Incentive Effects of Recall Elections: Evidence from Criminal Sentencing in California Courts

Sanford C. Gordon* and Sidak Yntiso†

Abstract

39 U.S. states authorize recall elections, but the incentives they create are not well understood. We examine how changes in the perceived threat of recall alter the behavior of one set of officials: judges. In 2016, outrage over the sentence imposed on a Stanford athlete following his sexual assault conviction sparked a drive to recall the presiding judge. Using disposition data from six California counties and arrest records for a subset of defendants, we find a large, discontinuous increase in sentencing severity associated with the recall campaign's announcement. Additional tests suggest that the observed shift may be attributed to changes in judicial preferences over sentencing and not strategic adjustment by prosecutors. We also demonstrate that the heterogeneous effects of the announcement did not mitigate preexisting racial disparities. Our findings are the first to document the incentive effects of recall and suggest that targeted political campaigns may have far-reaching, unintended consequences.

*Department of Politics, New York University. 19 West 4th Street, 2nd Floor. New York, New York, 10012. Email: sanford.gordon@nyu.edu

†Department of Politics, New York University, New York, New York, USA. Email: sidak.yntiso@nyu.edu

1 Introduction

When evaluating the performance of incumbents, voters may be only dimly aware of those officials' choices and their consequences, and the availability to voters of relevant information may itself be contingent on incumbent behavior. Either consideration may distort the incentives of elected officials. Accordingly, assessing the extent to which electoral institutions mitigate or exacerbate such distortions is a critical task for empirical political science.

One electoral institution that enjoys widespread use in the United States is the recall election, in which voters may remove an incumbent from office before the expiration of his or her ordinary term. 39 states authorize recall elections for at least some offices. In the 19 states that allow for gubernatorial recall, eleven governors have faced recall since 2018 (Greenblatt, September 17, 2020). And of the 45 state-level recall elections in U.S. history, nearly half have occurred in the past ten years.[1]

Despite the availability and increasing prominence of the recall option, there exists, to our knowledge, no study systematically assessing how the threat of recall affects incumbent behavior. This lacuna may stem in part from a host of methodological challenges. Most obviously, recall campaigns are not randomly assigned, and so comparing officials who do and do not experience them is likely to suffer from substantial omitted variables bias. Relatedly, if public officials rationally anticipate the consequences of a recall threat, they may take pains to avoid it. And finally, the behavior of officials in jurisdictions with and without the recall institution may differ in innumerable ways besides the availability of that specific institution, and the durability of the institution makes within-jurisdiction comparisons infeasible.

To circumvent these difficulties, we examine the effect of a shock to the salience of the recall threat brought by a widely publicized and ultimately successful recall campaign. In

[1]Source: `https://www.ncsl.org/research/elections-and-campaigns/recall-of-state-officials.aspx`.

June 2016, Santa Clara Superior Court Judge Aaron Persky achieved notoriety for imposing an unusually lenient sentence on Brock Turner, an affluent, white Stanford student athlete convicted of two counts of sexual assault and one count of attempted rape. Two years later, 61.5% of Santa Clara voters elected to remove Persky from office.

Several weeks before the 2018 vote, Persky delivered a speech in which he lamented, "The judicial recall, if successful, will be a silent force, a silent corrupting force. A force that will enter the minds of judges as they contemplate difficult decisions." A host of elected officials, political activists, and legal academics echoed this warning about the incentive effects of the recall effort, arguing that it might push judges to become more punitive in their sentencing decisions, even while condemning the leniency of the specific sentence that instigated the campaign.[2] Several of these observers argued that the burden of any change in electoral incentives would be borne disproportionately by minority defendants rather than white ones – a cruelly ironic prediction given that the behavior the recall aimed to sanction was leniency toward a white defendant.

Using data on almost 20,000 sentences handed down by over 158 Superior Court Judges in six California counties from 2015 to 2018, we examine whether critical events in the recall campaign were accompanied by corresponding changes in other judges' sentences. Specifically, using a regression discontinuity in time (RDit) approach (Hausman and Rapson, 2018), we examine the effects of two specific events: the initial announcement of the recall petition; and the recall election itself.

Our main results point to an instantaneous increase in average sentence length of over 30% in the immediate aftermath of the recall petition announcement. This result is robust to the inclusion of judge- and charge-level fixed effects, and a battery of placebo and specification tests. We demonstrate that the effect is driven by increases in sentencing on non-sexual

[2]In a similar spirit, we strongly caution against interpreting the findings presented in this paper as speaking to the merits of the specific sentence that motivated Judge Persky's recall.

violent crimes, and is unlikely to be an artifact of strategic adjustment by prosecutors. By contrast, we find no evidence that the recall election itself induced changes in sentencing. The conjunction of these findings suggests that the announcement of a well-organized, well-funded recall campaign against a Superior Court Judge signaled a new political reality for judges that was "priced in" by judges by the time the election took place.

Next, we consider whether the effects of the observed shift were borne disproportionately by minority defendants. Drawing on recent research decomposing the sources of racial disparities in sentencing (e.g., Rehavi and Starr, 2014), we hypothesize direct and indirect channels through which disproportionate burdens might manifest themselves. We find little evidence for either channel: sentencing increases were larger for white than minority defendants, but this does not appear to have mitigated preexisting, longer-term racial disparities.

In the last part of the paper, we estimate the aggregate effect of the change in judicial precipitated by the petition announcement over a narrow (45 day) time frame. Our most conservative estimates suggest that the petition announcement led to approximately 88 years of additional prison time in the six counties for which we have data.[3]

In the most immediate sense, our findings corroborate concerns that the campaign to remove a sitting a judge would affect the behavior of other judges, and amplify preexisting criminal justice disparities. In doing so, these findings complement prior research on the impact of retention on judicial behavior, while contributing to our understanding of the role recall elections may play in contemporary political life. Arguments in favor of recalling an elected official invariably focus on a selection function: the recall gives voters the opportunity to remove a specific malfeasant public official. The findings presented here suggest broader incentive effects that may extend beyond the official in question, and operate counter to the objectives of the recall's proponents.

[3]Absent a more comprehensive account of appropriate levels of sentencing, we are reluctant to draw any normative conclusions from this result.

2 Background

2.1 Institutional Setting

Recall elections in California and elsewhere. 39 states allow recall elections – those in which voters have the opportunity to remove a public official prior to the expiration of his or her term – in some form. Of the states that permit recall, seven specifically permit recall of judges. California, the setting of our main analysis, adopted recall elections in 1911. Since then, there have been 165 attempts to recall statewide officials, of which six were ultimately successful.[4] Far more ubiquitous in the state are recall efforts for local officials. Elected state legislators have been removed by voters in safe as well as competitive districts (Morton, 2006). Since 1995, recall attempts for 333 local officials have qualified for the ballot (reflecting a fraction of the full set of recall attempts); of these 244 have been successful.[5] The anti-Persky campaign represents the only attempt to recall a trial court judge to qualify for the ballot anywhere in the U.S. since 1982 (Spivak, 2020).[6]

Judges and judicial discretion in California. California has the largest judicial system in the nation, with 1,743 authorized superior court judges sitting in 58 county courts. During 2016–2017, approximately 6 million cases were filed in these courts. Superior courts in California have jurisdiction over civil and criminal cases. Since 1998, superior courts are the only consolidated general jurisdiction trial courts. Judges run in non-partisan competitive elections for six-year terms. In the event of a vacancy, judges are appointed by the Governor.

Judicial discretion over sentencing in California is constrained by a complex array of

[4]Source: Complete List of Recall Attempts, California Secretary of State, `https://www.sos.ca.gov/elections/recalls/complete-list-recall-attempts/`.

[5]Source: California Election Data Archive, `http://hdl.handle.net/10211.3/210187`.

[6]Three Los Angeles County judges were recalled in 1932 (Smith, 1951). To our knowledge, this precedent was never cited in contemporary coverage of the Persky recall.

considerations. Since 1977, sentencing for most crimes operates according to a triad system, in which the judge is given the choice between upper, middle, and lower "base" terms. For example, Assault with a Deadly Weapon (§245(a)(1) of the California Penal Code) carries a base term of 2, 3, or 4 years in prison. Despite a presumption in favor of the middle term in the absence of aggravating or mitigating factors, few sentences precisely match the three prescribed base terms, for three reasons. First, judges have discretion over whether the sentences for convictions on multiple counts run consecutively or concurrently. Second, judges can issue sentencing enhancements for aggravating factors such as gang or hate crimes, or prior convictions. Third, since 2011, judges have been granted discretion to issue suspended or split sentences for certain felonies.[7]

As is generally the case in the United States, the vast majority of cases are resolved via plea bargain. Plea agreements consist of a guilty plea and a sentencing recommendation to the judge, who has ultimate discretion on whether to accept or reject it.[8] Even still, a threat to inference that must be addressed is the possibility that the Persky recall induced changes in the behavior of *prosecutors* rather than in the behavior of judges. We discuss this issue, and our strategy for circumventing it, in detail below.

2.2 The Persky Recall

Our empirical analysis focuses on a shock to the salience of the recall threat to judges in California brought about by the campaign to recall Judge Aaron Persky from 2016 to 2018. The campaign was initiated in response to Judge Persky's sentencing decision in a widely publicized sexual assault case. On January 28, 2015, Brock Turner, a white Stanford student

[7]Effective since 2015, many crimes that are neither sexual crimes, violent crimes nor serious crimes are also eligible for county jail sentencing.

[8]See, e.g., *People vs. Orin*, 13 Cal. 3d 937 (1975) and *People vs. Clancy*, 56 Cal. 4th 562 (2013).

athlete, sexually assaulted Chanel Miller,[9] a visiting student, and was arrested. Five days later, Turner was indicted on two rape counts, two felony sexual assault counts, and one attempted rape count. The rape charges were later dropped, and in March 2016, Turner was convicted on the sexual assault and attempted rape charges.

Turner faced a maximum sentence of 14 years for these convictions, but on June 2, 2016, Judge Persky sentenced Turner to six months in prison and three months probation. The lenient sentence and Miller's impact statement, published by Buzzfeed, sparked widespread national attention.[10] On June 6, 2016, Stanford Law School Professor Michele Dauber announced the formation of a committee and began the process of collecting signatures to recall Judge Persky. With 94,000 verified signatures collected, the Santa Clara Registrar certified the signature threshold had been met on January 24, 2018. Finally, Judge Persky was recalled (with 61.5% supporting removal) on June 5, 2018. According to the *Palo Alto Daily Post*, the campaign to remove Persky raised more than one million dollars.[11]

Criticisms of the recall campaign were immediate. 95 Californian law professors signed an open letter in August 2017 opposing the recall petition. Californian mayors, state legislators, former Supreme Court justices, and hundreds of Superior Court judges supported the Retain Judge Persky Campaign.[12] Critics were primarily concerned with judicial independence and

[9]Miller has specifically expressed a preference *not* to remain anonymous, both in public appearances and her memoir (aptly titled *Know My Name*).

[10]`https://www.buzzfeednews.com/article/katiejmbaker/heres-the-powerful-letter-the-stanford-victim-read-to-her-ra`.

[11]`https://padailypost.com/2018/05/27/recall-persky-campaign-raises-more-than-1-million/`. By contrast, an attempt to recall an Orange County judge in 2015 raised less than $25,000 and did not achieve the required signatures (Source: `https://www.nbcnews.com/news/us-news/group-pushing-recall-effort-stanford-rape-case-judge-it-long-n590431`).

