-term effects of CCP reform in Montevideo *outcome*: daily number of crimes reported to police *actual adoption time*: 1 November 2017

TR Assignment Mechanism

Joint test of hypotheses H_{TR}, for $\tau = 1, \ldots, K$

	Max		Hotelling		Mean	
	Statistic	*p*-value	Statistic	*p*-value	Statistic	*p*-value
$\tau = 1, \ldots, 7$ ($K = 7$)	0.839	0.064	12.995	0.074	0.292	0.172
$\tau = 1, \ldots, 14$ ($K = 14$)	0.839	0.064	37.748	0.004	0.302	0.059

AT Assignment Mechanism

Joint test of hypotheses H_{AT}, for $\tau = 2, \ldots, K$

	Max		Hotelling		Mean	
	Statistic	*p*-value	Statistic	*p*-value	Statistic	*p*-value
$\tau = 2, \ldots, 7$ ($K = 7$)	0.403	0.124	6.893	0.105	0.283	0.233
$\tau = 2, \ldots, 14$ ($K = 14$)	0.403	0.015	26.375	0.003	0.327	0.021

Note. Sample is 62 neighbourhoods in Montevideo, each observed before and after the adoption of the CCP reform. The *p*-values are randomization-based for $\widehat{\theta}_\tau = \frac{1}{n}\sum_{i=1}^{n} \bar{Y}_{i,\tau,1} - \frac{1}{n}\sum_{i=1}^{n} \bar{Y}_{i,\tau,0}$ with 10,000 simulations. τ denotes the half-length of a symmetric window around adoption time. TR and AT refer to Treatment Reversal and Adoption Timing mechanisms, respectively. In the AT mechanism, $\mathcal{A} = \{-K + 1, -K + 2, \ldots, 0\}$.

effects at artificial cutoffs is interpreted as potentially casting doubt on the main identifying assumptions.

We intend these analyses as an illustration of how to provide empirical support for the main assumptions, not as a formal test—the assumptions are fundamentally untestable, and the lack of effects at artificial adoption times is neither necessary nor sufficient for the assumptions to hold. Nonetheless, observing no effects at artificial adoption times can sometimes rule out some violations. For example, if another factor that affects crime other than the new CCP changes discontinuously between $t = a_0 - 1$ and $t = a_0$, the causal interpretation of the results would be incorrect. In our study, this could happen, for example, if $t = a_0 - 1$ fell on a Sunday and $t = a_0$ on a Monday, and crime were higher during the weekend (see Prieto Curiel, 2023, for evidence and discussion regarding the temporal concentration of crime). Observing no significant day-of-the-week effects in the absence of the intervention might dissipate fears about this potential violation.

Our falsification analyses are an example of negative control analyses or a placebo population test, where an experiment is repeated under conditions in which null results are expected and the researchers verify that a null result is indeed observed (Lipsitch et al., 2010). In general, for this type of analysis to be informative, researchers make the assumption that failure to obtain a null result implies a violation of the main assumptions invoked.

Table 4 presents results with the adoption time artificially set to midnight on 1 November for the years 2015, 2016, and 2018. The outcome, windows, and assignment mechanisms are identical to those used in Table 2; the only change is the use of artificial adoption times. The results show that there are no distinguishable differences in crime reports around artificial adoption times in the same windows considered in Table 2: the null hypothesis fails to be rejected in all cases except for $\tau = 1$ in 2016 under the TR mechanism, but this corresponds to a negative value of test statistic (-0.855).

In Section S3 in the online supplementary material, we show that the falsification results in Table 4 remain robust when we consider alternative values of τ, and also when we set the artificial adoption time to the same day of the week as the actual adoption (first Wednesday in November) rather than to the same date. We also report results for additional years.

In sum, the rejections that we see in small windows around the actual adoption time are not seen when we consider windows of the same length around the same date or time in prior years. Moreover, because our analysis is based on the exact same units before and after adoption in all falsification and actual analyses, the results in Table 2 cannot be explained by any time-constant unit-specific characteristics or 'fixed effects'.

