Does Judicial Capacity Affect Appellate Outcomes? Evidence from State Court Expansions

Parker Howell

April 07, 2026

Abstract

Appellate courts in the United States affirm lower court decisions in the vast majority of cases. If judges under heavy caseloads rely on this base rate as a decisional heuristic, expanding judicial capacity could improve the quality of appellate review and increase reversal rates. I construct a novel panel dataset combining over 700,000 published opinions from 24 state intermediate appellate courts (1970–2024) with hand-compiled data on legislative court expansions. Using continuous two-way fixed effects and event study estimators, I find no statistically significant effect of court expansions on reversal rates, though the estimates are imprecise and cannot rule out economically meaningful effects. This imprecision reflects the confluence of publication bias in the underlying database, few large expansion events, noisy text-based outcome classification, and unreliable judicial appointment dates. I document these data limitations and argue that the hypothesis remains viable but requires richer administrative data to test convincingly. The dataset and empirical framework developed provide a foundation for future work on judicial capacity and the quality of justice.

1 Introduction

The ability to appeal a decision perceived as unfair or incorrect is a central feature of legal systems across the developed world. Although the formal status of appellate rights varies—ranging from constitutionally embedded guarantees to statutory entitlements—nearly all advanced judicial systems provide litigants with at least one opportunity to seek review of lower-court decisions. In the United States, all 50 states and the federal system provide at least one appeal as a matter of statute.

The ability to appeal, however, is useful only insofar as the appeal is taken seriously. Here, again, the specific legal mechanisms used to ensure a fair appeal vary widely. In systems with weak judicial independence the right of appeal exists formally but offers little practical recourse.[1] The United States appellate system, by contrast, is designed to make reversal a live possibility: appellate judges serve long or life terms insulated from political removal, panels are assigned without executive direction, and published reversals carry binding precedential force.

In practice, appellants in the United States prevail[2] in fewer than 10 percent of cases (Edwards, 2019; Eisenberg, 2004), making it possible to correctly predict the outcome of a randomly selected appeal with high accuracy by always guessing that the lower court will be affirmed, without knowing anything about the merits of the case.

This paper asks whether this pattern is partly an artifact of judicial resource constraints. The hypothesis is straightforward: judges deciding appeals under heavy caseloads may implicitly rely on the low rate of affirmance as a shortcut. When an appellate court is expanded—that is, when the legislature adds judgeships—per-judge caseloads fall, potentially giving each judge more time to evaluate cases on their merits. If the rate of affirmance reflects, in part, a cognitive heuristic rather than the true distribution of meritorious appeals, then court expansions should increase the rate at which appellants prevail.

This hypothesis draws on a long tradition in behavioral economics documenting that decision-makers facing uncertainty rely on heuristics that, while efficient, can produce systematic errors (Tversky and Kahneman, 1974). The literature has extensively documented the influence of extralegal factors on judicial decisions (Cohen, 2025; Berdejó and Yuchtman, 2013; Holden et al., 2021; Eren and Mocan, 2025; Ash et al., 2025). Despite this, the effect of judicial workload on case outcomes remains largely unstudied.

To test this hypothesis, I construct a novel panel dataset combining two sources. The first is the CourtListener database maintained by the Free Law Project, which contains more than 99% of precedential appellate opinions in the United States. I extract the full text of over 700,000 published opinions from 24 state intermediate appellate courts spanning 1970 to 2024 to classify case types (e.g., civil or criminal) and outcomes (e.g., reversed or affirmed). The second source is a hand-compiled dataset of legislative court expansions—statutes that

[1] Solomon (2015) shows that Russian appellate courts, for instance, reverse criminal convictions in well under one percent of cases.

[2] I define prevailing as obtaining a reversal, remand, or vacatur of any part of the lower court's decision. I describe my classification approach in detail in Section 3.

add judgeships to specific courts—which I assembled primarily from state legislative records and secondary sources.

My lead empirical specification is a continuous two-way fixed effects (TWFE) regression of court-year reversal rates on the number of seats added by legislative action, with court and year fixed effects. This specification does not require an arbitrary threshold to define large expansions. Using the statute-authorized number of judges instead of the actual court capacity at any point in time gives this specification a natural interpretation as an intent-to-treat (ITT) approach. The ITT approach is less sensitive to fluctuations in court capacity due to events like the retirement or death of a sitting judge. I supplement this with heterogeneity-robust event study estimators following Sun and Abraham (2020), synthetic control methods, and sensitivity analysis across treatment definitions.

The results are null but imprecise. In my lead specification, each additional judgeship is associated with a 0.50 percentage point increase in the reversal rate (standard error = 0.41 percentage points, $p = 0.24$). The 95 percent confidence interval ranges from -0.31 to $+1.31$ percentage points per seat—a range that includes both zero and effects large enough to be economically meaningful. Binary event study estimates using the Sun and Abraham (2020) estimator are similarly inconclusive, and the estimated effects are sensitive to the definition of treatment: different thresholds for what constitutes a "large" expansion produce estimates that vary in sign, magnitude, and statistical significance. Federal circuit court results, where identification is more challenging due to the lack of large expansion events,[3] are also null.

I argue that these imprecise results reflect fundamental limitations of the available data rather than evidence against the hypothesis. Four limitations are particularly consequential. First, the CourtListener database contains a mix of published and unpublished opinions that varies across courts, creating cross-court heterogeneity in what the dependent variable measures and potentially introducing spurious trends if publication composition changes over time. Second, only four state appellate courts experienced expansions large enough to plausibly affect per-judge caseloads during the sample period, leaving very few treated units for identification. Third, the text-based outcome classifier I develop introduces measurement error in the dependent variable: approximately 6 percent of opinions cannot be classified, and manual validation suggests a disagreement rate of roughly 14 percent. Fourth, 68 percent of judges in the CourtListener database have appointment dates recorded only at the year level, introducing substantial noise in the timing of capacity changes.

To support my claim that poor data quality is the binding constraint, I replicate Huang (2011), a paper which shows that a surge immigration cases reduced reversal rates for civil appeals in affected circuits. For this, I use the same data used in that paper and CourtListener data. I recover the same estimates when I use Huang (2011)'s data, but not with the CourtListener data, even though CourtListener's coverage is much better at the federal level.

[3]The role of judges with senior status may also play a role at the federal level since the number of senior status judges affects the effective per-judge caseload. Because I observe senior status judges I am able to account for their impact. Even after accounting for senior status judges I do not observe any significant response in the reversal rate to changes in the effective number of judges. I omit this analysis since my focus in this paper is on state courts, noting only that this challenge is not relevant at the state level where very few states (e.g., Georgia) have a system whereby retired judges hear any cases at all.

Despite its limitations, this paper makes three contributions. First, I provide the first systematic empirical test of the relationship between judicial capacity and appellate outcomes using modern causal inference methods.[4] Second, I construct a novel dataset linking appellate outcomes to judicial capacity for 24 state courts over five decades, which can support future research on a range of questions about appellate court behavior. Third, I provide a detailed accounting of the data quality challenges that must be overcome to credibly test hypotheses about judicial capacity, offering a roadmap for future work with richer administrative data.

The remainder of the paper proceeds as follows. The next subsection reviews related literature. Section 2 describes the institutional setting. Section 3 presents the data and descriptive statistics. Section 4 develops the empirical strategy. Section 5 presents results. Section 6 discusses limitations and alternative interpretations. Section 7 concludes.

1.1 Related Literature and Contribution

This paper connects to several strands of literature. The first is the large body of work documenting extralegal influences on judicial decision-making. Credible causal evidence shows that judges' decisions are affected by racial bias (Arnold et al., 2018), the ideology of their appointing party (Cohen, 2025), proximity to retention elections (Berdejó and Yuchtman, 2013), the ideology of their law clerks (Bonica et al., 2019), social movements (Cai et al., 2025), the ideology and gender of their peers (Holden et al., 2021; Eren and Mocan, 2025), and even their own exposure to economic theory (Ash et al., 2025).[5] Harris and Sen (2019) and Rachlinski and Wistrich (2017) discuss judicial bias broadly. Despite this extensive literature, the effect of judicial workload on case outcomes has received little attention from economists.

A second strand studies how appellate courts respond to growing caseloads through institutional adaptations. Marvell (1989) documents that state appellate courts responded to caseload growth in the 1970s and 1980s through a combination of adding judges, restricting oral arguments, increasing reliance on staff attorneys, and issuing more unpublished opinions. Moody and Marvell (1987) find that the output capacity of trial courts, measured by the number of judges, has a strong impact on appellate filings. These studies describe the supply-side response of courts to demand pressure but do not estimate the causal effect of capacity changes on case outcomes.

The study closest to mine is Huang (2011), a law review article that examines whether a post-2001 surge of immigration appeals cases into the Second and Ninth Circuits reduced reversal rates for civil appeals in those circuits. Huang finds suggestive evidence that it did. My study differs in at least four ways: I examine capacity *expansions* (adding judges) rather than caseload *shocks* (adding cases), I employ modern econometric methods designed for

[4] While Huang (2011) documents that a surge of immigration cases into certain federal circuits depressed civil appeal reversal rates—a finding consistent with the hypothesis—that study examines an exogenous caseload *increase* rather than a capacity expansion, does not employ contemporary difference-in-differences methods, and restricts his sample to a very specific context.

[5] And possibly even the time since a judge's last meal break (Danziger et al., 2011), although this finding has not been replicated.

staggered treatment timing, I include all cases (as opposed to civil cases only), and I study state appellate courts in addition to federal circuits.

Third, this paper relates to the broader literature on how institutional resources affect the quality of public-sector decision-making. While I am not aware of prior work linking court capacity to appellate outcomes specifically, Black et al. (2023) document that aging federal judges—who may face cognitive resource constraints analogous to institutional capacity constraints—rely increasingly on cognitive shortcuts when interpreting law and casting votes. Bhuller and Sigstad (2025) show that Norwegian trial court judges update their behavior in response to appellate reversals, suggesting that the feedback channel between appellate and trial courts is active. If appellate court capacity affects case outcomes, this would provide a novel justification for court funding that extends beyond reducing backlogs to improving the quality of justice itself.