[12]Archived version of the associated website, Voices Against Recall available at `https:`

impartiality (Santa Clara County Association, June 14, 2016; Law Professors Statement, August 15, 2017). Some also predicted an increase in judicial punitiveness, with disproportionate effects on minority defendants (Butler, July 11, 2016; Gersen, June 17, 2016; Woolf, June 24, 2016). These predictions were bolstered by the empirical literature, cited below, documenting how concerns with reelection induce trial judges to impose longer sentences; as well as the significant literature, also discussed below, documenting the disproportionate burden imposed by the criminal justice system on minority defendants.

Which events during the recall campaign would we expect to have the largest effect on judicial behavior? On the one hand, a rational expectations account would anticipate that the petition announcement, insofar as it was the earliest and most surprising to the judges, would have the largest effect. (By contrast, the effect of later developments would have already been "priced in" to judges' electoral calculations.) On the other hand, the election itself may have had a larger effect if there was greater attention paid to it than to the announcement. To assess the plausibility of this countervailing mechanism, we collected data on Judge Persky's media salience over time using data from Google trends. This analysis (see Figure A.1 in the Appendix) suggest that there was 5.6 times as much interest in Judge Persky during the week of the announcement than the week of the recall election. Data from Lexis-Nexis on the number of news articles mentioning Persky tell the same story: 4.4 times more stories about Persky around the announcement than the election.

2.3 Related Research

Electoral incentives. The current research contributes to a rich literature on the incentive effects of electoral institutions on the behavior of incumbents generally (see, especially, Besley and Case, 1995; Alt, Bueno de Mesquita, and Rose, 2011; Ferraz and Finan, 2011) and judges specifically (Besley and Payne, 2013; Brace and Boyea, 2008; Brace and Hall, 1995;

//web.archive.org/web/20180423164925/http://www.voicesagainstrecall.org/.

Canes-Wrone, Clark, and Kelly, 2014; Calderone, Canes-Wrone, and Clark, 2009; Huber and Gordon, 2004; Gordon and Huber, 2007; Hall, 1992; Lim, 2013; Berdejo and Yuchtman, 2013; Matsusaka et al., 2010). One feature of judicial elections that makes them particularly noteworthy in the empirical analysis of electoral incentives is the nature of the informational environment in which they occur. Voters often lack verifiable information to evaluate judicial performance, a problem further complicated by the fact that judges often face voters in retention elections (in which there are no challengers) and nonpartisan elections (in which voters lack clear cues such as party labels). As a result, voters may be highly responsive to well-publicized examples of apparent judicial "error," as revealed by the media, organized interest groups, or challengers. If being perceived by voters as overly lenient is either more likely or more politically costly than being perceived as overly punitive, judges will face electoral pressures to become more punitive than they would be otherwise, *even if* under ordinary circumstances voters know little or nothing about judicial behavior.

Recall Elections. To our knowledge, there exists no extant empirical research on the incentive effects of recall elections. Political science research on recall elections has instead focused on voter behavior in recall elections – see, e.g., Ho and Imai (2006); Segura and Fraga (2008); Masket (2011); Shaw, McKenzie, and Underwood (2005). One explanation for this lacuna might be that the most straightforward research designs available to researchers do not translate well to the recall setting. Because the institution of recall is not randomly assigned, comparing the behavior of officials in states with and without recall is likely to be confounded by numerous other interstate differences. There are also issues characterizing variation in the "treatment" of officials within the same state because the timing and occurrence of recall attempts are random and idiosyncratic. Finally, studying changes in the behavior of an individual official subject to a recall effort will afford essentially no statistical power. More generally, a challenge to studying the effects of recall elections on official behavior is that the threat of recall will be "priced into" the behavior of the officials. Unanticipated shocks,

should they occur, are likely to be exceptionally rare and highly localized.[13]

Judicial bias and asymmetric burdens of criminal justice system African Americans face a six-fold greater rate of imprisonment than whites in the United States (Bronson and Carson, 2019). While noting potential racial differences in criminal behavior, a number of recent papers have highlighted the influence of disparities induced by judicial and prosecutorial discretion, even among defendants facing similar charges and of similar criminal backgrounds. Evidence from randomly assigned cases indicates that judges differ in the degree to which race influences their likelihood of incarceration (Abrams, Bertrand, and Mullainathan, 2012). In Kansas, retention pressures, discussed above, induce increased judicial punitiveness but only in cases involving black felons (Park, 2017). Capital punishment sentences involving white victims are significantly more likely to be overturned by appellate courts when the defendant is African American, providing evidence that lower courts discount the wrongful convictions of black defendants (Alesina and La Ferrara, 2014). Racially disparate judicial decision-making is in turn compounded by racial disparities in charging and plea bargaining (Rehavi and Starr, 2014), jury decision-making (Bayer, Hjalmarsson, and Anwar, 2012) and policing (Grogger and Ridgeway, 2006).

3 Data and Approach

3.1 Data on Sentencing in California

Unlike in other states, at the time of writing there is no publicly accessible, centralized repository for sentencing data. To overcome this limitation, we scraped criminal cases with hearing dates between January 2015 and December 2018 from the websites of six superior courts that make these data available electronically: Fresno, Napa, Sacramento, Santa

[13]In this vein, the unusual nature of the Persky recall should be viewed as an essential feature of our research design rather than a flaw.

Barbara, San Bernardino, and Santa Cruz.[14] Our search produced a total of 19,744 cases encompassing 21,939 felony charges with initial sentencing dispositions in the remaining six courts.[15] The sample counties represent 19% of California's total incarcerated population.[16]

Each count on which the defendant is convicted is associated with a sentence length in days. 92% of cases have only one count in the conviction. For the remaining 8% of cases with multiple counts, we encountered a number of inconsistencies in the data: generally, whether sentences run concurrently or consecutively is not evident – in some cases, the total sentence is entered for a top count (in clear excess of the legal maximum for that count), while in others the same sentence is entered for counts with dramatically different sentencing ranges. To reduce the effect of this issue, in our main analysis we restrict attention to the sentence entered for the top count in the conviction, and for specifications in which we adjust for covariates, we include the total number of counts. We also conduct a robustness test in which we restrict attention to convictions with only one count.

For each offense code, we acquired base terms from the State of California Attorney General's office operative for the period of our sample.[17] Additional case information in our final dataset include the charge (410 unique offenses) and sentencing judge (157 unique judges). We categorized crimes as nonviolent or violent based on offense codes from the California Attorney General: 82.7% of cases in the sample are classified as nonviolent crimes; 4.3% as (violent) sex crimes, and the remaining 13% as other violent crimes.

To explore heterogeneity by race, we linked defendants in our data to publicly available

[14]Alameda has online data but it is radically incomplete.

[15]Sentences may be amended – for example, in cases of probation violations.

[16]While we have no a priori reason to believe that the effect of the recall should be unusual in these counties, we emphasize that our estimates are local average treatment effects, and specific to those counties.

[17]`https://oag.ca.gov/law/code-tables`.

arrest records sourced from county and municipal law enforcement agencies in California.[18] We crawled 201,066 arrest records. Defendants were matched based on first name, middle name, last name, county of arrest and arrest date. Across the six counties, 12,844 defendants could be matched to arrest records, of which 11,184 defendants have race identified.

3.2 Empirical Approach

In the main part of our analysis, we look for sharp increases in judicial punitiveness immediately following key moments during the recall campaign. In particular, we consider two critical events: the announcement of the campaign itself, on June 6, 2016; and the recall election itself, on June 5, 2018.[19] Studying events in the timeline of the Persky recall permits us to evaluate the extent to which those events perturbed judges' perceptions about their own electoral vulnerabilities. Our main specification is the following local linear regression:

$$y_{ijt} = \beta_0 + \beta_1 \mathbb{I}(t > t_k) + \beta_2 f(t - t_k) + \varepsilon_{ijt} \tag{1}$$

Where t_k is the calendar date of a critical event k; $y_{ijt} \equiv \min\{s/\overline{s}, 1\}$ is the normalized sentence of conviction i by judge j at time t; and $f(\cdot)$ is smooth function of time. The normalization divides the sentence length in days s by the upper base term $\overline{s}$, creating a fractional measure of judicial discretion expressable in percentage terms and comparable across different offenses (cf., Lim, 2013). So, for example, a sentence of six months on an assault with a deadly weapon charge with an upper base term of four years would be coded as 0.125. The measure is censored at one so as not to be skewed by cases with unusual aggravating factors (or multiple charges) that increase the sentence above the upper base term for the top count. In point of fact, more than 95% of cases fall at or below the upper

[18] Source: `https://www.localcrimenews.com/`.

[19] Another candidate is the date on which signatures were certified by the Santa Clara Registrar: January 24, 2018. Results from this event may be found in Appendix Figure C.1.

base term. In robustness tests we use the uncensored measure as well as the raw sentence (in days) as outcome measures. We report RD estimates using the bias correction method and bandwidth selection approach described in Calonico, Cattaneo, and Titiunik (2014).

In addition to this unadjusted specification, we also present results throughout that adjust for a vector of judge- and offense-specific fixed effects, as well as the number of counts in the conviction. The adjusted estimates discard sentencing data from Sacramento County, whose data do not include judge identifiers. For both sets of estimates, we weight observations using a triangular kernel (as is standard). Standard errors for the adjusted specification are clustered at the judge-statute level and for the unadjusted specifications (given the absence of judge identifiers in Sacramento) at the county-statute level.

As noted above, we wish to rule out the possibility that observed changes in sentencing associated with the Persky recall are driven by prosecutorial, rather than judicial behavior. Note that from the perspective of the analyst, a more stringent plea offer made by a prosecutor in expectation of increased judicial stringency is observationally equivalent to a relatively lenient plea offer that is rejected by a judge, necessitating a second round of negotiation between the prosecutor and defendant. Both, however, are consistent with the account of stronger electoral incentives for judges brought about by the recall campaign. Another possibility, however, is that the recall campaign induced an increase in prosecutorial rather than judicial stringency. To adjudicate between these two accounts, we first assess whether the recall campaign led to a reduced willingness of prosecutors in a domain over which they exert greater autonomy: charge reduction, that is, the decision to drop higher counts in an indictment as a condition of a guilty plea. Our measure of charge reduction is equal to one minus the ratio of the maximum sentence at conviction to the maximum sentence at arraignment (restricted to the set of cases that did not go to trial and for which arraignment occurred prior to the critical date). To be sure, a prosecutor's inflexibility in this area will reflect expectations about the likelihood of conviction on more severe charges. To the

extent that the recall campaign influenced *jury* decisionmaking at trial, it would bias any finding of the effect of the recall on charging away from zero. Accordingly, a null finding for this test would strongly argue against the importance of the prosecutorial channel. We also conduct supplemental tests to see whether critical dates in the recall process were followed by compositional changes in the set of top charges reached at conviction.

The identifying assumption of regression discontinuity designs is that treatment assignment is conditionally ignorable sufficiently close to the cutoff (the critical event in the RDit setting). We examine threats to inference arising from shocks that vary discontinuously within the treatment windows. A sequence of placebo regressions for all dates in each calendar year alleviates the concern that the findings result from some confounding structural break (for instance, the ratification of two laws in September 2016 requiring mandatory sentences for sexual assault). To bolster further our claim that the recall events do not coincide with unrelated shocks to judicial decision-making, we examine contemporaneous sentencing patterns in the nearby state of Washington. Finally, we assess robustness to various specifications of the outcome and bandwidth.