4.2 Potential mechanisms

As shown in Table 2, the new CCP led to an average rise of 25 police reports per day ($\approx 0.403 \times 62$) in Montevideo during the week following adoption. The CCP reform had multiple components. The new system made plea bargaining easier, reduced pretrial detention, and affected coordination between the police and prosecutors, all of which could have plausibly affected crime rates through decreases in deterrence and/or incapacitation. Moreover, the new system could have affected whether and how individuals report crime, how the police patrol neighbourhoods, and how crime data is collected.

We start by considering the possibility that the effects we see reflect a change in how crime was reported rather than an actual deterioration in public safety. Although we cannot rule this out entirely, we have found no compelling evidence of a change in crime reporting practices. First, the reform did not alter how crime data were collected, managed, or disseminated; all processes continued to be overseen by the Ministry of the Interior through its Public Security Management System (SGSP). As part of the implementation of the new CCP, SGSP data began to be shared in real time with the Office of the Attorney General through the Uruguayan Adversarial Criminal Procedure Information System (SIPPAU). The connection between SGSP and SIPPAU reflected prosecutors' expanded role in the investigative and judicial stages of criminal proceedings under the new adversarial system. We have seen no evidence that this data transfer undermined the credibility of the crime statistics produced by the Ministry of the Interior.

Moreover, institutional trust trends do not support a sharp increase in the willingness to report crimes. According to Latinobarómetro (2016, 2018), trust in the police remained stable at elevated levels: around 60% of respondents reported 'some' or 'a lot' of trust in both 2016 and 2018. This is the highest level observed in Latin America, where the 2018 regional average stood at just 35%. In contrast, trust in the judiciary declined by 8 percentage points, from 47% in 2016 to 39% in 2018, possibly reflecting dissatisfaction with procedural changes, such as limits on pretrial detention and the rollout of alternative procedures to oral trial. If anything, this decline could have discouraged reporting, suggesting our estimates may understate the reform's true impact. Moreover, in Section S4 of the online supplementary material, we show that the increase in crime reports does not appear to have been driven by enhanced street-level policing.

We believe our estimates reflect a rise in actual crime driven the various components of the bundle of interventions introduced by the new CCP. The reform may have led to less severe sanctions and a lower probability of conviction, factors that have been shown to increase crime. First, the new system seems to have resulted in less severe penalties, as it introduced procedural alternatives

Table 4. Short-term effects of CCP reform in Montevideo around *placebo* adoption times (date) *outcome*: daily number of crimes reported to police *actual adoption time*: 1 November 2017

| | 2015 | | | 2016 | | | 2018 | | |
| | | *p*-value | | | *p*-value | | | *p*-value | |
τ	$\widehat{\theta}_\tau$	TR	AT	$\widehat{\theta}_\tau$	TR	AT	$\widehat{\theta}_\tau$	TR	AT
1 day	0.048	0.934	–	−0.871	0.036	–	0.661	0.170	–
7 days	0.090	0.544	0.879	−0.041	0.806	0.761	0.002	1.000	0.994
14 days	0.015	0.914	0.878	0.035	0.761	0.921	−0.113	0.401	0.468