My contribution is to bring rigorous causal inference methods to a question that has been discussed qualitatively in the legal literature but has not been tested empirically with modern tools. The null result I obtain is informative not because it settles the question, but because it clarifies what data and research design would be needed to answer it convincingly.

2 Institutional Context

2.1 State Intermediate Appellate Courts

Every U.S. state has a court of last resort (typically called a supreme court), and 42 states also have one or more intermediate appellate courts (IACs) that hear the first appeal as of right from trial courts.[6] These intermediate courts handle the vast majority of appeals: in most states, the supreme court exercises discretionary review over only a fraction of cases that pass through the IAC. For the typical litigant, the IAC decision is effectively final.

State IACs vary considerably in size, structure, and jurisdiction. Some states operate a single statewide appellate court (e.g., Colorado, Virginia), while others divide the court into geographic divisions or districts (e.g., Illinois, which has five appellate districts). Court sizes in my sample range from 3 to 28 judges. Cases are typically heard by three-judge panels drawn from the court's active judges, with *en banc* review reserved for cases of exceptional importance.[7]

State appellate courts are substantially understudied in the economics literature relative to federal courts, despite handling a far greater volume of cases. Federal circuit courts collectively terminate approximately 50,000 appeals per year, while state appellate courts dispose of several hundred thousand.[8] The decisions of state IACs directly affect more litigants and shape more areas of law—including family law, criminal sentencing, and commercial disputes—than their federal counterparts.

[6] The remaining states route all appeals directly to the state supreme court.

[7] Several states either do not practice *en banc* review or explicitly prohibit it.

[8] See the National Center for State Courts, *Court Statistics Project*, and the Administrative Office of the U.S. Courts, *Table B-5* (Administrative Office of the U.S. Courts, 2024).

2.2 Court Expansions

A court expansion occurs when a state legislature enacts a statute creating new judgeships on an existing appellate court. Expansions are typically motivated by rising caseloads, although political considerations also play a role: the governor or appointing authority fills the new seats, creating an opportunity to influence the court's ideological composition. Importantly, the decision to expand is made by the legislature, not by the court itself, which provides some separation between the treatment and the treated unit.

The process from legislation to full capacity typically involves a lag. After a statute is enacted, the governor must nominate candidates, who then go through a confirmation or appointment process that varies by state. The new judges must then be assigned to panels, and appeals already in the pipeline continue to be decided by pre-expansion panels. For these reasons, I would not expect any effect on reversal rates to appear immediately at the time of the statutory expansion; a lag of one to two years is plausible.

Table 1: Court Expansions of Four or More Judgeships

State	Expansion Year	Seats Added	Seats Before	Seats After
Colorado	1974	4	6	10
Michigan	1974	6	12	18
Virginia	2021	6	11	17
Arizona	2022	6	22	28

This table lists the first expansion event for each treated court that experienced an increase of at least four judges. Colorado and Michigan also experienced subsequent smaller expansions. Arizona experienced five prior expansions of 1–3 seats each. Virginia had a 1-seat expansion in 2000.

Table 1 lists the four state appellate courts that experienced expansions of four or more seats during the sample period[9]. Two courts—Colorado and Michigan—were treated in 1974, while Arizona and Virginia were treated in 2021–2022. Several other courts in the sample experienced smaller expansions (one to three seats) that are used as variation in the continuous specification but fall below the binary treatment threshold.

Oregon's Court of Appeals, which doubled from 5 to 10 judges in 1977, is excluded from the treated group because this expansion coincided with the court's establishment in 1969—only six years of pre-treatment data are available, all from the court's startup phase, producing severe upward pre-trends that violate the parallel trends assumption. Oregon remains in the sample as a control court. Minnesota's Court of Appeals experienced a 6-seat expansion in 1984 that coincides with the court's first complete year in existence, leaving no pre-treatment observations; it is therefore excluded from the treated group but retained as a control. I discuss the inclusion or exclusion of other states in my sample in Section 3.2.

[9] I use this sample asthe treated group under my baseline binary specification

3 Data and Descriptive Statistics

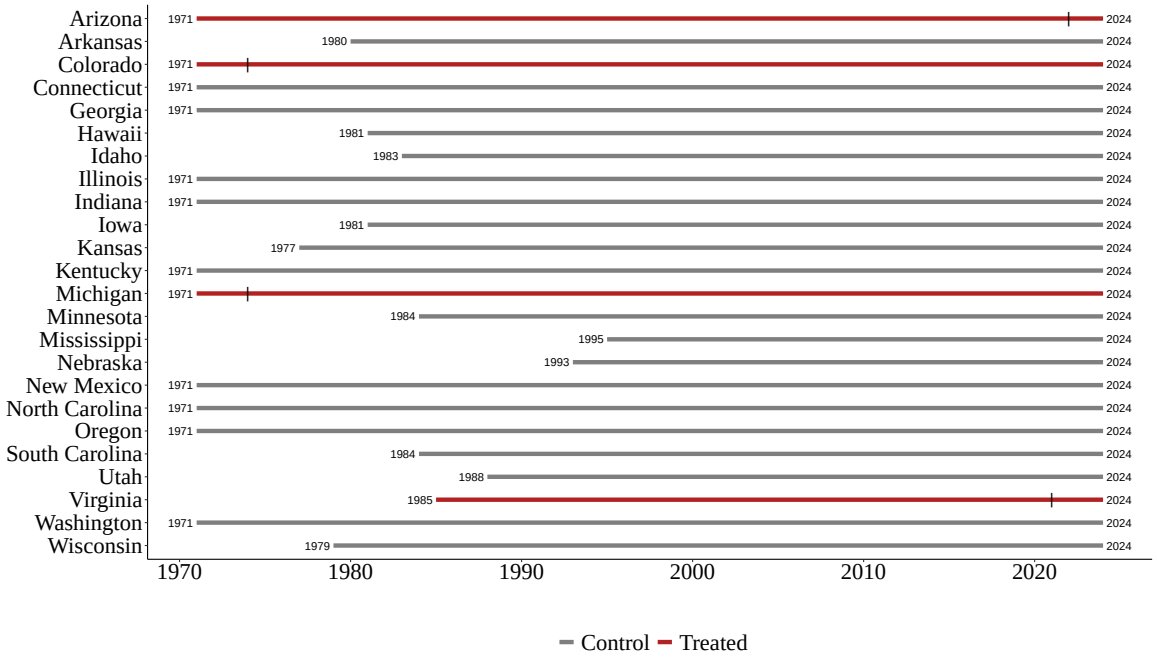
3.1 Data Sources

3.1.1 Appellate Opinions

The primary data source is the CourtListener database maintained by the Free Law Project, a nonprofit organization that aggregates legal data from court websites, PACER, and other public sources. CourtListener contains the full text of virtually all published (precedential) appellate opinions issued by U.S. state and federal courts, along with a variable and court-dependent collection of unpublished (non-precedential) opinions. I extract all available opinions from 24 state intermediate appellate courts, yielding approximately 713,000 opinions spanning 1970 to 2024.[10]

Figure 1 displays the years of data availability for each court. Coverage varies substantially: some courts (e.g., Illinois, Michigan) have opinions dating to the early 1970s, while others (e.g., Virginia, Hawaii) enter the database much later. I require a minimum of 25 published opinions per court-year for inclusion, and I drop the first and last observed year for each court to avoid partial-year artifacts.

Figure 1: Data Availability by Court



Notes: This figure shows the first and last year for each appellate court in the sample for which the CourtListener database contains at least 25 published opinions.

[10]The exact count varies by sample definition; this figure reflects all opinions—both published and unpublished—from courts in my analysis sample.

3.1.1.1 Publication Bias in the CourtListener Database An important caveat about the CourtListener data is that its coverage of published versus unpublished opinions varies substantially across courts. While CourtListener aims to capture the near-universe of published appellate opinions, many courts also issue large numbers of unpublished (non-precedential) decisions that CourtListener has ingested to varying degrees. Table 2 reports the breakdown of published and unpublished opinions in the CourtListener database for each court in the sample. Seven courts appear to have only published opinions in the database—most likely because CourtListener has not yet ingested their unpublished dispositions—while others have substantial unpublished holdings. Virginia, one of the four treated courts, has only 33 percent of its CourtListener opinions tagged as published; Michigan has 61 percent.

Table 2: Publication Status of Opinions in the CourtListener Database

State	Total in CL	Published	Non-Published	% Published
Virginia*	19,207	6,308	12,899	32.8
South Carolina	18,162	6,276	11,886	34.6
Idaho	10,332	5,075	5,257	49.1
Colorado*	29,055	15,132	13,923	52.1
New Mexico	14,536	8,262	6,274	56.8
Michigan*	48,506	29,355	19,151	60.5
Arizona*	25,002	16,457	8,545	65.8
Kentucky	15,192	11,629	3,563	76.5
Minnesota	16,128	12,669	3,459	78.6
Kansas	14,856	11,789	3,067	79.4
Wisconsin	18,924	15,456	3,468	81.7
Washington	86,198	71,955	14,243	83.5
Nebraska	6,241	5,230	1,011	83.8
Illinois	110,773	93,874	16,899	84.7
Indiana	56,019	51,127	4,892	91.3
Oregon	45,184	43,007	2,177	95.2
North Carolina	45,003	43,781	1,222	97.3
Arkansas	13,551	13,551	0	100.0
Connecticut	20,493	20,493	0	100.0
Georgia	84,926	84,926	0	100.0
Hawaii	16,123	16,123	0	100.0
Iowa	21,802	21,802	0	100.0
Mississippi	13,947	13,947	0	100.0
Utah	7,661	7,661	0	100.0

Counts reflect all opinions in the CourtListener database for each court, before sample restrictions. "Published" refers to opinions tagged as precedential in CourtListener; "Non-Published" includes opinions tagged as unpublished, unknown, or errata. Courts marked with * are treated under the baseline binary specification. Courts showing zero non-published opinions may reflect incomplete ingestion of non-precedential decisions or simply that those states publish nearly all opinions, regardless of precedential status.