Next, we examine whether any observed effects of key events on punitiveness are driven by sentencing for sexual, non-sexual violent or nonviolent crimes. As the recall campaign centered around Judge Persky's sentencing in a sex crime case, judges might have anticipated that voters would pay greater attention to perceived leniency on similar cases.

Third, we assess whether any increase in judicial punitiveness induced by the recall campaign placed a disproportionate burden on minority defendants. It is important to note that there are (at least) two channels through which this might operate. One involves judges apprehensive that a racially biased electorate might react especially negatively to perceived leniency toward minority defendants. In that case, we would anticipate that the instantaneous effect of the recall on sentencing would be larger for them than than for white defendants. Call this the *direct racial burden* hypothesis.

Testing the direct burden hypothesis is subtle, as the following example demonstrates: suppose that racial disparities in discretionary sentencing were *already present* prior the electoral shock, so that white defendants were sentenced at the lower, and minority defendants at the upper, ends of judges' discretionary sentencing range. In that case, and even in the presence of the direct channel, we might observe a (weakly) larger effect of the electoral shock for white than minority defendants. To assess this possibility, we examine whether race-based disparities in discretionary sentencing were in evidence immediately prior to the electoral shocks. Note that this is a descriptive exercise aimed at clarifying the mechanism – here we lack a strong identification strategy for assessing the causal effect of race on sentencing discretion directly (clearly a critical task, but one beyond the scope of this paper).

The absence of evidence for the direct burden hypothesis does not eliminate the possibility that minority defendants are hit harder by the consequences of an electoral shock brought about by the recall drive: it could be that white and minority defendants are charged with crimes that vary in severity. Suppose judges increase their discretionary sentencing in an apparently race-neutral way – e.g., from 0.5 to 0.6 on the normalized sentencing scale – but that minority defendants tend to be convicted of crimes with higher statutory maximum penalties. Then the electoral shock will mechanically lead to a higher cumulative sentencing load for minorities. Call this the *indirect racial burden* hypothesis. To assess the indirect burden hypothesis, we examine whether minority defendants in our sample are convicted of crimes with higher maximum penalties, and test for heterogeneous effects in total sentencing.

The RDit approach identifies a local average treatment effect (LATE) at the precise moment of the critical event in question. In the final part of our analysis, we compute aggregate effects, which require estimating longer-term consequences of electorally-induced shifts in judicial behavior. Doing so requires more stringent identifying assumptions than those necessary to identify the LATE. Rather than committing ourselves to one set of assumptions, we adopt four separate approaches: (1) assuming the estimated LATE persists as an average

treatment effect in a window of time after the announcement; (2) a fully parametric approach that attributes any post-announcement time trends to decay or growth in the effect of the announcement itself; (3) a linear reweighting estimator; and (4) a propensity score estimator (the latter two approaches suggested by Angrist and Rokkanen (2015)).

4 Empirical Results

4.1 Main Results: Instantaneous Effects of Critical Events

Graphical Evidence. Before proceeding to our local linear estimates, our first step in assessing the effect of the events described above is graphical. Figure 1 illustrates the main effects at the core of the paper, presenting binned averages of our normalized sentence measure, within the 90 day window surrounding each event.[20] Linear predictions and local polynomial smoothers are fit separately on either side of the event date.

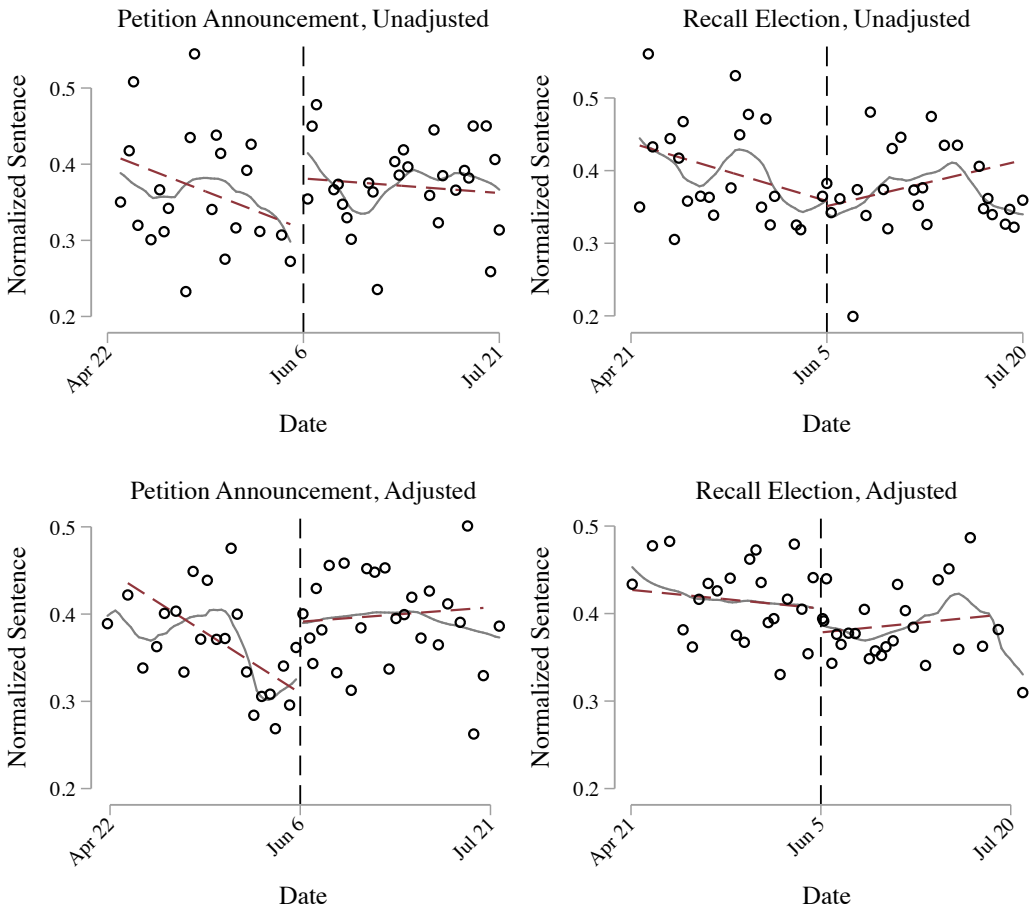
The two panels in the top row present plots of data unadjusted for covariates. We observe a large, discontinuous increase in average normalized sentence transitioning from immediately prior to, to immediately following the June 6, 2016 petition announcement. This increase corresponds to an increase of approximately 10 percentage points on the normalized sentencing scale, reflecting a proportionate increase of approximately 29 percent over pre-announcement levels. Turning to June 5, 2018, the date of the recall election itself, we observe no change in average sentence length from before to after that date.

Plots in the bottom row depict binned means residualized using judge- and offense-specific fixed effects and the number of conviction counts. The graphical analysis of the petition announcement adjusting for the fixed effects reveals a similar pattern to that in the unadjusted panel: an increase of around eight percentage points. Using the adjusted estimates, we again find no apparent difference before and after the election itself.[21]

[20] The ± 45 day window approximates the MSE-optimal bandwidth; see below.

[21] A declining pre-announcement trend in average sentences justifies our local linear re-

Figure 1 Effect on Sentencing of Critical Events in Persky Recall: Graphical Analysis



The left panels depict average normalized sentence lengths (as tokens) in equally-sized bins. The panels to the right depict binned means residualized using judge- and offense-specific effects and number of counts at conviction. Linear fit (maroon) and local polynomial smoothers (gray) fit separately on each side of event.

Local Linear Regression Results To interrogate the preliminary inferences suggested by the graphical analysis in a more rigorous fashion, we next present local linear regression estimates of the local average treatment effect (LATE), β_1 from equation (1). The LATE estimates appear in Table 1.[22]

gression specification, which accounts for this trend.

[22]Following the recommendation of Gelman and Imbens (2019), we report results from a local linear specification rather than estimating higher-order polynomials.

Table 1 Effect on Sentencing of Critical Events in Persky Recall: RD Estimates

	Petition Announced		Recall Election	
RD estimate	0.09	0.103	0.014	0.019
	(0.044)	(0.032)	(0.058)	(0.047)
Left-side intercept	0.303	0.311	0.341	0.357
	(0.03)	(0.023)	(0.049)	(0.039)
Bandwidth	45.1	47.3	40.5	36.8
Adjusted	N	Y	N	Y
Effective observations	1,476	1,289	1,209	954

Dependent variable in each column is the normalized sentence length (see text for description). Estimates in the second and fourth columns adjust for the number of counts and judge- and statute-specific fixed effects, and exclude Sacramento County (which does not report judge identifiers).

Reported estimates corroborate the results from the graphical analysis. We estimate a large, statistically significant effect of the June 6 petition announcement: unadjusted (first column), the estimated effect is 9 percentage points on the normalized sentencing scale; adjusted for judge- and offense-specific fixed effects (second column), the estimate increases to 10.3 points. To give a sense of the estimates' substantive import, immediately prior to the announcement, the estimated average normalized sentence length (the left-side intercept in the Table) was around 0.3; hence, these effects correspond to a proportionate increase of 29.8 percent. (Using the same baseline, the adjusted estimate implies a 33.1 percent increase.) Both RD estimates easily surpass conventional thresholds for statistical significance.

The third and fourth columns of the table display the analogous estimates for the recall election date. In contrast to the announcement estimates, the estimated effect, whether adjusted or unadjusted, is small and statistically indistinguishable from zero.

4.2 Prosecutors or Judges?

The next step in our analysis is to assess whether the main finding reflects adjustment by prosecutors rather than judges. We proceed in three steps. First, to minimize the role of

Table 2 Assessing the Prosecutorial Adjustment Hypothesis: RD Estimates

	Petition Announced		Recall Election	
A. Sentence Normed to Top Arraignment Count				
RD estimate	0.107	0.117	0.04	0.041
	(0.044)	(0.031)	(0.052)	(0.04)
Left-side intercept	0.259	0.268	0.306	0.32
	(0.028)	(0.022)	(0.042)	(0.032)
Bandwidth	39.4	40	38.5	37.2
Adjusted	N	Y	N	Y
Effective observations	1285	1079	1181	966
B. Charge Reduction				
RD estimate	-0.022	-0.031	-0.017	-0.023
	(0.027)	(0.019)	(0.023)	(0.022)
Left-side intercept	0.104	0.097	0.064	0.072
	(0.022)	(0.014)	(0.018)	(0.018)
Bandwidth	45.6	45.5	45	38.2
Adjusted	N	Y	N	Y
Effective observations	1268	1088	1209	915

idiosyncratic variation in prosecutorial discretion in our central results, we re-ran the main analysis, substituting the statutory maximum sentence for the top count at *arraignment* for its analog at *conviction* as the denominator of the outcome variable, and restricting the sample to cases arraigned before the critical events under study. Results appear in Panel A of Table 2. Given the fact that the maximum sentence at arraignment is weakly larger than the maximum sentence at conviction, it is unsurprising that the left-side intercepts are smaller than the analogous estimate in Table 1. More notable is that the RD estimates associated with the petition announcement are slightly larger. The combination of these effects implies that normalized to the arraignment maximum, the estimates imply a proportionate increase (relative to baseline) on the order of 41.4 to 43.7 percentage points.