Note. Sample is 62 neighbourhoods in Montevideo, each observed before and after the adoption of the CCP reform. The *p*-values are randomization-based for $\widehat{\theta}_\tau = \frac{1}{n}\sum_{i=1}^{n} \bar{Y}_{i,\tau,1} - \frac{1}{n}\sum_{i=1}^{n} \bar{Y}_{i,\tau,0}$ with 10,000 simulations. τ denotes the half-length of a symmetric window around adoption time. TR and AT refer to Treatment Reversal and Adoption Timing mechanisms, respectively. In AT mechanism, $\mathcal{A} = \{-6, -5, \ldots, 0\}$ for $\tau = 7$ and $\mathcal{A} = \{-13, -12, \ldots, 0\}$ for $\tau = 14$.
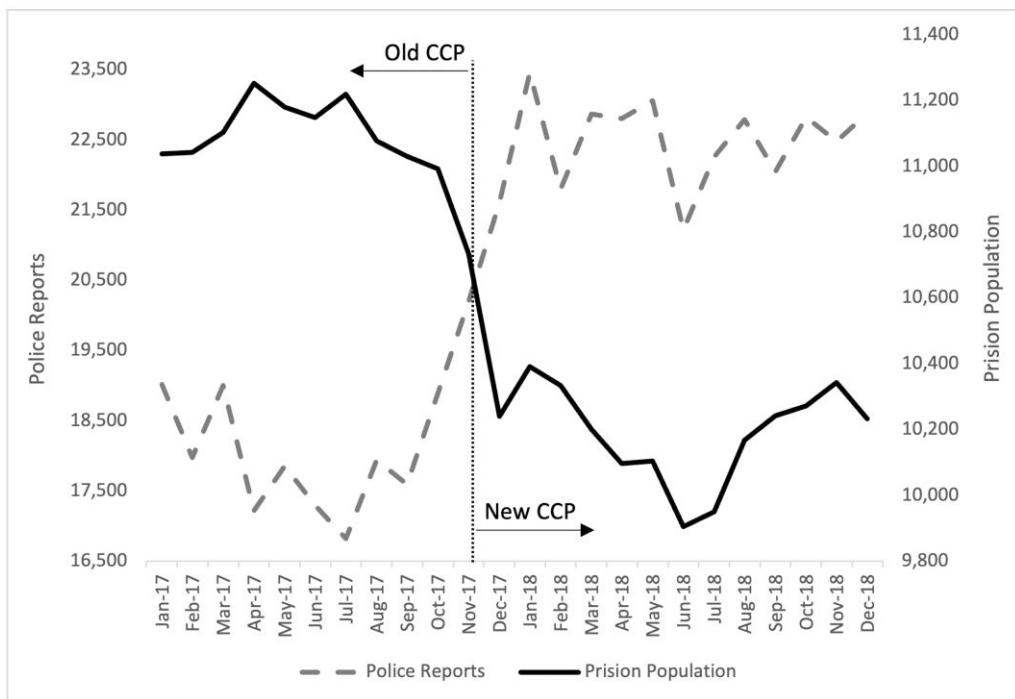
to oral trials that are associated with lighter sentences. These alternatives, which were employed in 90% of solved cases, help lawyers and defendants resolve their cases faster while allowing public prosecutors to avoid lengthy and demanding criminal trials. Most of these solved cases are the result of an abbreviated process (i.e. a type of plea bargaining introduced by the new CCP) that implies an agreement between the defendant and the prosecutor whereby the former pleads guilty to a particular charge in return for a more lenient sentence. Moreover, preventive detention (i.e. detention while the process lasts until there is a sentence) ceased to be the norm, in contrast to the old system, which might have created the expectation of less severe punishment. Second, the reform might have resulted in a lower probability of conviction. Under the new system, prosecutors faced a significant increase in their workload, as they were now exclusively responsible for leading the investigation and carrying the evidence to judges. Moreover, the police must conduct investigations under new supervisors (prosecutors instead of judges) and different rules, which could have created coordination challenges.

Figure 3 provides some evidence supporting the hypothesis that the reform affected crime incentives. For the first time in more than a decade, the average number of people in prisons decreased in 2018, from 11,005 to 10,179 inmates (a 7.5% decrease). According to Figure 3a, this sharp reduction in prison population starts at the same time as the implementation of the new CCP and contrasts with the sudden rise in the number of crimes reported to police that we documented in the prior section. Meanwhile, the number of criminal indictments (i.e. formal accusations made by public prosecutors) also experienced a strong month-to-month decrease in November 2017, from 1,001 to 584 cases (i.e. a 42% reduction). Figure 3b illustrates the evolution of indictments relative to police reports. The average ratio for the first 2 months of the new CCP is 3.2% (i.e. November and December 2017), well below its value of 5.9% during the rest of 2017 in the last months of the old CCP.