This variation matters for my analysis in two ways. First, because I include all opinions regardless of precedential status in my analysis sample, the composition of published versus unpublished opinions differs across courts.[11] Since published opinions disproportionately represent reversals—courts are more likely to write and publish an opinion when they reverse

[11]Restricting to published-only opinions would discard large fractions of the data for courts like Virginia, and the "published'' tag itself may not be consistently applied across courts and time periods.

a lower court than when they affirm—courts with a higher share of unpublished opinions in the database will mechanically have lower measured reversal rates, all else equal. Second, if a court's publication practices change over time—for example, if CourtListener begins ingesting unpublished opinions from a particular court at a certain date—this could create spurious trends in the measured reversal rate that coincide with, but are unrelated to, court expansions.

I cannot fully resolve this issue with the available data. While the continuous TWFE specification (which relies on within-court over-time variation) is robust to cross-court differences in publication composition, results may be severely biased if each court's composition of published versus unpublished opinions changes over time within court. I return to this limitation in Section 6.

3.1.2 Case Type Classification

I classify cases as civil or criminal using a two-stage approach. In the first stage, I apply regular expressions to case captions, disposition metadata, and opinion text to identify cases with unambiguous markers—for example, case names beginning with "People of the State of" are classified as criminal, while opinions containing terms like "summary judgment" or "breach" are classified as civil. In the second stage, I train a penalized logistic regression (LASSO) on term-frequency features from the first-stage labels and use it to classify remaining unlabeled opinions, retaining only predictions with high confidence (predicted probability ≥ 0.70). This two-stage pipeline classifies approximately 85 percent of opinions. Among classified cases, roughly two-thirds are criminal and one-third are civil, broadly consistent with the preponderance of criminal appeals on state appellate dockets.[12]

Because this classification is not integral to the success of this paper, I considered it a poor use of time and funds to use an LLm to help increase the classification rate above 85%. However, as a proof of concept to ensure the remaining 15 percent of unclassified cases could be classified relatively easily, I selected a random sample of five unclassified opinions to, submitting the full text of these opinions via API to be classified by Anthropic's Claude Opus 4.6 model. I then manually reviewed each of these opinions, agreeing with the LLM's classification in all five cases.

3.1.3 Outcome Classification

I consider a remand, vacatur, or reversal of any part of the lower court's decision a win for the appellant. If the opinion affirms on all arguments or dismisses the appeal entirely, that is a loss for the appellant.

I leverage the full text of every part of every opinion to determine whether the appellant won or lost. The CourtListener database includes the full text of concurrences, dissents, and majority opinions. When different opinion types are uploaded separately, classification is

[12]See the National Center for State Courts, *Court Statistics Project*, for national data on state appellate caseload composition.

easier since only the majority opinion is relevant. When they are combined, extra caution is required to avoid misclassifying cases in which, for example, a dissenting judge says, "I would reverse" even though the majority affirms.

I classify dispositions using a multi-pass regular expression approach applied to the full text of each opinion. For each opinion, I search for disposition keywords—*affirm*, *reverse*, *vacate*, *remand*, and *dismiss*—in three successive passes with expanding search windows: the first 1,000 characters, the last 1,000 characters, and the last 5,000 characters of the opinion text. Dispositions are most commonly stated near the end of an opinion, so the backward-looking passes capture the majority of cases. A case is classified as a win for the appellant if the extracted disposition contains reversal, remand, or vacatur language, and as a loss if it contains only affirmance or dismissal language. Approximately 6 percent of opinions cannot be classified because the disposition language is absent or ambiguous.

This automated classification introduces measurement error. I constructed a validation sample of 100 randomly selected opinions and compared my automated classification to a manual reading. The automated classifier disagreed with my manual classification in approximately 14 percent of cases. As with the case type classification, considerable improvement is possible here, especially with the use of modern large language models. As before, I have constructed a pipeline through which I could reduce the number of mis- or unclassified opinions, but I have not followed through since I believe that the marginal benefit of such an exercise is small given the other challenges to identification in this setting. I discuss the implications of this measurement error for my estimates in Section 6.

3.1.4 Judicial Capacity

I measure judicial capacity—the number of active judges on each court—using CourtListener's database which records the start and end dates of each judge's service. I count a judge as active beginning in the month they start service and ending when they retire, resign, or transfer to a different court. I then aggregate to court-year (or court-quarter) totals.

This measure has important limitations. First, 68 percent of judges in the database have appointment dates recorded at only the year level. This introduces substantial measurement error in the timing of capacity changes, particularly at sub-annual frequencies. Second, the database is incomplete for some courts: Iowa, Oregon, and Washington have judge records that are sufficiently unreliable that I exclude them from specifications using observed capacity (while retaining them for ITT specifications that rely on the hand-compiled expansion data).

3.1.5 Court Expansion Data

I hand-compiled a dataset of legislative court expansions from state legislative records, court histories, and secondary sources (primarily Wikipedia and state court websites). For each expansion event, I record the court, the statute date, the number of seats added, and the number of seats before the expansion. This intent-to-treat (ITT) measure has the critical advantage of being unaffected by gaps or errors in the CourtListener judge database: it

captures when the legislature authorized new positions, regardless of whether CourtListener accurately records when those positions were filled.

The distinction between ITT and observed capacity measures is important throughout this paper. The ITT measure (seats added by statute) is clean but captures only legislatively authorized changes, not actual staffing. The observed measure (year-over-year change in active judges from CourtListener) captures actual staffing but inherits all the noise in the judge database. I present results using both measures, with the ITT specification as the lead.

3.2 Sample Construction

I begin with every published opinion at the intermediate appellate level for each of the 50 U.S. states.

I exclude states from the sample for one of three broad reasons (see Table A7 in the appendix for a complete list). First, I exclude states where the intermediate appellate court is organized into multiple geographically independent districts, each with its own judges and caseloads—California, Florida, New York, Texas, and Missouri. Since I cannot reliably observe district-level judicial capacity, treating these multi-district systems as single courts would introduce severe measurement error. Second, I exclude states with separate civil and criminal appellate courts (Alabama, Oklahoma, Tennessee), where an expansion of one court would not affect caseloads on the other and the treatment is therefore not comparable to a unified-court expansion. Third, I exclude states with institutional features that make them noncomparable: Louisiana (the only U.S. civil law jurisdiction), Maryland (whose intermediate court was originally limited to criminal cases), Nevada (which uses a "deflective" model in which the supreme court selectively assigns cases to the court of appeals), and Pennsylvania (which has two statewide intermediate courts with overlapping jurisdiction). Several additional states are excluded because they publish few precedential opinions.

After restricting to the subset of states with comparable state systems I am left with over 700K opinions across 24 courts.[13] I also drop the first year for each court to avoid partial years. The resulting panel is unbalanced, with courts entering and exiting the sample as data availability dictates. For analysis, I aggregate from the opinion level to the court-period level (court-year or court-quarter), computing the mean reversal rate and total case count for each court-period cell. This yields an unbalanced panel of approximately 1,100 court-year observations across 24 courts.

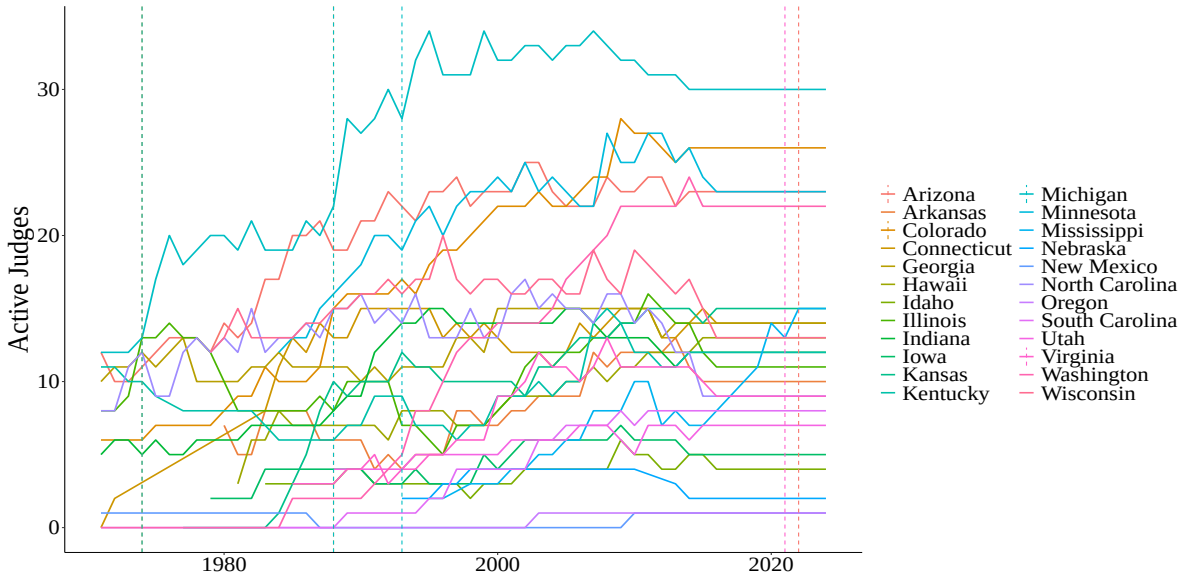
3.3 Descriptive Statistics

Figure 2 displays the number of active judges on each court over time, per the CourtListener database (i.e., not according to statute). The figure illustrates both the variation in court

[13]The 24 courts are: Arizona, Arkansas, Colorado, Connecticut, Georgia, Hawaii, Idaho, Illinois, Indiana, Iowa, Kansas, Kentucky, Michigan, Minnesota, Mississippi, Nebraska, New Mexico, North Carolina, Oregon, South Carolina, Utah, Virginia, Washington, and Wisconsin.

sizes and the expansion events that I exploit for identification. The four treated courts (Colorado, Michigan, Virginia, Arizona) are visible as courts experiencing discrete jumps in capacity.

Figure 2: Number of Active Judges by Court and Year



Notes: This figure shows the number of active judges in each of the 24 state intermediate appellate courts in the sample. A judge is counted as active beginning in the month of appointment and ending at retirement, death, or transfer.

Table A6 presents opinion-level descriptive statistics. The overall reversal rate among classifiable opinions is approximately 27 percent. Criminal cases account for roughly 56 percent of classified opinions and civil cases for 29 percent, with 15 percent unclassified. The mean opinion length is approximately 15,700 characters (or slightly more than ten double-spaced pages, assuming approximately 3000 characters to page.).