Second, we assess the effects of the petition announcement and recall on charge reduction by prosecutors, which is outside of direct control of sitting judges. Panel B of Table 2 reports

RD estimates using the measure of charge reduction as the outcome. A significant, negative RD estimate would reflect a reduced willingness of prosecutors to drop higher counts as a condition of plea bargaining following the event in question. While the estimates are negative, they are small and imprecisely estimated. Hence, we cannot reject the null hypothesis that charge reduction practices were unaffected by the petition announcement.

Finally, we look for evidence of any change in the composition of the set of cases around the petition announcement. If, for example, prosecutors expedited convictions for more severe offenses in response to the petition announcement, we might observe a mechanical, positive effect on sentencing severity. To assess balance on the charges' distributions, we present RD estimates for the daily count of each crime against its severity (logged maximum possible sentence in days) in Figure B.1 in the Appendix. We find no evidence of an imbalance in charge severity that might induce the purported effect.

4.3 Additional Robustness Checks

Placebo tests for contemporaneous shocks. Our main analysis implies that the announcement of the recall petition caused a substantial and immediate increase in the length of felony sentences in California. One threat to inference is the possibility that other events may have been taking place around the time of the announcement. An event that is particularly relevant for our analysis is the 2016 California primary, which took place on June 7. A second is Persky's actual sentence of Brock Turner on June 2.

With respect to the primary, there are two immediate responses. First, a Superior Court judge who faced a challenger in 2016 did so initially in a top-two primary, and would only need to face the voters in the general election upon placing in the top two but receiving less than 50% of the vote. Owing to California's unusual electoral rules, the vast majority of judges would thus see the sway of electoral incentives *diminish* following a contested primary or remain roughly constant following an uncontested one. The anticipated behavioral response

(given the prior research cited above) would be a reduction in average sentence length; hence, the overall effect would be to bias the above results *downward*. In point of fact, only one incumbent judge in our sample (in San Bernardino County) faced a primary challenge, and she did not hand down a sentence in the sample period.

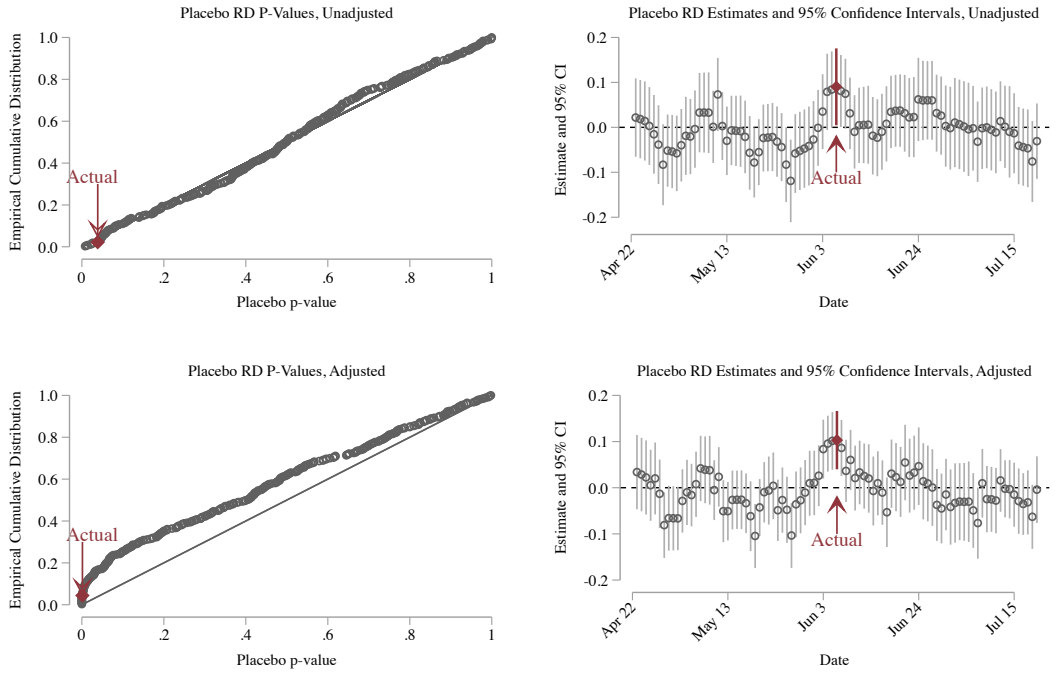
With respect to the Brock Turner sentence, it is less clear what the direction of the bias might be. Judges may, anticipating the electoral backlash from outrage over the sentence, ratchet up sentencing in their courtrooms in response; perhaps this is what our main results are capturing. This would confirm the power of anticipated electoral threat, but complicate our efforts to make inferences about the specific effect of the petition announcement. On the other hand, perhaps Turner's sentence signaled the acceptability of unusual sentences. In the first account, our main estimates are biased upward; in the second; biased downward.

Another threat to inference with which to contend is that there may be numerous structural breaks throughout the year that affect sentencing, some associated with the explicitly political stimuli discussed above and others associated with, inter alia, changes in sentencing guidelines, news accounts of prison overcrowding (or litigation on that issue), or shifts in prosecutorial behavior. The relevant question then becomes whether the shift associated with the June 6, 2016 cutoff was particularly unusual relative to other candidate breakpoints (including the Turner sentence) – recognizing that those other breakpoints might exist.

We therefore conducted a sequence of placebo tests, using every day of calendar year 2016 as a breakpoint in a sequence of regression discontinuity analyses. Figure 2 displays the results. The left panels display the empirical cumulative distribution of p-values for the placebo tests, along with the June 6 p-value (labeled "Actual"). For the unadjusted (adjusted) estimates, the figure demonstrates that the p-value for the RDit using June 6 as the breakpoint is smaller than 98% (96%) of its placebo analogs.

The right panels zoom in, for clarity, to the 90 day period surrounding the June 6, and plots 90 placebo estimates plus their associated 95% confidence intervals, along with the June

Figure 2 Placebo Tests for Main Effect of Petition Announcement



The left panels displays empirical cumulative distributions of estimated placebo p-values (in gray), with the actual petition announcement p-value overlaid in maroon. The right panels displays placebo RD estimates and associated 95% confidence intervals (in gray), with the actual estimate and confidence interval in maroon, for the 90 day period surrounding June 6, 2016.

6 estimate and its confidence interval (again labeled "Actual"). Because they are estimated using nearly identical data, any adjacent RD estimates are very unlikely to be statistically significantly different from each other. That being said, it is instructive that the actual June 6 estimate is larger than any of the surrounding placebo estimates – including June 7 (the primary) and June 2 (the sentence). In fact, the placebo estimates for the date of the Brock Turner sentence are statistically indistinguishable from zero.

Washington as a placebo state. To further assuage concerns that the petition announcement coincided with an unrelated shock to judicial decision-making, we assess shifts in judicial punitiveness in nearby Washington state. Like judges in California, judges in

Washington face nonpartisan elections (four year terms) and have broad discretion to issue sentences within the appropriate sentencing guidelines. However, the Washington state constitution does not allow for recall of judges. Using data sourced from the Washington State Department of Corrections, we extracted the sentencing judge, charge and sentence length associated with 67,441 charges. In Appendix Figure B.3, we present binned averages of the normalized sentence within 45 days of the petition announcement date. Neither the unadjusted averages nor averages adjusted for judge and charge fixed effects significantly change after the announcement date. Local linear regression results confirm the null finding.

Tests for bandwidth artifacts. While the estimates above employ a principled means of selecting the optimal bandwidth for the RD estimates, we wish to make sure that the significance of our results is not overly dependent on the breadth of the interval employed in the analysis. Accordingly, we re-ran our analysis of the effect of the petition announcement for different bandwidths, ranging from one week to 90 days. Results of this exercise appear in Figure B.2 in the Appendix. For very short bandwidths, of course, the sample size declines dramatically, substantially diminishing the precision of the estimates. However, past around a two-week bandwidth for both adjusted and unadjusted specifications, our main results are robust to a wide range of different bandwidths.

Alternative measures of the outcome. Finally, we consider whether our estimates are influenced by the choice of outcome variable. Table B.1 in the Appendix replicates the main analysis in Table 1 using the same normalization but without top-coding at one. This operationalization will pick up increases in judicial punitiveness that result from, e.g., decisions to let sentences for multiple charges run consecutively instead of concurrently. Using this alternative coding leads to slight changes in coefficient magnitudes, but reproduces the main results: substantial, statistically significant increases in sentencing associated with the petition announcement, and no significant change associated with the recall election.

Table B.2 in the Appendix uses the raw sentence (in days) rather than the normalized

measure. Here, the offense-specific fixed effects are particularly important, picking up mean sentence length for specific charges. Using the non-normalized time scale as the outcome, the substantive import remains unchanged, with our fixed effects estimates suggesting that the petition announcement had an average (within-charge, within-judge) effect of about 125 days additional incarceration (relative to a baseline of slightly more than a year).

As discussed in Section 3.1, we detected a few inconsistencies in how court clerks recorded sentences for multiple conviction counts arising in the same case. Appendix Table B.3 restricts attention to the 92% of cases that contain only one count. The substantively identical findings strongly imply that our main results are invariant to the corrections we employed to handle the data entry idiosyncrasies.

4.4 Effects by Type of Crime

A natural question to consider in assessing the above result concerning the petition announcement is the extent to which it is driven by increases in sentencing for different crimes. To the extent that the precipitating event for the Persky recall was a lenient sentence for a violent sex crime, we wish to consider whether the incentive effect of the petition was confined to sex crimes. Accordingly, we partition the sample of felony cases into sex crimes, non-sexual violent crimes, and nonviolent crimes,[23] and run the local linear regression estimator (unadjusted and adjusted for judge and offense fixed effects) separately for each of the three categories. Results appear in Table 3.

Turning first to the analysis of sex crimes (the first and second columns), one immediately notes the small sample size. This contributes to marked imprecision in the estimated coefficients, especially for the unadjusted specification. Both estimates are negative and nowhere close to statistical significance. By contrast, we observe highly significant estimates nearly twice the magnitude of the pooled estimates for non-sexual violent crimes (third and fourth

[23]We rely on California Penal Code §667.5 to categorize crimes as violent or non-violent.

Table 3 Heterogeneous Effects of the Petition Announcement: RD Estimates by Crime Type

	Sex Crimes		Other Violent Crimes		Nonviolent Crimes	
RD estimate	-0.033	-0.003	0.248	0.188	0.077	0.106
	(0.17)	(0.048)	(0.098)	(0.056)	(0.048)	(0.037)
Left-side intercept	0.29	0.448	0.211	0.2	0.318	0.319
	(0.133)	(1.8e-09)	(0.046)	(0.031)	(0.028)	(0.023)
Bandwidth	65.5	33.9	43.8	38.1	46.5	46
Adjusted	N	Y	N	Y	N	Y
Effective observations	85	29	207	151	1,232	1,041

See notes in Table 1 for estimation details.

columns). For non-violent crimes, RD estimates are in the vicinity of the pooled estimates, but only reach statistical significance in the covariate-adjusted specification (despite the far larger sample of non-violent felonies). Taken together, the effects described in the main analysis appear to be driven largely by increases in sentences for non-sexual violent crimes, and possibly for nonviolent crimes. This is consistent with the prediction by critics of the recall that any resulting increases in sentencing would extend beyond sex crimes.