The above trends suggest that less severe crimes could have increased due to a change in crime incentives (i.e. less deterrence) and also to a rise in the number of active criminals (i.e. less incapacitation). The reduction in prison population might be explained by the new constraints imposed on the use of preventive detention by prosecutors. As mentioned above, preventive prison was used extensively under the old system, in particular for cases of recidivism (55% of the cases). This was the norm for both property crimes (e.g. thefts) and violent crimes (e.g. robberies and domestic violence). However, under the new system, preventive detention is applicable only when there is sufficient evidence that the defendant might attempt to escape, obstruct the investigation, or pose a risk to society (i.e. severe and typically violent crimes). Consequently, a convicted offender who re-offends by committing a lesser crime could be immediately released under the new CCP, whereas under the old CCP, they would likely have served preventive detention.

If the increase in crime reports is due to a more selective use of preventive detention, we should not observe any impact on domestic violence (a crime considered very severe by prosecutors under both procedural regimes), while the effects should be present for thefts, as they are nonviolent. In

(a)

(b)
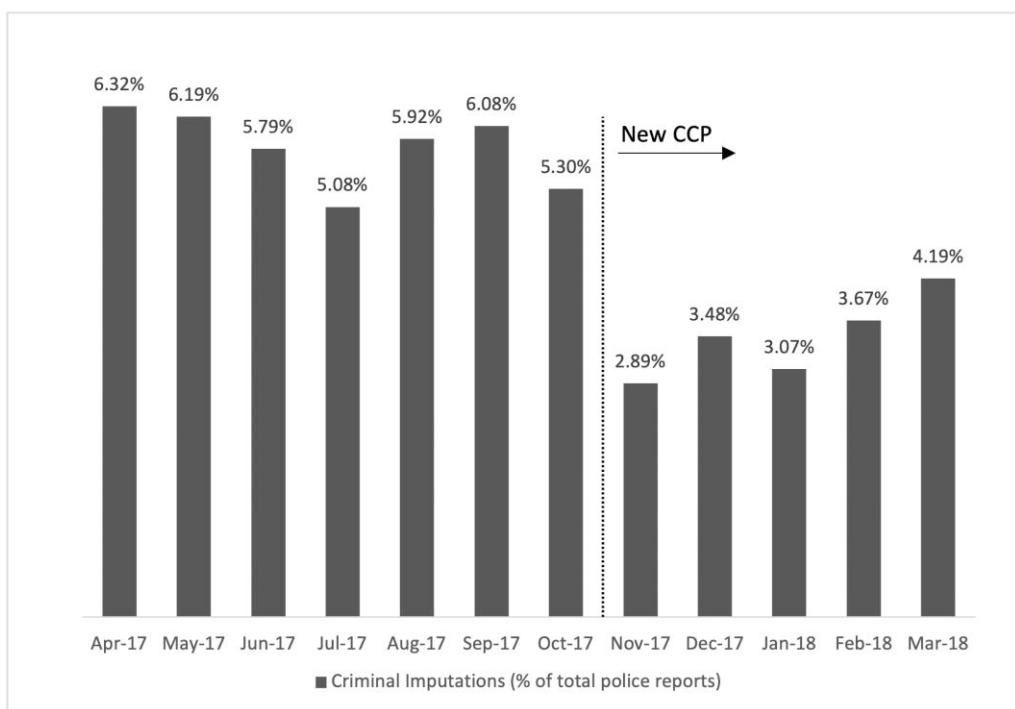
Figure 3. Police Reports, Prison Population and Criminal Imputations. (a) Police Reports and Prison Population. (b) Criminal Imputations.

Section S4 of the online supplementary material, we report disaggregated effects by type of offence and show that the overall effects are not driven by the the two most frequent violent crimes (domestic violence and robberies).

5 Conclusion

We developed a randomization-based framework for short-term causal inference, designed to analyze before-and-after event studies where an intervention is given to all cross-sectional units at the same time, and all units are observed for several periods before and after the intervention. Our set-up assumes that the units' potential outcomes are nonstochastic, and considers two possible assignment mechanisms governing how units receive the treatment in a small window around the time of adoption of the intervention. We used this framework to study the effects of a reform to the code of criminal procedure introduced in Uruguay on 1 November 2017, which was simultaneously implemented in the entire country. The reform increased total crime reports in Montevideo by about 8.2 percent in the week immediately after adoption. Our randomization-based framework rejects hypotheses of no effect in small windows around adoption time, providing evidence that the reform increased crime reports. Moreover, we see no evidence of such increase when the analysis is conducted in prior years, before the intervention is active.