Table 3 compares pre-treatment characteristics of treated and control courts.[14] Because the unit of treatment assignment is the court, I first average each variable within court across pre-treatment years and then compute normalized differences (the difference in group means divided by the root-mean-square of the two standard deviations) following Imbens and Rubin (2015). Values below 0.25 in absolute value are conventionally considered well-balanced.

[14]Table A5 reports summary statistics for the court-year panel for the combined sample. The mean court-year reversal rate is 28 percent, with substantial cross-court and over-time variation (standard deviation of 10 percentage points). Courts publish a median of roughly 500 opinions per year, though this varies widely from 10 to over 3,800.

Table 3: Balance Table: Pre-Treatment Court-Level Means

	Treated		Control		
	Mean	SD	Mean	SD	Norm. Diff.
Reversal Rate	0.27	0.14	0.28	0.07	-0.08
Opinions Published	556.92	382.47	597.36	467.69	-0.09
Appointed Judges	12.75	6.70	12.81	10.13	-0.01
Active Judges	12.50	7.42	9.39	5.60	0.47
Opinions/Judge	43.26	23.51	62.95	42.86	-0.57
Year	1995.75	27.43	2000.31	10.70	-0.22
Number of courts	4		24		
Court-year observations	12		932		

Each court contributes one observation (its pre-treatment mean). Control includes both never-treated and not-yet-treated courts. The normalized difference is $(\bar{X}_T - \bar{X}_C)/\sqrt{(S_T^2 + S_C^2)/2}$; values below 0.25 in absolute value suggest adequate balance (Imbens and Rubin 2015).

The outcome of interest, the reversal rate, is well-balanced: treated courts average 27 percent versus 28 percent in the control group ($\Delta = -0.06$). Opinions published per year are similarly close ($\Delta = -0.04$). The number of appointed (statutory) judgeships—compiled from legislative records independently of the CourtListener database—is also well-balanced, with treated courts averaging 13 judges in both groups ($\Delta = -0.01$). This indicates that treated and control courts are of similar authorized size prior to the expansion events I study.

The number of *active* judges as recorded in the CourtListener database, however, shows imbalance ($\Delta = 0.49$), with treated courts appearing to have more active judges (13 vs. 9). Because statutory capacity is balanced, this discrepancy likely reflects differences in how completely CourtListener records judge tenures across courts rather than a genuine difference in court size. The opinions-per-judge variable inherits this measurement issue ($\Delta = -0.55$). In any case, level differences in court size do not threaten identification: court fixed effects absorb time-invariant characteristics, and the parallel trends assumption concerns trends, not levels.

The mean pre-treatment year also differs ($\Delta = -0.27$), reflecting the fact that the two earliest-treated courts (Colorado and Michigan, both treated in 1974) contribute pre-treatment observations from the 1960s and early 1970s, pulling the treated-group mean earlier. This imbalance is mitigated by the fact that each treated court's pre-treatment observations are compared to contemporaneous control observations.

Appendix 1 provides additional descriptive statistics, including opinion counts by court-year (Figure A6), a balance table comparing treated courts to never-treated courts only (Table A8), and the relative lack of seasonal patterns in opinion publication (Figure A7).

4 Empirical Strategy

4.1 Identification

The identifying idea is to compare reversal rates before and after a court expansion in courts that experience expansions, using courts without contemporaneous expansions as controls. The key assumption is that, conditional on court and year fixed effects, the timing of court expansions is independent of trends in appellant outcomes. Under this assumption, courts that did not expand provide a valid counterfactual for the trajectory that expanding courts would have followed absent the expansion.

Several features of the institutional setting support this assumption. Court expansions are legislated, typically in response to growing caseloads and backlogs, not in response to trends in reversal rates. The legislative process is slow, often spanning multiple years, so the precise timing of an expansion reflects political dynamics (e.g., budget cycles, partisan control) rather than short-run changes in case outcomes. Moreover, since the outcome I study—the reversal rate—is not a salient metric in legislative debates about court expansion, it is unlikely that legislators time expansions to coincide with changes in this variable.

The assumption could fail if the political factors that drive expansions also directly affect how judges decide cases. For instance, if a political party that favors appellants also tends to expand courts, then expansion and higher reversal rates could be spuriously correlated. Court and year fixed effects absorb time-invariant court characteristics and common shocks, but time-varying confounders correlated with both expansion timing and reversal rate trends would bias my estimates. I address this concern through pre-trend tests (Section 5) and by noting that the parallel trends assumption is more plausible at shorter horizons and for the ITT specification, which captures the legislative decision rather than the possibly endogenous timing of actual appointments.

A distinct identification concern relates to the mechanism through which expansions might affect reversal rates. My hypothesis is that additional judgeships reduce per-judge caseloads, giving incumbents more time per case and thereby improving the quality of review. But court expansions also change the composition of the bench: the newly appointed judges may differ systematically from incumbents in ideology, experience, or judicial philosophy. If new judges are more (or less) reversal-prone than the judges they join, the expansion could change the court's aggregate reversal rate through this composition channel regardless of any caseload effect. The cleanest test of the caseload hypothesis would restrict the sample to cases decided entirely by judges who were on the court before the expansion—if those incumbent judges reverse more often after the expansion, that would isolate the caseload mechanism from the composition channel.

I do not perform this restriction in this paper. Because CourtListener does not systematically record which judges participated in each decision, I would need to extract judge names from the text of each opinion. This is possible, and I made initial steps toward doing this systematically, but because data limitations made it clear that I would not be able to fully test my hypothesis, I opted to stop before completing this exercise. As a result, the estimates in this paper should be interpreted as the reduced-form effect of court expansions

on aggregate reversal rates—combining any caseload effect with compositional effects of new appointments—rather than as identified estimates of the caseload mechanism alone. This is an important qualification: even if the parallel trends assumption holds, my estimates do not isolate the hypothesized channel.

4.2 Specifications

4.2.1 Lead Specification: Continuous TWFE

My lead specification is a continuous two-way fixed effects regression:

$$Y_{ct} = \beta \cdot \text{SeatsAdded}_{ct} + \alpha_c + \lambda_t + \varepsilon_{ct}, \tag{1}$$

where Y_{ct} is the reversal rate (share of opinions classified as reversals) in court c and year t; SeatsAdded_{ct} is the number of judgeships added to court c by legislative action in year t (the ITT measure); α_c are court fixed effects; λ_t are year fixed effects; and ε_{ct} is an error term clustered at the court level.

The coefficient β is interpretable as the change in the reversal rate associated with each additional legislatively authorized judgeship. This specification has two advantages over binary treatment definitions. First, it uses all variation in expansion size, including small expansions that would fall below any reasonable binary threshold. Second, it avoids the researcher degree of freedom involved in choosing a threshold for "large" versus "small" expansions—a choice to which my results are sensitive (Section 5.3).

One caveat is in order: as Callaway et al. (2024) discuss, continuous TWFE may not recover a well-defined causal parameter when treatment effects are heterogeneous across dose levels or over time. In my setting, treatment effect heterogeneity is plausible: the marginal effect of adding a sixth judge to a 5-judge court likely differs from adding a sixth judge to a 22-judge court. The continuous TWFE coefficient should therefore be interpreted as a weighted average of marginal effects, with weights determined by the variation in SeatsAdded_{ct}.

As a robustness check, I also estimate Equation 1 replacing SeatsAdded_{ct} with ΔJudges_{ct}, the year-over-year change in the number of active judges observed in the CourtListener database. This observed measure captures actual staffing changes but inherits measurement error from the judge database.

4.2.2 Event Study Specification

To examine the dynamic path of treatment effects and test for pre-trends, I estimate a heterogeneity-robust event study using the interaction-weighted estimator of Sun and Abraham (2020):

$$Y_{ct} = \sum_{k \neq -1} \beta_k \cdot 1\{t - T_c = k\} + \alpha_c + \lambda_t + \varepsilon_{ct}, \tag{2}$$

where T_c denotes the first year in which court c is treated (i.e., experiences an expansion of ≥ 4 seats, the baseline binary threshold); $1\{t - T_c = k\}$ is an indicator equal to one when year t is k periods from treatment for court c; and $k = -1$ is the omitted reference period. The specification includes court fixed effects α_c and year fixed effects λ_t, with standard errors clustered at the court level.

The Sun and Abraham (2020) estimator avoids the bias that arises in conventional TWFE event studies when treatment effects are heterogeneous across cohorts, by using only never-treated and not-yet-treated units as controls for each cohort-period combination. This is important in my setting because the two treatment cohorts (1974 and 2021–2022) are separated by nearly five decades, during which the legal and institutional environment changed substantially.

The pre-treatment coefficients β_k for $k < -1$ test the parallel trends assumption: under the null of no differential pre-trends, these coefficients should be jointly indistinguishable from zero. The post-treatment coefficients β_k for $k \geq 0$ trace out the dynamic treatment effect. I expect any effect to emerge with a lag of one to two years, reflecting the time required for new judges to be appointed, assigned to panels, and for the resulting opinions to be published.

The average treatment effect on the treated (ATT) is computed by averaging the post-treatment coefficients, excluding the treatment year itself ($k = 0$) and the year immediately preceding it ($k = -1$). I exclude these periods since treatment effects should not appear for approximately a year after the expansion due to the typical delay (often exceeding one year) between the time a judge is assigned a case and the date the case is published.

I estimate this specification using a 3-year pre-treatment and 5-year post-treatment window at the annual frequency, and a 12-quarter pre / 12-quarter post window at the quarterly frequency.

4.2.3 Synthetic Control

As an additional robustness check, I implement a stacked synthetic control design following Arkhangelsky et al. (2021). For each treated court, I construct a synthetic control from the pool of never-treated and not-yet-treated courts that best matches the treated court's pre-treatment reversal rate trajectory. The stacked design aggregates across treated courts to produce a pooled estimate. I present these results in Appendix 5.

4.3 Inference

With 24 courts and only 4 treated, conventional cluster-robust standard errors may be unreliable due to the small number of clusters. To address this, I report wild cluster bootstrap confidence intervals and p-values for the Sun–Abraham event study specifications, using the Rademacher weight distribution and imposing the null hypothesis (Webb, 2023). Specifically, I follow the approach put forth and implemented by Roodman et al. (2019) and MacKinnon

et al. (2023) to construct 95 percent confidence intervals and to conduct a joint test that all pre-treatment coefficients equal zero.