4.5 The Recall Petition and Disproportionate Burden by Race

We next assess the argument made by critics of the recall effort that notwithstanding the aim of sanctioning a judge for imposing a lenient sentence for a White defendant, any increase in judicial punitiveness driven by the recall itself would likely be disproportionately borne by Black or Hispanic defendants. As discussed above, doing so requires adjudicating between the direct and indirect racial burden hypotheses.

The direct racial burden hypothesis. Two patterns in the data would be consistent with the direct burden mechanism: (1) a strictly more severe instantaneous effect of the petition announcement on normalized sentences for minority defendants than White defendants; or (2) a weakly more severe instantaneous effect for White than minority defendants, and a higher average normalized sentences for minority defendants prior to the announcement. Critically, the second pattern, while consistent with the direct burden hypothesis, would not

definitively confirm it. This is because the same pattern would also be expected if there were preexisting racial disparities, and either no racial differences in the effect of the petition announcement, or a larger effect for White defendants (for example, if outrage at the Turner sentence pushed judges to mitigate underlying racial biases in sentencing).

Panel A of Table 4 displays local linear RD estimates of the petition announcement reported separately for Black, Hispanic, and White defendants. The first thing to note is that relative to our main analysis, the effective sample size is considerably smaller, owing to the difficulty matching arrest and sentencing records. Second, coefficient estimates are highest for White defendants, followed by Black and then Hispanic defendants. (In neither specification can we reject the null hypothesis of no effect for Hispanic defendants.) Panel B of the table displays results from a sequence of hypothesis tests comparing the race-specific RD estimates. These tests permit us to reject null hypotheses of no racial differences in each of the covariate-adjusted tests, but none of the unadjusted tests. In other words, the effect of the announcement for white defendants is significantly higher than for Black or Hispanic defendants in the adjusted specification.

That the effects of the petition announcement are apparently largest for White defendants essentially rules out the first pattern consistent with the direct burden hypothesis. To assess the second, we consider the left-side intercepts associated with each RD estimation – these correspond to the expected sentence immediately prior to the announcement for defendants of different races. Hypothesis tests comparing left-side intercepts appear in Panel C of Table 4. Consistent with expectations, the intercept is lower for White than either Black or Hispanic defendants. However, as the test statistics indicate, in no case can we reject the null hypothesis that they are equal across defendant race. Taken together, these results suggest scant evidence for the direct burden hypothesis.

The indirect racial burden hypothesis. The indirect burden hypothesis suggests that comparable effects of the petition announcement across defendant racial categories could

Table 4 Assessing the Direct Racial Burden Hypothesis: RD Estimates by Defendant Race

A. RD Estimates by Race of Defendant						
	Black		Hispanic		White	
RD estimate	0.136	0.24	0.098	0.062	0.206	0.545
	(0.12)	(0.074)	(0.059)	(0.045)	(0.096)	(0.072)
Left-side intercept	0.355	0.37	0.302	0.31	0.247	0.225
	(0.066)	(0.055)	(0.036)	(0.026)	(0.064)	(0.05)
Bandwidth	64.2	41.6	65.9	54.1	56	26.6
Adjusted	N	Y	N	Y	N	Y
Effective observations	304	136	689	516	326	126
B. Hypothesis Tests of Equality of RD Estimates						
$H_0 : RD_{Black} = RD_{White}$	0.07	0.305				
	(0.153)	(0.103)				
$H_0 : RD_{Black} = RD_{Hispanic}$	0.039	0.178				
	(0.134)	(0.086)				
$H_0 : RD_{Hispanic} = RD_{White}$	0.108	0.483				
	(0.112)	(0.084)				
C. Hypothesis Tests of Equality of Intercepts						
$H_0 : LSI_{Black} = LSI_{White}$	0.108	0.146				
	(0.109)	(0.094)				
$H_0 : LSI_{Black} = LSI_{Hispanic}$	0.052	0.061				
	(0.091)	(0.075)				
$H_0 : LSI_{Hispanic} = LSI_{White}$	0.056	0.085				
	(0.085)	(0.073)				

See notes in Table 1 for estimation details. LSI is the left-side intercept, i.e., the value of the regression function estimated using data prior to the petition announcement at the date of the announcement.

obscure a disparate impact that would emerge if minority defendants tend to be charged with more severe crimes (and are thus eligible for higher sentences generally). Table 5 displays estimates from a regression of a case's *statutory maximum sentence* (in days) – a measure of crime severity – on indicator variables for race (Black and Hispanic – the omitted category is White),[24] adjusting in some specifications for judge-specific fixed effects. (As the primary instrument for manipulating charging severity is the choice of offense itself, we omit

[24]Only a tiny fraction of defendants are identified as Asian or Native American.

Table 5 Assessing Indirect Racial Burden: Racial Disparities in Crime Severity, as Measured by Statutory Maximum Penalties

	Pre-Announcement		Post-Announcement		Full Sample	
Black	73.0	60.5	79.4	80.0	78.1	76.2
	(14.3)	(56.6)	(21.1)	(16.0)	(19.0)	(15.7)
Hispanic	48.5	5.2	44.4	21.6	44.9	19.7
	(38.6)	(21.1)	(15.7)	(12.8)	(18.4)	(12.6)
Intercept	1300.0	1338.8	1283.4	1302.0	1286.2	1307.1
	(29.6)	(15.3)	(12.9)	(7.8)	(15.2)	(7.6)
Judge fixed effects	N	Y	N	Y	N	Y
N	1,805	1,155	9,343	7,304	11,148	8,459

The dependent variable in each column is the statutory maximum sentence (in days) associated with each charge. The excluded category is white defendants. Standard errors are clustered at the county (Columns 1, 3, and 5) or judge level (Columns 2, 4, and 6).

statute-specific fixed effects for this portion of our analysis.)

In the full sample, we find descriptive evidence that African American defendants indeed tend to be sentenced to more severe crimes than their white counterparts, with sentences for Hispanic defendants occupying an intermediate position. This pattern holds when we partition the data into pre- and post-announcement periods, although the statistical significance is attenuated for the pre-announcement period.

Next, recall that our earlier analysis suggested that the effect of the petition announcement on sentencing was larger for White than Black or Hispanic defendants (although the statistical significance of that disparity differed by specification). Thus, in evaluating the indirect burden hypothesis we are interested in assessing whether the apparently larger effect for Whites attenuated or reversed the underlying disproportionate burden for minority defendants. To assess this, we ran our unadjusted RD estimator, disaggregated by racial category, with the (logged) sentence length as the outcome variable. This will capture the total effect, by race, of the petition announcement on average sentencing.

Figure 3 displays the left- and right-side intercepts (and associated 95% confidence intervals) from this analysis for Black, Hispanic, and White defendants. The figure implies that we cannot reject the null hypothesis of no racial differences in sentence length *immediately* before, or after, the petition announcement. Coupled with the descriptive results in Table 5, which suggest that the disparities do exist in the time periods both before and after the announcement, the results imply that the apparently larger effect of the petition announcement on judicial sentencing for White defendants documented in Table 4 neither mitigated, nor exacerbated, any long-term discriminatory treatment in sentencing.

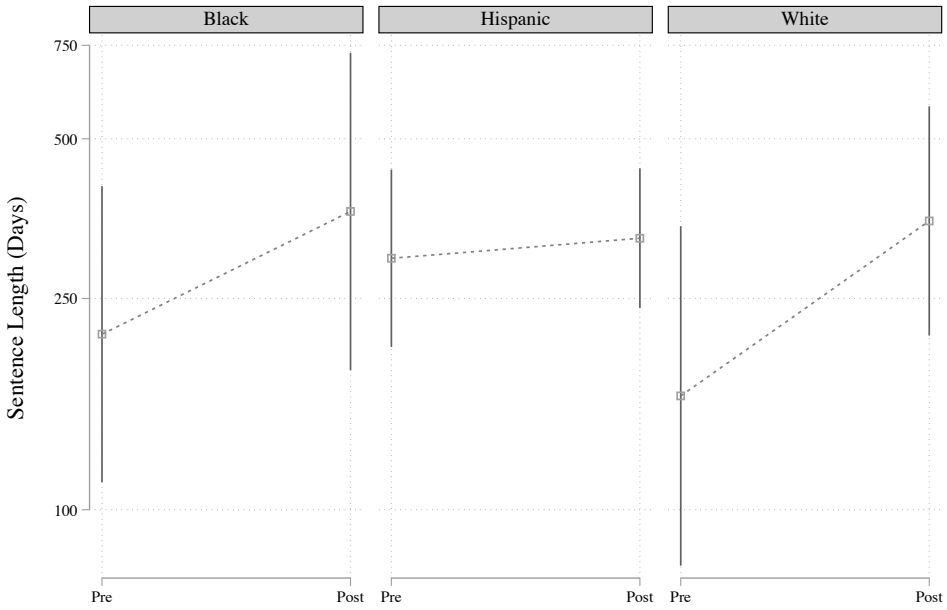
Note that recall campaign critics (e.g., Butler (July 11, 2016)) anticipated a disproportionate racial burden even in the absence of any immediate change in discriminatory treatment by judges. Specifically, these critics emphasized how the over-representation of Black citizens in courts and prisons,[25] implies that even a race-neutral increase in overall severity will place a disproportionate burden on minority communities. Our findings are consistent with this interpretation.

4.6 Aggregate Effects

An advantage of the regression discontinuity in time approach is that it precisely identifies a local average treatment effect at the time of the critical event under consideration under relatively weak assumptions. However, insofar as effects are only identified at the boundary, interpreting their broader substantive implications requires additional assumptions. In the current application, the most relevant consideration – both in terms of cost to defendants

[25]African Americans comprise 6.5% of California's population, but 28.3% of California's incarcerated population and 21.4% of the defendants in our court data. This pattern may reflect racially disparate treatment at earlier stages in criminal justice (including the disparities in charge severity evidenced in Table 5) as well as the disparate impacts of facially neutral laws and procedures.

Figure 3
Assessing Indirect Racial Burden:
Effect of Petition Announcement on Sentence Length, by Race



Each panel displays left- and right-side intercepts (and associated 95% confidence intervals) from a MSE-optimal bandwidth regression discontinuity in time around the petition announcement.

and cost to the state of California – is a counterfactual one: how does the shift in judicial behavior following the petition announcement translate into additional days, months, or years of additional assigned prison time?

In this section we adopt four alternative approaches to guard against the possibility that the assumptions behind any one of them drives our conclusions. Likewise, rather than extrapolate over a prolonged period of time (in which, per our placebo tests above, a sequence of additional factors not pertaining to the Persky recall may have affected judicial punitiveness), we restrict ourselves to the 45 day window following the petition announcement (with the 45 day length approximating the optimal bandwidth from the RD estimates above).

The first approach is to assume that the identified LATE is the average treatment effect over the 45 day window. This approach assumes no decay or growth in the effect of the an-

nouncement on sentencing considerations. We proceed by multiplying the LATE estimate for the increase in raw sentence days, expressed as a percentages of a case's statutory maximum by the total number of cases in the 45 day window. We report results using the unadjusted LATE estimate and the estimate adjusted for judge- and offense-specific fixed effects.