Our framework is tailored to study immediate or short-term effects, which are essential to the practical implementation of institutional reforms. Our case study illustrates how a shift in the criminal adjudication system can trigger immediate and unintended consequences that materially shape the reform's long-term trajectory. While the full impact of reforms can only be assessed over a medium- to long-term horizon, early signals play a critical role in shaping public opinion and securing political support. From a public policy perspective, our findings underscore the importance of embedding institutional change within a framework of real-time empirical monitoring. Our proposed framework was designed specifically for short-term causal inference, facilitating evidence-based monitoring of reform outcomes from the outset. By enabling policymakers to detect early shocks, diagnose implementation challenges, and adjust course in real time, this approach enhances the capacity to steer complex reforms toward long-term success.

Data availability

Data and replication files can be found at https://github.com/mdcattaneo/replication-CDT_2026_JRSSA.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series A*.

References

Acemoglu D., & Johnson S. (2005). Unbundling institutions. *Journal of Political Economy*, *113*(5), 949–995. https://doi.org/10.1086/432166

Acosta C., Mejia D., & Zorro Medina A. (2023). *On the tension between due process protection and public safety: The case of an extensive procedural reform in Colombia*. Unpublished manuscript.

Atkins R. A., & Rubin P. H. (2003). Effects of criminal procedure on crime rates: Mapping out the consequences of the exclusionary rule. *The Journal of Law and Economics*, *46*(1), 157–179. https://doi.org/10.1086/345582

Becker G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy*, *76*(2), 169–217. https://doi.org/10.1086/259394

Bushway S. D., & Redlich A. D. (2012). Is plea bargaining in the "shadow of the trial" a mirage? *Journal of Quantitative Criminology*, *28*(3), 437–454. https://doi.org/10.1007/s10940-011-9147-5

Cattaneo M. D., Frandsen B., & Titiunik R. (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the U.S. senate. *Journal of Causal Inference*, *3*(1), 1–24. https://doi.org/10.1515/jci-2013-0010

Cattaneo M. D., Idrobo N., & Titiunik R. (2020). A practical introduction to regression discontinuity designs: Foundations. In *Cambridge elements: Quantitative and computational methods for Social Science*. Cambridge University Press. https://doi.org/10.1017/9781108684606

Cattaneo M. D., Idrobo N., & Titiunik R. (2024). A practical introduction to regression discontinuity designs: Extensions. In *Cambridge elements: Quantitative and computational methods for Social Science*. Cambridge University Press. https://doi.org/10.1017/9781009441896

Cattaneo M. D., & Titiunik R. (2022). Regression discontinuity designs. *Annual Review of Economics*, *14*(1), 821–851. https://doi.org/10.1146/economics.2022.14.issue-1

Cattaneo M. D., Titiunik R., & Vazquez-Bare G. (2017). Comparing inference approaches for RD designs: A re-examination of the effect of head start on child mortality. *Journal of Policy Analysis and Management*, *36*(3), 643–681. https://doi.org/10.1002/pam.2017.36.issue-3

Cook T. D., Campbell D. T., & Shadish W. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Dalla Pellegrina L. (2008). Court delays and crime deterrence. *European Journal of Law and Economics*, *26*(3), 267–290. https://doi.org/10.1007/s10657-008-9076-4

Dušek L. (2015). Time to punishment: The effects of a shorter criminal procedure on crime rates. *International Review of Law and Economics*, *43*(1112), 134–147. https://doi.org/10.1016/j.irle.2014.04.007

Fandiño M., & González Postigo L. (2020). Adversarial criminal justice in Latin America: Comparative analysis and proposals. *Justice Studies Center of the Americas (JSCA)*.

Freyaldenhoven S., Hansen C., Pérez J. P., & Shapiro J. M. (2025). Visualization, identification, and estimation in the linear panel event-study design. In V. Chernozhukov, J. Hörner, E. L. Ferrara, & I. Werning (Eds.), *Advances in economics and econometrics: Twelfth world congress* (Vol. 2, Chapter 7). Cambridge University Press.