For the all specifications, I report standard errors clustered at the court level.

5 Results

5.1 Continuous TWFE

Table 4 presents results from the lead specification, Equation 1. Column (1) reports the main estimate: each additional legislatively authorized judgeship is associated with a 0.50 percentage point increase in the reversal rate. The estimate is positive, consistent with the hypothesis that judicial capacity improves the quality of appellate review, but it is not statistically significant ($p = 0.24$). The 95 percent confidence interval, $[-0.31, +1.31]$ percentage points, cannot rule out either a zero effect or an effect large enough to be substantively important. To put the magnitude in context: at the baseline reversal rate of 27 percent, a 0.50 percentage point increase represents a roughly 2 percent relative increase per seat added. A typical large expansion of 6 seats would imply a 3 percentage point (11 percent relative) increase, but with wide uncertainty.

Column (2) uses the observed year-over-year change in active judges from the CourtListener database instead of the ITT measure. The point estimate is slightly smaller (0.41 percentage points) and less precise ($p = 0.43$), consistent with attenuation from measurement error in the judge database. Columns (3) and (4) report the ITT specification separately for civil and criminal cases. Both estimates are positive but statistically insignificant, with the civil estimate (0.45 pp) slightly larger than the criminal estimate (0.36 pp).

Table 4: Effect of Judicial Capacity on Reversal Rates (Continuous TWFE)

	All (ITT)	All (Obs.)	Civil (ITT)	Criminal (ITT)
Seats Added (ITT)	0.003		0.003	0.002
	(0.003)		(0.003)	(0.004)
Δ Active Judges (Obs.)		0.004		
		(0.005)		
Observations	951	951	949	951
Adj. R-Squared	0.454	0.455	0.253	0.458
Court FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes

* p $<$ 0.1, ** p $<$ 0.05, *** p $<$ 0.01

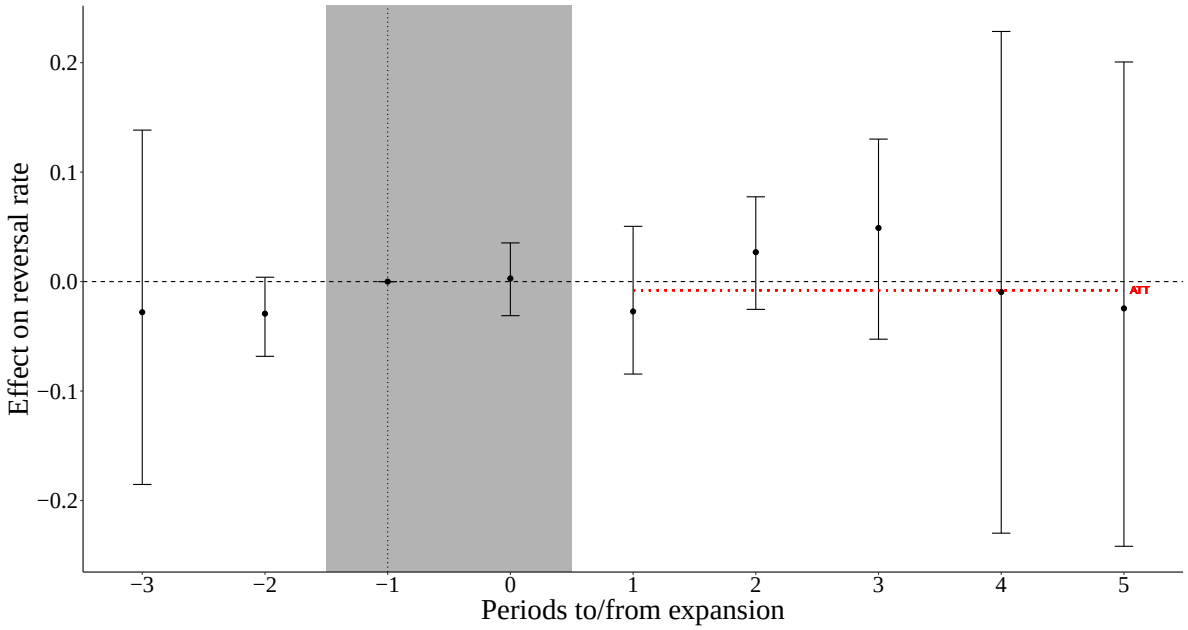
Each column reports the coefficient from a regression of court-year reversal rate on the indicated capacity measure, with court and year fixed effects. The dependent variable is the share of opinions classified as reversals, remands, or vacaturs. ITT columns use the number of legislatively authorized seats from hand-compiled expansion data. The "Observed" column uses the year-over-year change in active judges from the CourtListener judge database. Standard errors clustered by court in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

5.2 Event Study

Figure 3 presents event-study coefficients from the Sun and Abraham (2020) estimator for all case types at the annual frequency, under the baseline binary treatment definition (≥ 4 seats added). The shaded gray region highlights periods -1 and 0, which are excluded from the ATT calculation. The horizontal red line indicates the estimated ATT over periods 1 through 5.

The pre-treatment coefficients are mostly small and statistically indistinguishable from zero, consistent with the parallel trends assumption. The post-treatment coefficients fluctuate around zero without a clear pattern of positive effects. The ATT is small and not statistically significant.

Figure 3: Effect of Court Expansion on Reversal Rates (Sun–Abraham, Yearly)



Notes: Event-study coefficients and 95% wild cluster bootstrap confidence intervals from the Sun and Abraham (2020) interaction-weighted estimator. Treatment is defined as adding ≥ 4 judges to the court. The gray shaded region highlights periods -1 (reference) and 0 (partially treated). The horizontal red dotted line shows the estimated ATT averaged over post-treatment periods 1 through 5. The sample is a court-year panel of 24 state appellate courts, 1970–2024.

I present the sub-sample event studies in Appendix 2. Figure A8 reports the event study for civil cases; Figure A9 for criminal cases. Quarterly event studies, which provide higher-frequency evidence on the timing of effects, are reported in Appendix 4 (Figure A12). These figures are even noisier than the results using a yearly granularity. I also estimate a continuous event study specification—interacting event-time indicators with the continuous capacity-change measure rather than using a binary treatment threshold—in Figure A11.

5.3 Sensitivity to Treatment Definition

A central challenge in this analysis is defining what constitutes a treatment. The binary event study requires a threshold: how many seats must be added for an expansion to count as treatment? Ideally, the choice would not matter and results would be robust to all reasonable specifications.

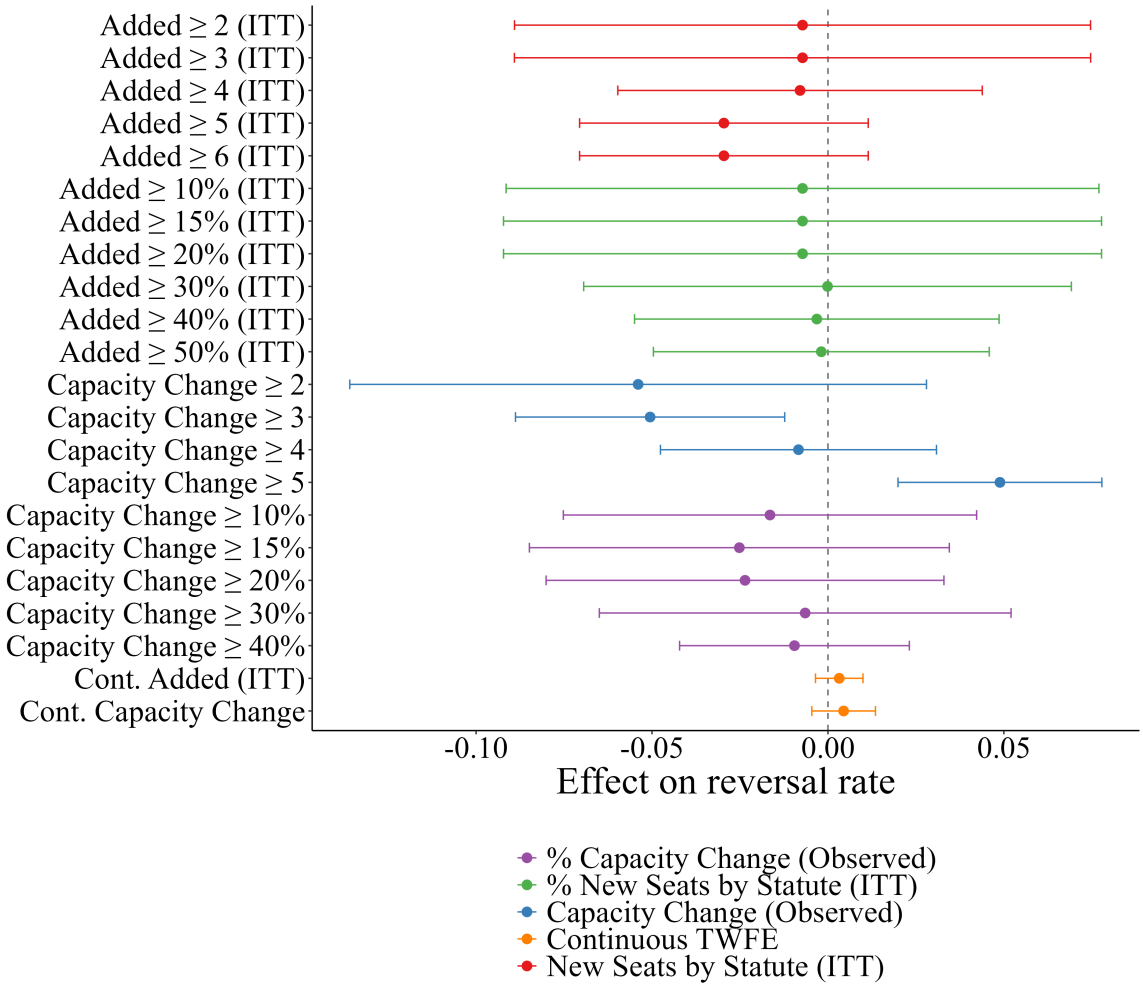
Figure 4 displays estimated treatment effects across 20 binary specifications and 2 continuous specifications. The binary specifications vary the threshold for treatment along two dimensions: the absolute number of seats added and the percentage increase in court size,

for both ITT and observed capacity measures. The continuous specifications (shown at the bottom) impose no threshold.