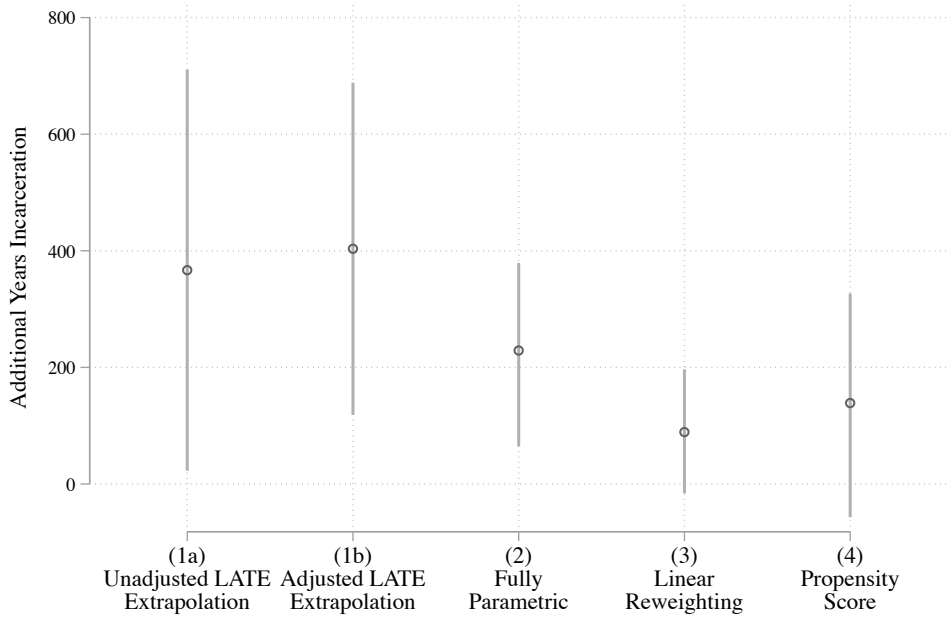
The second approach is to estimate a fully parametric regression model that adjusts for time trends before and after the announcement, and use the predicted values from that model to estimate the aggregate effect over the 45 day window. This approach may capture growth or decay in the effect over the interval following the announcement; however, it may also erroneously attribute factors unrelated to the announcement to the announcement itself. Because downward pre-announcement trends might artificially inflate anticipated sentencing differences, we constrain the trend to zero when predicting counterfactual sentences.

The third and fourth approaches employ estimators recommended by Angrist and Rokkanen (2015), which rely on a critical feature of the regression discontinuity design: that failure to control for the running variable (time) is the only source of omitted variables bias. Their approach to estimating average treatment effects away from the boundary is to test whether, conditioning on covariates, a relationship exists between the outcome and running variables; and if not, to estimate average treatment effects for an interval using either a linear reweighting or propensity score estimator. We report results using both approaches.

Figure 4 displays the estimated additional assigned incarceration (in years) for the counties in our sample using the four approaches. Point estimates suggest total effects of between 88 and 403 years total additional sentencing associated with the announcement. The larger figure comes from the adjusted LATE specification, and the smallest from the linear reweighting estimator. Note that the smaller estimates discard observations for which we lack covariate overlap pre- and post-treatment (e.g., sentences from the same judge both before and after the announcement), and are likely to be biased downward.

While the human cost to defendants is difficult to calculate without very strong assump-

Figure 4
Estimating Aggregate Effects of the Petition Announcement: 45 Day Window



Aggregate effect estimates in (1a) and (1b) assume the LATE is the ATE. (1a) reports the unadjusted ATE and (1b) reports the ATE adjusted for judge- and offense-specific fixed effects. Estimate (2) is the fully parametric aggregate effect, allowing for time trends before and after the announcement date and adjusting for judge- and offense-specific fixed effects. Estimates (3) and (4) employ two conditional independence assumption based estimators - a linear reweighting and a propensity score estimator respectively. Confidence intervals for (2), (3), and (4) are derived using the nonparametric bootstrap.

tions, a far easier calculation is total cost to the state: In 2016-17, the average annual cost of incarceration in California was $71 thousand per inmate. Using the most conservative 88 year estimate, our analysis suggests a total cost to the six counties in our sample of $6.25 million.[26] If the effect of the petition announcement persisted longer than the 45 day win-

[26]Defendants from the counties in our sample make up just 19% of the incarcerated population in the state. We make no claims respecting the representativeness of these counties. However, under fairly restrictive assumptions (most importantly, that the distribution of charges and the effect of the petition announcement are both uniform across the state) a back-of-the-envelope calculation using the most conservative estimate suggests that the total

dow under consideration, actual costs could be considerably higher. At the same time, one factor that may lower the total cost is parole, which was significantly expanded following the passage of Proposition 57 in November of 2016.

5 Discussion

In 2016, a Superior Court judge in California had not been recalled from office by voters in 84 years. But in the summer of that year, days after a Santa Clara County judge handed down a widely-publicized lenient sentence to an affluent, white defendant for a sexual assault and attempted rape conviction, an unanticipated recall campaign against that judge raised the threat of potential electoral sanctions for other judges. The research presented in this paper documents the far-reaching consequences of that threat for the criminal justice system. Using data from six California counties, we observe large, instantaneous increases in judicial punitiveness immediately following the announcement of the recall campaign, which are most readily apparent in sentencing for non-sexual violent crime. While we uncover no evidence that these instantaneous effects were larger for Minority than White defendants, we demonstrate that the petition announcement neither mitigated nor exacerbated observed longer-term racial disparities in sentencing.

The broader import of these findings – for our understanding of the criminal justice system in the United States and our understanding of electoral accountability – is twofold. First, they underscore the fact that even political campaigns targeting individual officeholders may have broad, unintended consequences. This is because such campaigns do not operate in a vacuum, and thus may alter the expectations of other officeholders that they themselves might be subject to such campaigns. The fact that we document no observable effects of the eventual recall election itself is consistent with this shift in beliefs to a "new normal" in the political environment of sitting trial judges, about which the ultimate (and widely-

effect statewide is 733 years, reflecting a total cost to the state of $52.1 million.

anticipated) electoral outcome conveyed no additional information. And critically, although the defendant in the precipitating case was White, and the crimes for which he was convicted were sex crimes, the use of the recall tool cannot be restricted to similar cases. And as such, neither can any anticipatory responses to that threat by judges in their courtrooms.

The events of the Persky recall campaign and the salience of law and order in the 2016 presidential campaign suggest that elected officials still face strong incentives to appear tough on crime. Elected judges and prosecutors may be ill-suited to lead efforts to reduce mass incarceration. Instead, sentencing reform may be more effective, including efforts to alter discretion by reducing statutory guideline severity, eliminating mandatory minimums, and employing more meaningful use of diversion programs. Moreover, given the composition of offenders in state prisons, fundamental changes to mass incarceration may require reducing imprisonment for serious and violent offenders (Gottschalk, 2016; Pfaff, 2017), precisely those cases for which electoral institutions, including the recall, incentivize greater severity.

Second, the research presented contributes to our understanding of the electoral incentives of public officials. We provide the first empirical evidence that the threat of recall affects the behavior of incumbent officials. In the current context, we provide evidence that an exogenous shock to judges' beliefs in the risk of recall affected their sentencing decisions. We document a substantial and immediate increase in sentencing severity following the highly-publicized announcement of a recall campaign, and calculate aggregate effects of that increase on the order of 88 years of additional incarceration for around 600 defendants in the 45 day period following the announcement. Insofar as we restrict our attention to a narrow window of time and only six counties, these estimates likely substantially underestimate the broader effects of this change in the behavior of these officials.

Finally, our analysis provides a roadmap for studying non-standard electoral institutions whose structure does not lend itself to standard research designs that exploit, inter alia, proximity to the next election, cross-sectional institutional variation, or term limits. This

is particularly valuable for an institution such as the recall, which, although widespread, is poorly understood. In the same vein, understanding the scope of incentive effects of recall efforts that vary in their intensity, and the political responses of incumbent officials,[27] is an important topic for future research.

References

Abrams, David S, Marianne Bertrand, and Sendhil Mullainathan. 2012. "Do judges vary in their treatment of race?" *The Journal of Legal Studies* 41 (2): 347–383.

Alesina, Alberto, and Eliana La Ferrara. 2014. "A Test of Racial Bias in Capital Sentencing." *American Economic Review* 104 (11): 3397–3433.

Alt, James, Ethan Bueno de Mesquita, and Shanna Rose. 2011. "Disentangling accountability and competence in elections: Evidence from US term limits." *The Journal of Politics* 73 (1): 171–186.

Angrist, Joshua D., and Miikka Rokkanen. 2015. "Wanna Get Away? Regression Discontinuity Estimation of Exam School Effects Away From the Cutoff." *Journal of the American Statistical Association* 110 (512): 1331–1344.

Bayer, Patrick, Randi Hjalmarsson, and Shamena Anwar. 2012. "The Impact of Jury Race in Criminal Trials." *The Quarterly Journal of Economics* 127 (2): 1017–1055.

Berdejo, Carlos, and Noam Yuchtman. 2013. "Crime, Punishment, and Politics: An Analysis of Political Cycles in Criminal Sentencing." *The Review of Economics and Statistics* 95 (3): 741–756.

Besley, Timothy, and Abigail Payne. 2013. "Implementation of anti-discrimination policy: does judicial selection matter?" *American Law and Economics Review* 15 (1): 212–251.

[27]For example, a group of Californian judges launched the Judicial Fairness Coalition shortly after the Persky recall (`https://www.caljudges.org/CommFairness.asp`), in part to provide resources for judicial officers facing recall threats.

Besley, Timothy, and Anne Case. 1995. "Does electoral accountability affect economic policy choices? Evidence from gubernatorial term limits." *The Quarterly Journal of Economics* 110 (3): 769–798.

Brace, Paul, and Brent D. Boyea. 2008. "State Public Opinion, the Death Penalty, and the Practice of Electing Judges." *American Journal of Political Science* 52 (2): 360–372.

Brace, Paul, and Melinda Gann Hall. 1995. "Studying courts comparatively: The view from the American states." *Political Research Quarterly* 48 (1): 5–29.

Bronson, Jennifer, and E Ann Carson. 2019. "Prisoners in 2017." *Office of Justice Programs Report.* Bureau of Justice Statistics (BJS), US Dept of Justice.

Butler, Paul. July 11, 2016. "Judicial Recall Will Inevitably Lead to Harsher Sentences.".

Calderone, Richard, Brandice Canes-Wrone, and Tom S. Clark. 2009. "Partisan Signals and Democratic Accountability: An Analysis of State Supreme Court Abortion Decisions." *Journal of Politics* 29 (2): 560–573.

Calonico, Sebastian, Matias D Cattaneo, and Rocio Titiunik. 2014. "Robust nonparametric confidence intervals for regression-discontinuity designs." *Econometrica* 82 (6): 2295–2326.

Canes-Wrone, Brandice, Tom S. Clark, and Jason P. Kelly. 2014. "Judicial Selection and Death Penalty Decisions." *American Political Science Review* 108 (1): 23–39.

Ferraz, Claudio, and Frederico Finan. 2011. "Electoral accountability and corruption: Evidence from the audits of local governments." *American Economic Review* 101 (4): 1274–1311.

Gelman, Andrew, and Guido Imbens. 2019. "Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs." *Journal of Business & Economic Statistics* 37 (3): 447–456.

Gersen, Jeannie Suk. June 17, 2016. "The Unintended Consequences of the Stanford Rape-Case Recall.".