Freyaldenhoven S., Hansen C., & Shapiro J. M. (2019). Pre-event trends in the panel event-study design. *American Economic Review*, *109*(9), 3307–3338. https://doi.org/10.1257/aer.20180609

Hausman C., & Rapson D. S. (2018). Regression discontinuity in time: Considerations for empirical applications. *Annual Review of Resource Economics*, *10*(1), 533–552. https://doi.org/10.1146/resource.2018.10.issue-1

Ho D. E., & Imai K. (2006). Randomization inference with natural experiments: An analysis of ballot effects in the 2003 election. *Journal of the American Statistical Association*, *101*(475), 888–900. https://doi.org/10.1198/016214505000001258

Idrobo N., Kronick D., & Slough T. (2024). *The effect of oversight on the quantity and quality of policing*. Unpublished manuscript.

Imbens G. W., & Rosenbaum P. (2005). Robust, accurate confidence intervals with a weak instrument: Quarter of birth and education. *Journal of the Royal Statistical Society, Series A*, *168*(1), 109–126. https://doi.org/10.1111/j.1467-985X.2004.00339.x

Jain S., & Mukand S. (2005). Public opinion and the dynamics of reform. In *Sixth Jacques Polak Annual Research Conference*.

Kang H., Peck L., & Keele L. (2018). Inference for instrumental variables: A randomization inference approach. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *181*(4), 1231–1254. https://doi.org/10.1111/rssa.12353

Langer M. (2007). Revolution in Latin American criminal procedure: Diffusion of legal ideas from the periphery. *The American Journal of Comparative Law*, *55*(4), 617–676. https://doi.org/10.1093/ajcl/55.4.617

Latinobarómetro (2016). Latinobarómetro Report 2016: Results by Sex and Age - Uruguay, https://www.latinobarometro.org/latContents.jsp, Study LAT-2016. Accessed July 21, 2025.

Latinobarómetro (2018). Latinobarómetro Report 2018: Results by Sex and Age - Uruguay, https://www.latinobarometro.org, Study LAT-2018. Accessed July 21, 2025.

Lipsitch M., Tchetgen Tchetgen E., & Cohen T. (2010). Negative controls: A tool for detecting confounding and bias in observational studies. *Epidemiology*, *21*(3), 383–388. https://doi.org/10.1097/EDE.0b013e3181d61eeb

Maier J., & Struensee E. (2000). Introducción: Las reformas procesales penales en América Latina. In J. Maier, K. Ambos, & J. Woischnik (Eds.), *Las Reformas Procesales Penales en América Latina* (pp. 17–34). Buenos Aires: Ad-Hoc.

Miller D. L. (2023). An introductory guide to event study models. *Journal of Economic Perspectives*, *37*(2), 203–230. https://doi.org/10.1257/jep.37.2.203

Palumbo I. (2018). Para Bonomi, el nuevo código trajo aparejado contradicciones entre jueces, fiscales y policías, *Semanario Crónicas*, July 27th.

Prieto Curiel R. (2023). Weekly crime concentration. *Journal of Quantitative Criminology*, *39*(1), 97–124. https://doi.org/10.1007/s10940-021-09533-6

Rosenbaum P. R. (2002). *Observational studies* (2nd ed.). Springer.

Rosenbaum P. R. (2010). *Design of observational studies* (Vol. *10*). Springer.

Soares Y., & Sviatschi M. M. (2010). Does Court Efficiency Have a Deterrent Effect on Crime? Evidence for Costa Rica, *Unpublished Manuscript*.

Solomita M. (2019). Fiscales Agobiados: Turnos Interminables, Licencias por Estrés y Jubilaciones Anticipadas, *Diario El País*, Suplemento Qué Pasa, May 18th.

Stokes S. C. (1996). Public opinion and market reforms: The limits of economic voting. *Comparative Political Studies*, *29*(5), 499–519. https://doi.org/10.1177/0010414096029005001

Stokes S. C. (Ed.) (2001). *Public support for market reforms in new democracies*. Cambridge University Press.