The results are strikingly sensitive to specification choice. Across the binary ITT specifications, point estimates range from -8.7 percentage points (seats added $\geq 40\%$ of initial court size) to $+0.6$ percentage points (seats added ≥ 5). The observed capacity measures are similarly unstable. The sign, magnitude, and statistical significance of the estimated effect all change with the threshold, and no single pattern emerges. The continuous specifications, which avoid the threshold problem entirely, yield small positive point estimates (0.41–0.50 pp) with confidence intervals that include zero.

This sensitivity is perhaps the most important result in the paper. It demonstrates that any claim of a positive (or negative) effect of court expansions on reversal rates is contingent on a specific, and somewhat arbitrary, treatment definition. Different thresholds identify different sets of treated courts—ranging from 1 court to 16 courts—with different expansion magnitudes, in different time periods, against different comparison groups. The fragility of the binary results is what motivates my choice of the continuous specification as the lead.

Figure 4: Sensitivity of Estimated Effect to Treatment Definition



Notes: Each point reports an estimated effect on the reversal rate. For binary definitions, the point is the Sun–Abraham ATT (excluding periods 0 and −1) with 95% confidence intervals from wild cluster bootstrap. For continuous specifications, the point is the TWFE coefficient per seat added/changed with cluster-robust 95% confidence intervals.

5.4 Effect on Caseload

If court expansions increase judicial capacity, we might expect them to affect not only the quality of decisions but also the quantity of published opinions. I estimate Equation 1 with the number of published opinions per court-year as the dependent variable. The ITT estimate is −16.8 opinions per seat added ($p = 0.17$), and the observed-capacity estimate is −2.6 ($p = 0.84$). Neither is statistically significant. I also present the Sun–Abraham event study for the number of published opinions per court-year in Figure A10. In that specification, the ATT is positive, although also small and insignificant. As with the primary outcome of interest, the results of court expansions on caseloads appear null. Interpretation

is challenging. The lack of a result here may be because there legitimately is no effect. Alternatively, opposing forces could result in the null: perhaps judges facing lower caseloads in one court substitute away from unpublished summary dispositions toward more carefully written published opinions while judges in another simply catch up on the backlog of cases. Unfortunately, the small number of treated courts makes it difficult to properly estimate any heterogeneous treatment specification.

5.5 Federal Circuit Courts

Although my focus is on state courts, I also examine federal circuit court expansions as a supplementary analysis. As with the state level, I do this analysis using both the "ITT" approach (i.e., I use only the number of authorized judgeships) as well as the observed number of active judges[15] The largest expansion event at the federal level was the Omnibus Judgeship Act of 1978, which added seats to nearly every circuit simultaneously. This near-universal simultaneity severely limits the usefulness of a difference-in-differences design: with almost no untreated circuits in 1978, identification relies on variation in the *number* of seats added across circuits. I omit results from an analysis which uses only the post-1978 period to estimate a modified version of equation 1 since there is almost no meaningful variation in court expansions during this period.

The continuous TWFE estimate for federal circuits is -0.28 percentage points per seat added (standard error = 0.24), small, negative, and statistically insignificant. This null result is consistent with the state-level findings but could also reflect the weaker identification at the federal level. I also estimate this FJC data, finding an effect of -0.006 percentage points per seat added (standard error = 0.005).[16] I interpret these small, insignificant effects as indicative of a lack of an actual workload shock in the federal system, not a proper test of the hypothesis that large expansions may affect appellant outcomes.

I do, however, provide suggestive evidence of one shock at the federal level that likely was large enough to measurably affect caseloads, albeit in only one circuit. In 1979, the Ninth Circuit became the first federal judicial circuit to establish a Bankruptcy Appellate Panel (BAP), authorized by the Bankruptcy Reform Act of 1978. The BAP heard bankruptcy appeals that would otherwise have been decided by the circuit's three-judge panels, effectively reducing the Ninth Circuit's appellate caseload burden. If caseload pressure affects decision quality, establishing the BAP should have freed Ninth Circuit judges to devote more attention to their remaining cases, potentially increasing the reversal rate for non-bankruptcy appeals.

I estimate a simple difference-in-differences comparing reversal rates in the Ninth Circuit against all other federal circuits before and after the BAP's establishment. Figure A14 plots yearly reversal rates for the Ninth Circuit against the average for all other federal circuits. Table A9 reports the regression results. The court-year-level specification yields a positive,

[15] At the federal level I can measure the number of active judges using CourtListener's data or data from the Federal Judicial Center. The differences in these datasets are small and results look similar with both databases so I report only the numbers from my analysis with CourtListener data.

[16] Because my focus is on state courts, I omit these results, though they can be made available upon request.

but small and insignificant coefficent, while the opinion-level specification yields a coefficient of 1.5 percentage points ($p = 0.04$), statistically significant at the 5 percent level. This is suggestive evidence consistent with the capacity hypothesis, though it should be interpreted cautiously given the small number of clusters.

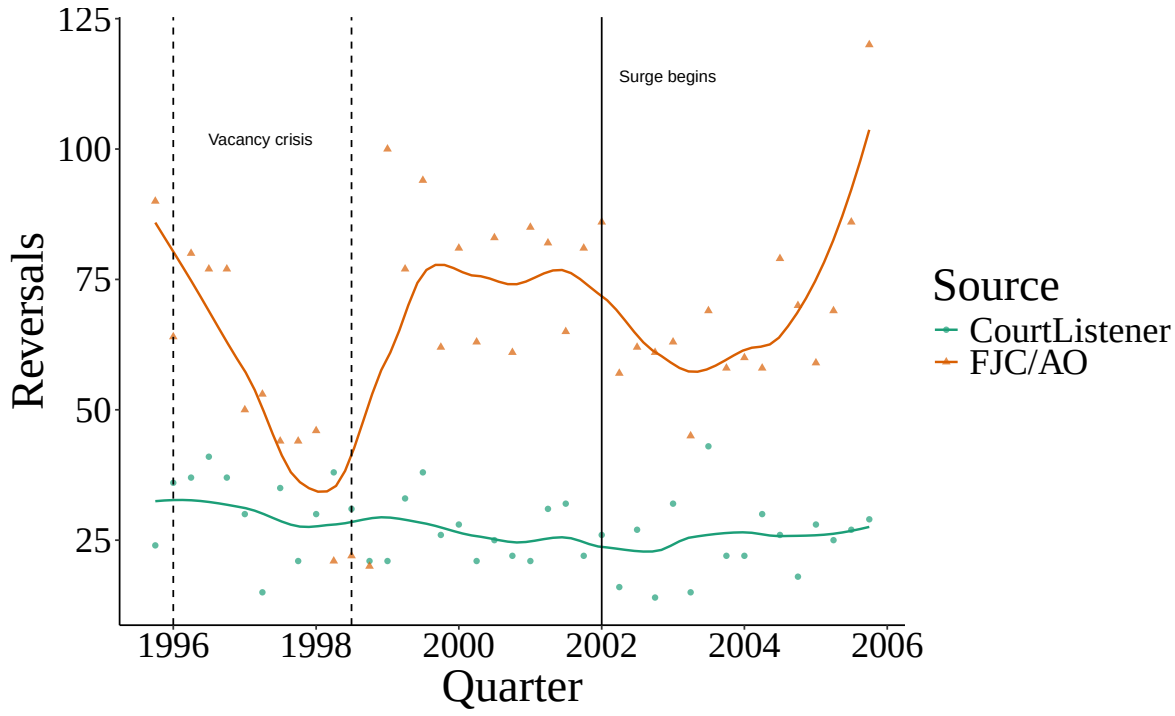
6 Discussion

The headline result of this paper is that I find no statistically significant effect of court expansions on reversal rates. This holds whether I use continuous or binary treatment measures, at the state or federal level, for all cases or for civil and criminal sub-samples. The question is whether this null reflects the absence of a true effect (potentially because the hypothesized behavioral biases may persist even if judges are given additional time), or simply the inability of the available data to detect an effect. While I am unable to rule out the possibility that judges simply do not respond to a reduced caseload, the following subsections present support for the claim that the available data is simply insufficient to answer this question. I conclude this section by asking what would constitute sufficient evidence.

6.1 Replication of Huang (2011)

Huang (2011) uses Administrative Office data provided by the Federal Judicial Center (FJC) covering all issued opinions (i.e., this data includes unpublished, or nonprecedential opinions), whereas CourtListener contains nearly all precedential opinions, but substantially fewer nonprecedential opinions. This means that even though absolute levels may differ, qualitative patterns should still be similar. Figure 5 plots the number of federal circuit court civil cases reversed on appeal between 1995 to 2005. The orange triangles use FJC data, closely mirroring the figure presented by Huang (2011). As documented, the number of reversals sharply declines leading up to the "vacancy crisis" in which "five out of thirteen judicial positions on the circuit had become vacant" leading the Second Circuit's chief judge to declare a "judicial emergency" before sharply rising as those vacancies were filled. The same pattern occurs a few years later when an influx of immigration cases overwhelmed the Second Circuit when "tens of thousands of appeals from the federal immigration agency" occupied Second Circuit judges. By contrast, the green dots, plotted with CourtListener civil appeals data remain flat during the entire period.

Figure 5: Replication Huang's (2011) Figure 5: Reversals In the Second Circuit



Notes: This figure attempts to replicate Figure 5 from Huang (2011) which presents the raw counts of reversals per quarter in the Second Circuit. The green dots use CourtListener civil appeals data. The orange triangles use data from the Administrative Office, made available by the Federal Judiciary Center. Both datasets are quarterly panels from 1995 Q4–2005 Q4. Solid vertical lines mark the BIA surge onset (2002 Q1); dashed vertical lines mark the Second Circuit vacancy crisis (1996–1998). LOESS smoothers are overlaid.

This exercise is informative about the limitations of the CourtListener database for studying the relationship between caseload pressure and appellate outcomes. Huang's original results, estimated on the FJC's comprehensive data, are clear. When I attempt the same analysis using CourtListener's published-opinion sample, the qualitative patterns are obscured by noise. If CourtListener data cannot cleanly reproduce a well-established finding in the federal context (where coverage is far better than at the state level) from a setting with a large, well-identified shock, it is unlikely to have the precision needed to detect the smaller and less sharply identified effects I seek in the state court expansion setting.