Gordon, Sanford C., and Gregory A. Huber. 2007. "The Effect of Electoral Competitiveness

on Incumbent Behavior." *Quarterly Journal of Political Science* 2: 107–138.

Gottschalk, Marie. 2016. *Caught: The prison state and the lockdown of American politics*. Princeton University Press.

Greenblatt, Alan. September 17, 2020. "Due to Pandemic, Dozens of Governors and Mayors Face Recall Efforts.".

Grogger, Jeffrey, and Greg Ridgeway. 2006. "Testing for racial profiling in traffic stops from behind a veil of darkness." *Journal of the American Statistical Association* 101 (475): 878–887.

Hall, Melinda Gann. 1992. "Electoral Politics and Strategic Voting in State Supreme Courts." *Journal of Politics* 54 (2): 427–446.

Hausman, Catherine, and David S Rapson. 2018. "Regression discontinuity in time: Considerations for empirical applications." *Annual Review of Resource Economics* 10: 533–552.

Ho, Daniel E, and Kosuke Imai. 2006. "Randomization Inference With Natural Experiments." *Journal of the American Statistical Association* 101 (475): 888–900.

Huber, Gregory A., and Sanford C. Gordon. 2004. "Accountability and Coercion: Is Justice Blind When It Runs for Office?" *American Journal of Political Science* 48: 247–263.

Law Professors Statement. August 15, 2017. "Law Professors' Statement for the Independence of the Judiciary and Against the Recall of Santa Clara County Superior Court Judge Aaron Persky.".

Lim, Claire S. H. 2013. "Preferences and Incentives of Appointed and Elected Public Officials: Evidence from State Trial Court Judges." *American Economic Review* 103 (4): 1360–97.

Masket, Seth E. 2011. "The Circus That Wasn't: The Republican Party's Quest for Order in California's 2003 Gubernatorial Recall Election." *State Politics & Policy Quarterly* 11 (2): 123–147.

Matsusaka, John G et al. 2010. "Popular control of public policy: A quantitative approach." *Quarterly Journal of Political Science* 5 (2): 133–167.

Morton, Rebecca. 2006. *Analyzing Elections*. WW Norton.

Park, Kyung H. 2017. "The Impact of Judicial Elections in the Sentencing of Black Crime." *Journal of Human Resources* 52 (4): 998–1031.

Pfaff, John. 2017. *Locked in: The true causes of mass incarceration-and how to achieve real reform*. Basic Books.

Rehavi, M. Marit, and Sonja B. Starr. 2014. "Racial Disparity in Federal Criminal Sentences." *Journal of Political Economy* 122: 1320–1354.

Santa Clara County Association. June 14, 2016. "SCCBA Statement on Judicial Independence.".

Segura, Gary M, and Luis R Fraga. 2008. "Race and the recall: Racial and ethnic polarization in the California recall election." *American Journal of Political Science* 52 (2): 421–435.

Shaw, Daron, Mark J McKenzie, and Jeffrey Underwood. 2005. "Strategic voting in the California recall election." *American Politics Research* 33 (2): 216–245.

Smith, Malcolm. 1951. "The California Method of Selecting Judges." *Stanford Law Review* 3 (4): 571–600.

Spivak, Joshua. 2020. "Recall Elections in the US: Its Long Past and Uncertain Future." In *The Politics of Recall Elections*. Springer pp. 73–93.

Woolf, Nicky. June 24, 2016. "Stanford sexual assault: public defenders support judge in open letter.".

Online Appendix for "Incentive Effects of Recall Elections"

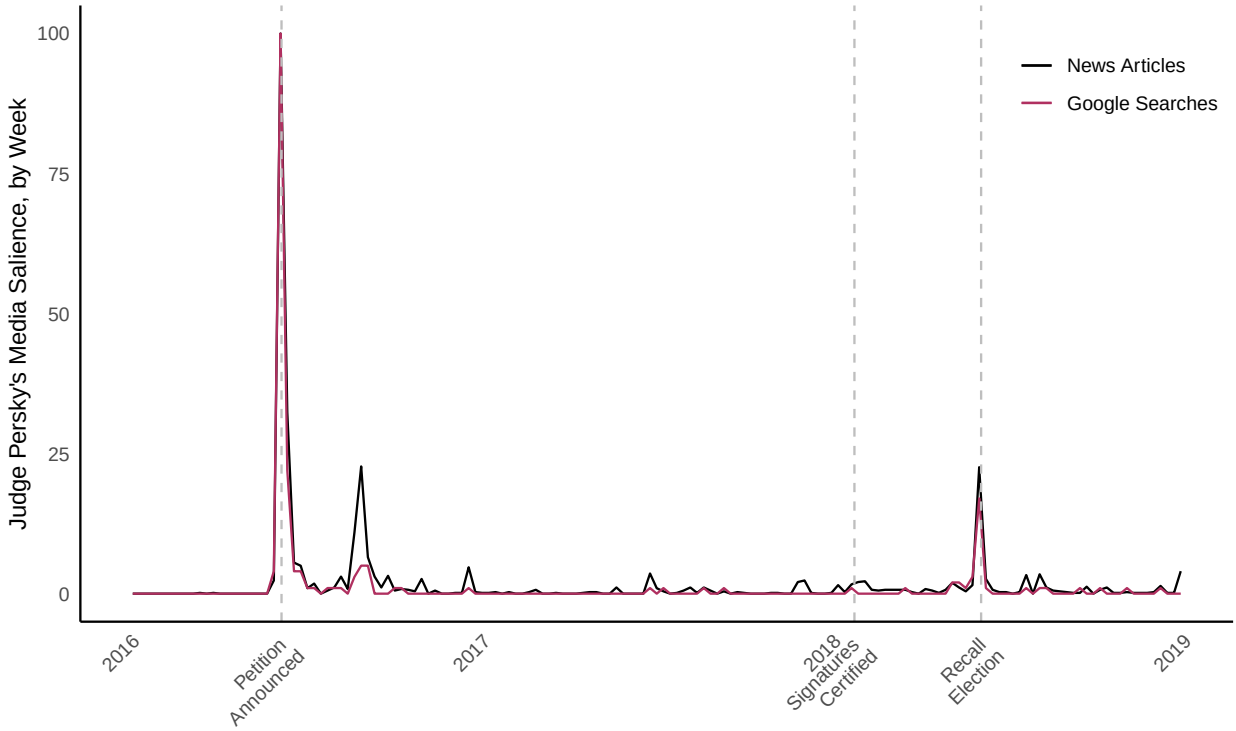
A Descriptive Analysis and Statistics

Table A.1 Descriptive Statistics

	All		Pre-&Post-Petition Announcement		Pre-&Post-Recall Election	
	Mean	SD	Mean	SD	Mean	SD
	Sentencing Characteristics					
Sentence Length (days)	559.165	1091.419	496.503	544.095	587.403	1039.722
Uncensored Normalized Sentence	0.417	0.764	0.385	0.383	0.415	0.423
Normalized Sentence	0.377	0.332	0.367	0.322	0.384	0.334
	Charge Characteristics					
Sex Crime	0.043	0.202	0.044	0.206	0.045	0.207
Violent Non-sex Crime	0.142	0.349	0.146	0.353	0.148	0.355
Non-violent Crime	0.827	0.378	0.817	0.387	0.822	0.383
	Defendant Characteristics					
Black	0.214	0.410	0.198	0.398	0.229	0.420
Hispanic	0.418	0.493	0.493	0.500	0.408	0.492
White	0.331	0.470	0.273	0.445	0.321	0.467
Male	0.831	0.375	0.861	0.346	0.837	0.369
Age	36.602	10.976	36.306	10.625	35.389	11.140
Num. Cases	19,832	19,832	1,476	1,476	1,383	1,383
Num. Defendants	18,293	18,293	1,434	1,434	1,342	1,342

This table presents summary statistics of our disposition data overall and within bandwidths relevant for the judicial recall campaign. Columns 3-4 report statistics for the sample restricted to within 45 days of the petition announcement date; Columns 5-6 report the corresponding values with respect to the recall election date.

Figure A.1
Judge Persky-related News Articles and Google Searches, by Week



Normalized Google searches and news article counts per week. Google searches refers to a Google trends search for the term "Judge Persky." News articles refers to 2,293 news articles collected from a Lexis-Nexis search for the term "Judge Persky." The normalization divides the count by the highest count, so that a value of 100 is the peak popularity between 2016-2018.

B Additional Robustness Checks Described in the Paper

Table B.1 Replication of Main Analysis Using Uncensored Normalized Sentences as Outcome

	Petition Announced		Recall Election	
RD estimate	0.137	0.162	-0.006	-0.011
	(0.058)	(0.052)	(0.068)	(0.051)
Left-side intercept	0.326	0.312	0.372	0.37
	(0.029)	(0.027)	(0.058)	(0.044)
Bandwidth	54.5	43.8	41.8	36.4
Judge fixed effects	N	Y	N	Y
Statute fixed effects	N	Y	N	Y
Effective observations	1533	1078	1176	919

The dependent variable in each column is the uncensored normalized sentence length (sentence length as a fraction of statutory maximum sentence). See notes in Table 1 for estimation details.

Table B.2 Replication of Main Analysis Using Non-Normalized Sentence Length as Outcome

	Petition Announced		Recall Election	
RD estimate	116.704	124.853	-60.407	-29.494
	(58.497)	(45.845)	(155.625)	(66.836)
Left-side intercept	387.113	374.297	589.261	583.316
	(39.902)	(31.134)	(135.467)	(57.366)
Bandwidth	62.3	50.2	51.1	38.9
Judge fixed effects	N	Y	N	Y
Statute fixed effects	N	Y	N	Y
Effective observations	1887	1244	1441	900

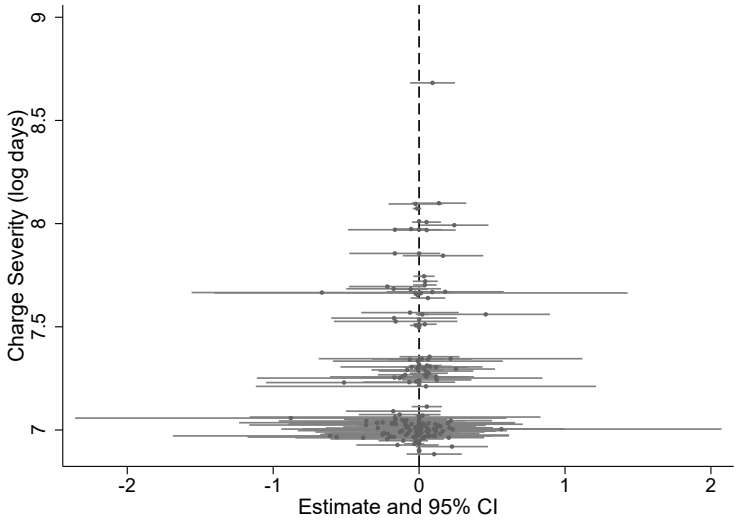
The dependent variable in each column is the sentence length (in days). See notes in Table 1 for estimation details.