This is the strongest corroboration that data limitations—not the absence of an underlying effect—are the binding constraint on my analysis.

6.2 Data Limitations as the Binding Constraint

My estimates are imprecise. The 95 percent confidence interval from the lead specification (-0.31 to $+1.31$ pp per seat) comfortably includes effect sizes that would be economically

meaningful. I lack the statistical power to distinguish a small positive effect from zero. Second, the treatment sensitivity analysis (Figure 4) reveals that the null is not a stable finding across specifications—point estimates range from -9 to $+7$ percentage points depending on the treatment definition—but rather an average over a distribution of noisy, specification-dependent estimates. Third, the hypothesis itself remains theoretically well-motivated: there is no reason to expect that judges are immune to the resource constraints that affect decision quality in other institutional settings. The imprecision of my estimates can be traced to the following data limitations:

Publication bias in the CourtListener database. As documented in Table 2, the CourtListener database contains a highly variable mix of published and unpublished opinions across courts. Seven courts in the sample appear to have only published opinions in the database, likely because CourtListener has not yet ingested their unpublished dispositions. At the other extreme, Virginia—one of the four treated courts—has only 33 percent of its opinions tagged as published, meaning the majority of its database entries are non-precedential decisions. Published opinions disproportionately contain reversals, since courts are more likely to write and publish full opinions when they overturn a lower court. This cross-court variation in the published/unpublished composition means that the measured reversal rate reflects a fundamentally different sample of decisions in different courts. Court fixed effects absorb time-invariant differences in publication composition, but if publication practices change differentially over time—for instance, because CourtListener began ingesting unpublished opinions from certain courts at certain dates—the resulting compositional shifts could generate spurious trends in the dependent variable. This concern is particularly acute for Virginia, where the high share of unpublished opinions may reflect a recent bulk ingestion that coincides with the 2021 expansion. Future work should either restrict to published opinions with a consistently applied definition or obtain administrative disposition data that covers the universe of cases regardless of publication status.

Few treated units. Only four state appellate courts experienced expansions of four or more seats during the sample period. Two (Colorado and Michigan) were treated in 1974, leaving just two post-2000 expansions (Virginia 2021, Arizona 2022) with at most three to four years of post-treatment data. This small number of treated units limits both statistical power and the ability to learn about treatment effect heterogeneity. A dataset covering more states, or incorporating expansions of trial courts or specialized courts, would substantially increase the number of identifying events.

Noisy outcome classification. My text-based classifier disagrees with manual classification in roughly 14 percent of cases. This includes mixed dispositions (e.g., "affirmed in part, reversed in part") that cannot be cleanly assigned to a binary reversal/affirmance variable. Classical measurement error in the dependent variable does not bias OLS coefficients, but it does inflate standard errors—exactly the pattern I observe. Administrative data on case dispositions, such as those maintained by the National Center for State Courts, would provide more accurate outcome measures.

Imprecise judicial appointment dates. Nearly 70 percent of judges in the CourtListener database have appointment dates recorded only at the year level. This introduces substantial measurement error in the timing of capacity changes, which matters most for the

observed-capacity specifications and for sub-annual (quarterly, monthly) analyses. Court administrative records typically maintain precise appointment dates, and access to such records would sharpen both the treatment timing and the capacity measure.

Inability to isolate the caseload mechanism. As discussed in Section 4, the estimates in this paper capture the reduced-form effect of court expansions on reversal rates, combining the hypothesized caseload channel with any compositional effects of newly appointed judges. Isolating the caseload mechanism requires restricting to cases decided by incumbent judges—those who were on the court before the expansion. I am unable to perform this restriction because panel composition data are not systematically available in the CourtListener database, and extracting judge names from opinion text is not yet complete. This is not merely a precision issue; it is a gap in the identification strategy. Without it, even a precisely estimated positive effect could not be attributed to reduced caseload pressure rather than to ideological or experiential differences between new and incumbent judges. Administrative court records, which typically identify the judges on each panel, would close this gap.

Beyond these limitations, the text-based civil/criminal classifier leaves 15 percent of cases unclassified, and the sub-sample event studies reveal troubling pre-trend patterns that may reflect compositional shifts in the case mix around expansion events.

6.3 What Would Constitute Convincing Evidence?

A convincing first stage would show a measurable reduction in per-judge caseload following the expansion (something that is impossible to observe in the CourtListener database, which contains only published opinions, not the full set of cases assigned to each judge). Crucially, the analysis would also restrict to cases decided by incumbent judges to isolate the caseload channel from the composition channel, as discussed in Section 4.

My analysis falls short of this standard. I cannot observe the first stage, I do not restrict to incumbent judges' cases, the outcome variable is measured with error, and the small number of treated courts leaves the estimates too imprecise to distinguish economically meaningful effects from zero.

7 Conclusion

This paper provides the first systematic empirical test of whether expanding judicial capacity on appellate courts affects the rate at which appellants prevail. I construct a novel dataset combining over 700,000 published opinions from 24 state intermediate appellate courts with hand-compiled data on legislative court expansions, and I apply continuous two-way fixed effects and heterogeneity-robust event study estimators.

The results are inconclusive. Point estimates are generally small and positive, consistent with the hypothesis that increased capacity improves the quality of appellate review, but they are not statistically significant. This imprecision is attributable to fundamental limitations

of the data: publication bias in the CourtListener database, too few treated courts, a noisy text-based outcome classifier, and unreliable judicial appointment dates. Moreover, because I cannot identify which judges participated in each decision, the estimates are reduced-form—they conflate the hypothesized caseload mechanism with compositional effects of new appointments—and should not be interpreted as causal estimates of the caseload channel. The sensitivity of binary event study results to the treatment definition underscores the fragility of any single specification in this setting.

Because of these shortcomings, the hypothesis that judicial workload affects case quality remains untested. If confirmed, such a finding would have direct implications for court funding policy. The dataset and empirical framework I develop provide a foundation for future work. The most promising path forward involves obtaining administrative case disposition data from state court systems, which would simultaneously resolve the outcome classification problem and provide the panel composition data needed to isolate the caseload mechanism from compositional effects. Combined with a larger set of expansion events—potentially incorporating trial courts, specialized courts, or courts in additional states—such data would substantially increase the power to detect effects that the current analysis does not.

References

Administrative Office of the U.S. Courts. 2024. "Table B-5—U.S. Courts of Appeals Statistical Tables For The Federal Judiciary (December 31, 2024)." https://www.uscourts.gov/data-news/data-tables/2024/12/31/statistical-tables-federal-judiciary/b-5, December.

Arkhangelsky, Dmitry, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager. 2021. "Synthetic Difference-in-Differences." *American Economic Review* 111 (12): 4088–4118. 10.1257/aer.20190159.

Arnold, David, Will Dobbie, and Crystal S Yang. 2018. "Racial Bias in Bail Decisions." *The Quarterly Journal of Economics* 133 (4): 1885–1932. 10.1093/qje/qjy012.

Ash, Elliott, Daniel L Chen, and Suresh Naidu. 2025. "Ideas Have Consequences: The Impact of Law and Economics on American Justice." *The Quarterly Journal of Economics* qjaf042. 10.1093/qje/qjaf042.

Berdejó, 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. 10.1162/REST_a_00296.

Bhuller, Manudeep, and Henrik Sigstad. 2025. "Feedback and Learning: The Causal Effects of Reversals on Judicial Decision-Making." *The Review of Economic Studies* 92 (4): 2359–2397. 10.1093/restud/rdae073.

Black, Ryan C., Ryan J. Owens, and Patrick C. Wohlfarth. 2023. "The Effects of Lifetime Tenure and Aging in the United States Federal Judiciary." August. 10.2139/ssrn.4555766.

Bonica, Adam, Adam Chilton, Jacob Goldin, Kyle Rozema, and Maya Sen. 2019. "Legal Rasputins? Law Clerk Influence on Voting at the US Supreme Court." *The Journal of Law, Economics, and Organization* 35 (1): 1–36. 10.1093/jleo/ewy024.

Cai, Xiqian, Shuai Chen, Zhengquan Cheng, and Emily E. Nix. 2025. "Gender-Based Violence and Judge Responses." October. 10.3386/w34345.

Callaway, Brantly, Andrew Goodman-Bacon, and Pedro H. C. Sant'Anna. 2024. "Difference-in-Differences with a Continuous Treatment." Technical Reportw32117, National Bureau of Economic Research. 10.3386/w32117.

Cohen, Alma. 2025. "The Pervasive Influence of Political Composition on Circuit Court Decisions." *Journal of Legal Analysis* 17 (1): 14–41. 10.1093/jla/laaf004.

Danziger, Shai, Jonathan Levav, and Liora Avnaim-Pesso. 2011. "Extraneous Factors in Judicial Decisions." *Proceedings of the National Academy of Sciences* 108 (17): 6889–6892. 10.1073/pnas.1018033108.

Edwards, Barry C. 2019. "Why Appeals Courts Rarely Reverse Lower Courts: An Experimental Study to Explore Affirmation Bias."

Eisenberg, Theodore. 2004. "Appeal Rates and Outcomes in Tried and Nontried Cases: Further Exploration of Anti-Plaintiff Appellate Outcomes." *Journal of Empirical Legal Studies* 1 (3): 659–688. 10.1111/j.1740-1461.2004.00019.x.

Eren, Ozkan, and Naci Mocan. 2025. "Judge Gender Peer Effects in the Courthouse." *American Law and Economics Review* ahaf014. 10.1093/aler/ahaf014.

Harris, Allison P., and Maya Sen. 2019. "Bias and Judging." *Annual Review of Political Science* 22 (1): 241–259. 10.1146/annurev-polisci-051617-090650.

Holden, Richard, Michael Keane, and Matthew Lilley. 2021. "Peer Effects on the United States Supreme Court." *Quantitative Economics* 12 (3): 981–1019. 10.3982/QE1296.

Huang, Bert I. 2011. "Lightened Scrutiny." *Harvard Law Review* 124 (5): 1109–1152.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York, NY: Cambridge University Press.