Table B.3 Replication of Main Analysis Restricted to Cases with One Count

	Petition Announced		Recall Election	
RD estimate	0.085	0.104	0.019	0.036
	(0.042)	(0.035)	(0.056)	(0.048)
Left-side intercept	0.291	0.297	0.321	0.332
	(0.029)	(0.025)	(0.046)	(0.040)
Bandwidth	47.509	45.807	40.836	37.490
Adjusted	N	Y	N	Y
Effective observations	1421	1140	1108	872

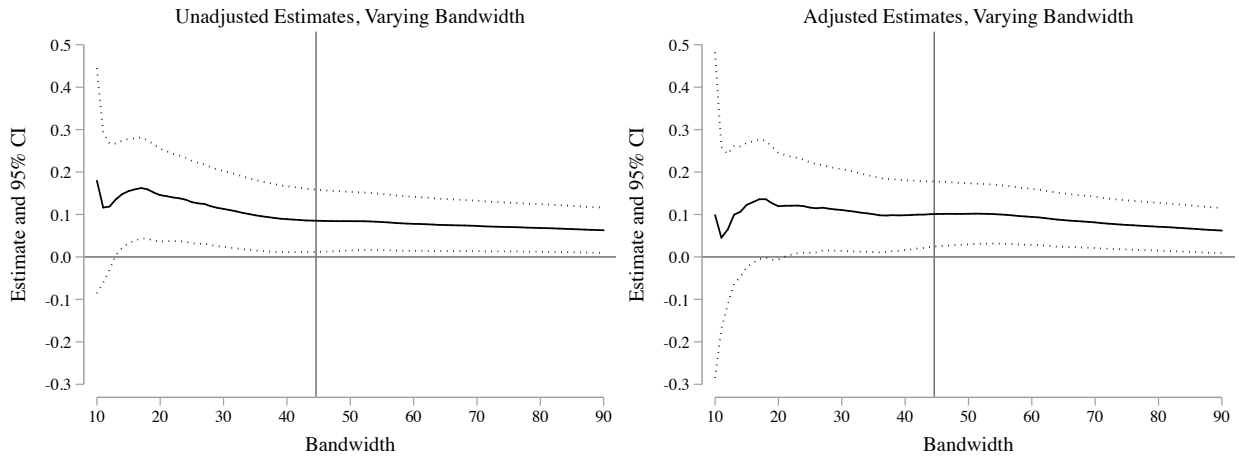
The dependent variable is the normalized sentence length (see text for description). Estimates in the second and fourth columns adjust for the number of counts and judge- and statute-specific fixed effects, and exclude Sacramento County (which does not report judge identifiers).

Figure B.1
Charge-FE RD Estimates and 95% Confidence Intervals



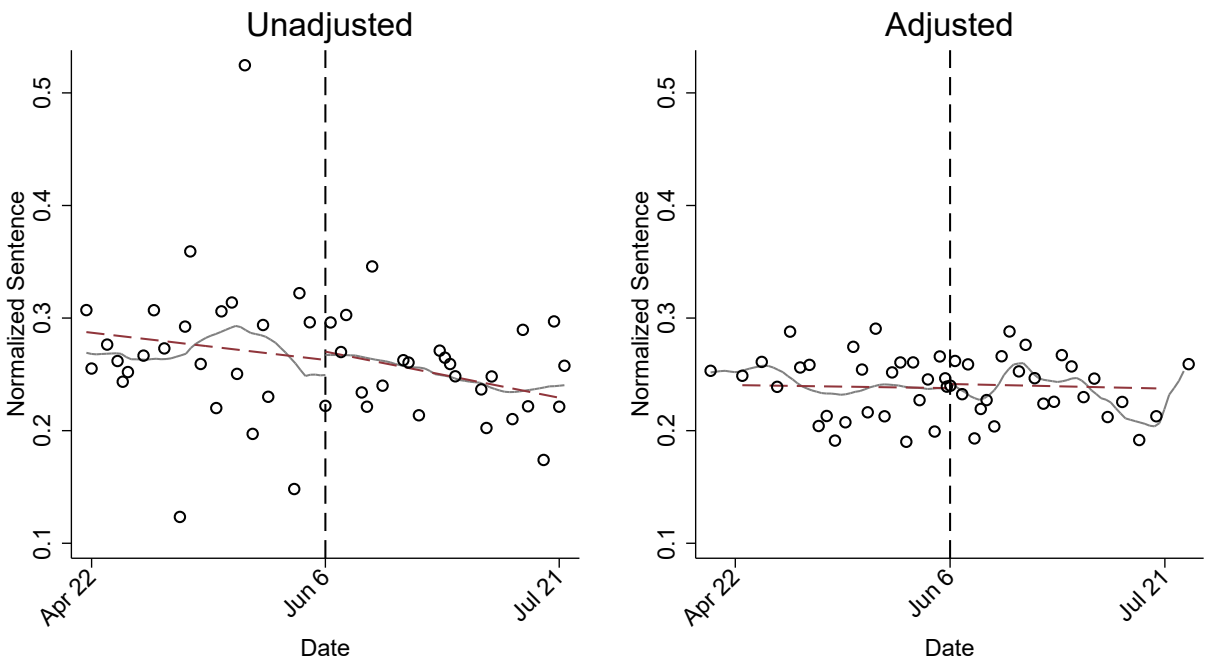
Each grey circle (and grey line) represents the RD estimate (and 95% confidence interval) associated with a unique crime's daily count.

Figure B.2
RD Estimates Varying Bandwidth



As in the main analysis, estimates employ triangular kernel, with standard errors clustered at the judge-charge level. The solid line denotes the MSE-optimal bandwidth.

Figure B.3
Effect on Sentencing of Petition Announcement in Washington State: Placebo Test



See notes in Figure 1.

C Additional Analyses

Heterogeneous Effects by Judicial Leniency. We stored the judge fixed effects from a regression of sentencing severity on statute and judge fixed effects (restricting attention to pre-announcement cases), and created a dummy indicator equal to one for judges with above median severity and zero for judges with below median severity. We then estimated the effect of the petition announcement, following the estimation procedures in the manuscript, and partitioning the sample into sentences handed down by more and less lenient judges.

Table C.1 reports the corresponding RD estimates. Contrary to our expectations, punitive judges (Column 1) were more responsive to the recall threat than lenient judges (Column 3), although the difference between these estimates is not statistically significant. We find a similar pattern in the adjusted specification (Columns 2 and 4).

Table C.1 Heterogeneous Effects of the Announcement: RD Estimates by Judicial Leniency

	Punitive		Lenient	
RD estimate	0.144	0.118	0.073	0.079
	(0.067)	(0.047)	(0.057)	(0.040)
Left-hand side intercept	0.361	0.355	0.253	0.253
	(0.046)	(0.033)	(0.038)	(0.029)
Bandwidth	71.506	65.571	86.242	95.160
Adjusted	N	Y	N	Y
Effective Observations	613	556	691	715

Dependent variable in each column is the normalized sentence length. Punitive judges issued above median sentences in the pre-announcement period, adjusting for statute fixed effects. Estimates in the second and fourth columns adjust for the number of counts and judge- and statute-specific fixed effects, and exclude Sacramento County (which does not report judge identifiers).

Appellate Judge Behavior. To assess the possibility that the appellate judges in California responded to the recall campaign (either directly or to changes in the behavior of trial judges), we collected appellate cases filed between January 2015 and January 2021 from the Judicial Branch of California's Appellate Court website. We were able to obtain records from all counties, with the exception of certain counties in 4th Appellate District (Inyo, Orange, Riverside, and San Bernardino Counties) for which the court interface does not permit searching by date.

We identified criminal appeals by searching cases that contained the string "The People" in the case name; we then removed any observations that the court system flagged as civil. This process produced 9,047 criminal appeals. Our outcome of interest is whether the defendant receives any relief from the appellate court. To generate this outcome, we classified decisions in which the court's ruling reversed, modified, or remanded a lower court decision. The final dependent variable is a binary outcome equal to one if either the defendant appeals and the appellate court reverses or if the prosecution appeals and the court affirms.

Approximately 20.65% of decisions are *pro-defendant* rulings according to this definition. The prosecution appeals a lower court decision in fewer than 1.2% of cases in our sample. The data also identify the appellate judges, which we used to construct 781 unique panels.[28]

Table C.2 presents estimates of the effect of petition announcement and recall election. In Columns (1)-(2), we find that appellate judges were 15-30% more likely to issue pro-defendant rulings immediately after the petition announcement. Only the latter estimate, which adjusts for panel fixed effects, is statistically significant at conventional thresholds. Columns (3)-(4) indicate that the recall election itself had a small and statistically indistinguishable effect on appellate judge behavior in either specification.

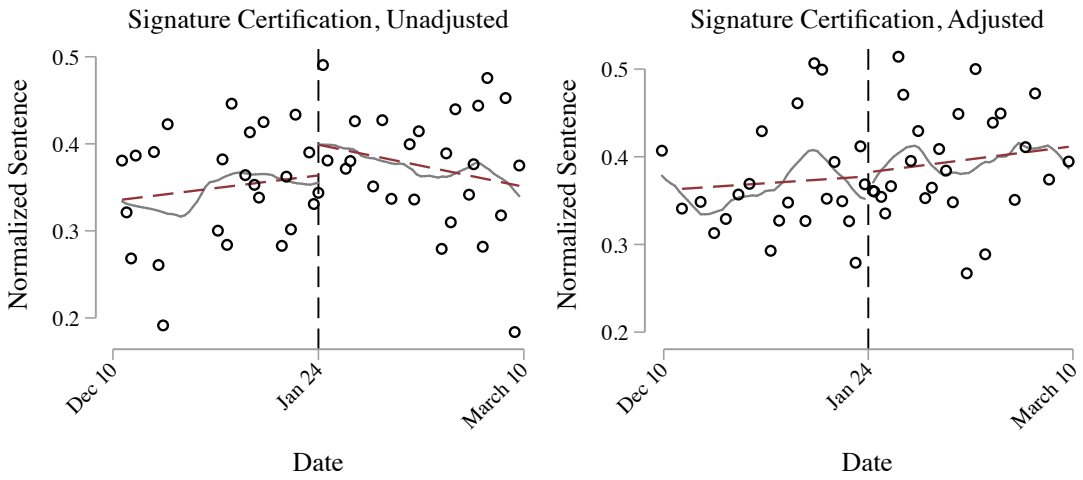
[28]In addition to information about the ruling, the appellate data also identify the sentencing judge, sentencing date, and sentencing case number. While these fields permit us to link our trial court data, fewer than 100 of our trial cases appear in the appellate data at all, and only two within the 45-day window around the petition announcement.

Table C.2 Effect of Critical Events on Prodefendant Rulings in Appellate Courts

	Petition Announced		Recall Election	
RD estimate	0.157	0.302	0.058	0.063
	(0.115)	(0.073)	(0.098)	(0.059)
Left-side intercept	0.114	0.079	0.095	0.086
	(0.077)	(0.043)	(0.051)	(0.036)
Bandwidth	66.4	65.6	88.3	79.6
Panel fixed effects	N	Y	N	Y
Effective observations	314	236	378	303

The dependent variable in each column is a prodefendant appellate outcome. Estimates in the second column and fourth columns adjust for appellate judge-panel fixed effects. Standard errors clustered on the appellate judge-panel level.

Figure C.1
Effect on Sentencing of Signature Certification Date in Persky Recall: Graphical Analysis



The left panel depicts average normalized sentence lengths (as tokens) in equally-sized bins. The panel to the right depicts binned means residualized using judge- and offense-specific fixed effects. Quadratic curve (maroon) and local polynomial smoothers (gray) are fit separately on each side of the signature certification date.