MacKinnon, James G., Morten Ørregaard Nielsen, and Matthew D. Webb. 2023. "Fast and Reliable Jackknife and Bootstrap Methods for Cluster-robust Inference." *Journal of Applied Econometrics* 38 (5): 671–694. 10.1002/jae.2969.

Marvell, Thomas B. 1989. "State Appellate Court Responses to Caseload Growth." *Judicature* 72 (5): 282–291.

Moody, Carlisle E., and Thomas B. Marvell. 1987. "Appellate and Trial Court Caseload Growth: A Pooled Time-Series?Cross-Section Analysis." *Journal of Quantitative Criminology* 3 (2): 143–167.

Rachlinski, Jeffrey J., and Andrew J. Wistrich. 2017. "Judging the Judiciary by the Numbers: Empirical Research on Judges." October. 10.1146/annurev-lawsocsci-110615-085032.

Roodman, David, Morten Ørregaard Nielsen, James G. MacKinnon, and Matthew D. Webb. 2019. "Fast and Wild: Bootstrap Inference in Stata Using Boottest." *The Stata Journal* 19 (1): 4–60. 10.1177/1536867X19830877.

Solomon, Peter H. 2015. "Understanding Russia's Low Rate of Acquittal: Pretrial Screening and the Problem of Accusatorial Bias." *Review of Central and East European Law* 40 (1): 1–30. 10.1163/15730352-40012001.

Sun, Liyang, and Sarah Abraham. 2020. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." *Journal of Econometrics*.

Tversky, Amos, and Daniel Kahneman. 1974. "Judgment under Uncertainty: Heuristics and Biases." *Science* 185 (4157): 1124–1131. 10.1126/science.185.4157.1124.

Webb, Matthew D. 2023. "Reworking Wild Bootstrap-Based Inference for Clustered Errors." *Canadian Journal of Economics/Revue canadienne d'économique* 56 (3): 839–858. 10.1111/caje.12661.

Appendix

1 Additional Descriptive Statistics

Table A5: Court-Year Panel Summary Statistics

	Mean	SD	Median	Min	Max
Reversal Rate	0.28	0.10	0.29	0.05	0.72
Opinions Published	649.27	643.79	458.50	10	3831
Appointed Judges	13.50	11.24	10	3	54
Active Judges	8.86	6.02	9	0	27
Opinions/Judge			57.99	1.83	
Year	2000.53	14.50	2001	1971	2024
Number of Courts	24				
Observations	932				

Court-year panel. Pre-treatment observations only. Includes never-treated courts.

Table A6: Opinion-Level Descriptive Statistics

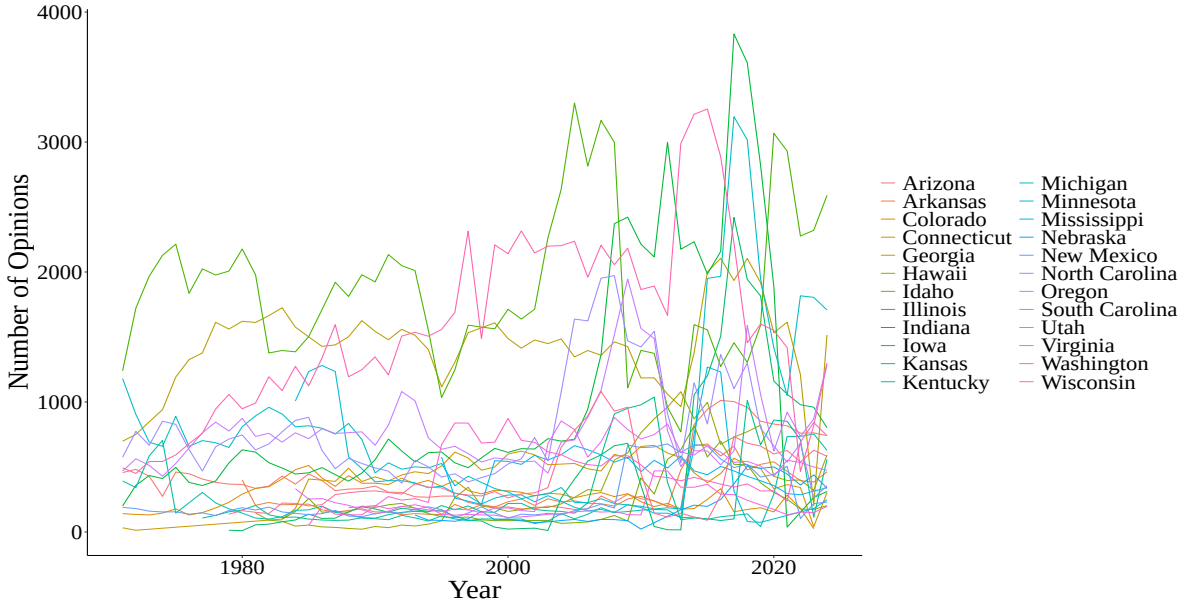
	Mean	SD	Median	Min	Max
Civil	0.30	0.46	0	0	1
Criminal	0.56	0.50	1	0	1
Not Yet Classified as Civil/Criminal	0.15	0.35	0	0	1
Appellant Won	0.27	0.44	0	0	1
Case Outcome Unclassified	0.06	0.24	0	0	1
Opinion Length (characters)	15967.72	16748.98	12034	16	1237837
Year Published	2002.72	14.99	2006	1971	2024
Number of Courts	24				
Number of Opinions	713236				

Full sample of published opinions from 24 state intermediate appellate courts available via CourtListener. Case type and outcome are classified using regular expressions applied to opinion text.

Table A7: States with Intermediate Appellate Courts Excluded from Sample

State	Reason for Exclusion
Alabama	Separate civil and criminal appellate courts
Alaska	Very few published opinions per year
California	Six geographically independent appellate districts
Florida	Five geographically independent district courts of appeal
Louisiana	Only state to use a civil law system
Maryland	Intermediate court originally limited to criminal jurisdiction
Massachusetts	Multiple courts with intermediate appellate jurisdiction
Missouri	Three geographic divisions with separate judge pools
Nevada	Deflective model: supreme court selectively assigns cases to court of appeals
New Jersey	Court structure not yet fully investigated
New York	Four geographically independent appellate divisions
North Dakota	Very few published opinions per year
Ohio	Court structure not yet fully investigated
Oklahoma	Separate civil and criminal appellate courts
Pennsylvania	Two statewide intermediate appellate courts with overlapping jurisdiction
Tennessee	Separate civil and criminal appellate courts
Texas	Fourteen geographically independent courts of appeals
West Virginia	Intermediate appellate court created in 2022; insufficient time series

Figure A6: Number of Opinions Published per Court-Year



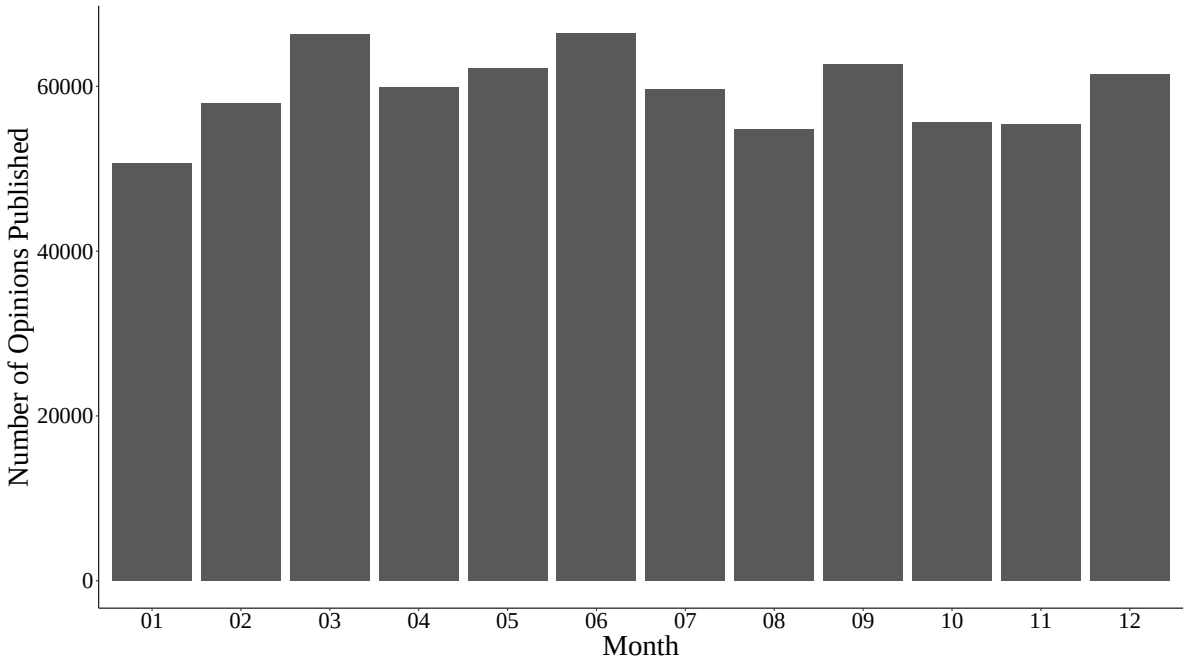
Notes: This figure shows the number of full-text published opinions available in CourtListener for each court-year in the sample.

Table A8: Balance Table: Treated vs. Never-Treated Courts

	Treated		Never-Treated		
	Mean	SD	Mean	SD	Norm. Diff.
Reversal Rate	0.27	0.14	0.28	0.06	-0.10
Opinions Published	556.92	382.47	605.44	491.20	-0.11
Appointed Judges	12.75	6.70	12.83	10.83	-0.01
Active Judges	12.50	7.42	8.76	5.18	0.58
Opinions/Judge	43.26	23.51	68.58	46.06	-0.69
Year	1995.75	27.43	2001.22	3.83	-0.28
Number of courts	4		20		
Court-year observations	12		920		

Each court contributes one observation (its pre-treatment mean). Excludes not-yet-treated courts from the control group. The normalized difference is $(\bar{X}_T - \bar{X}_C)/\sqrt{(S_T^2 + S_C^2)/2}$; values below 0.25 in absolute value suggest adequate balance, following Imbens and Rubin (2015).

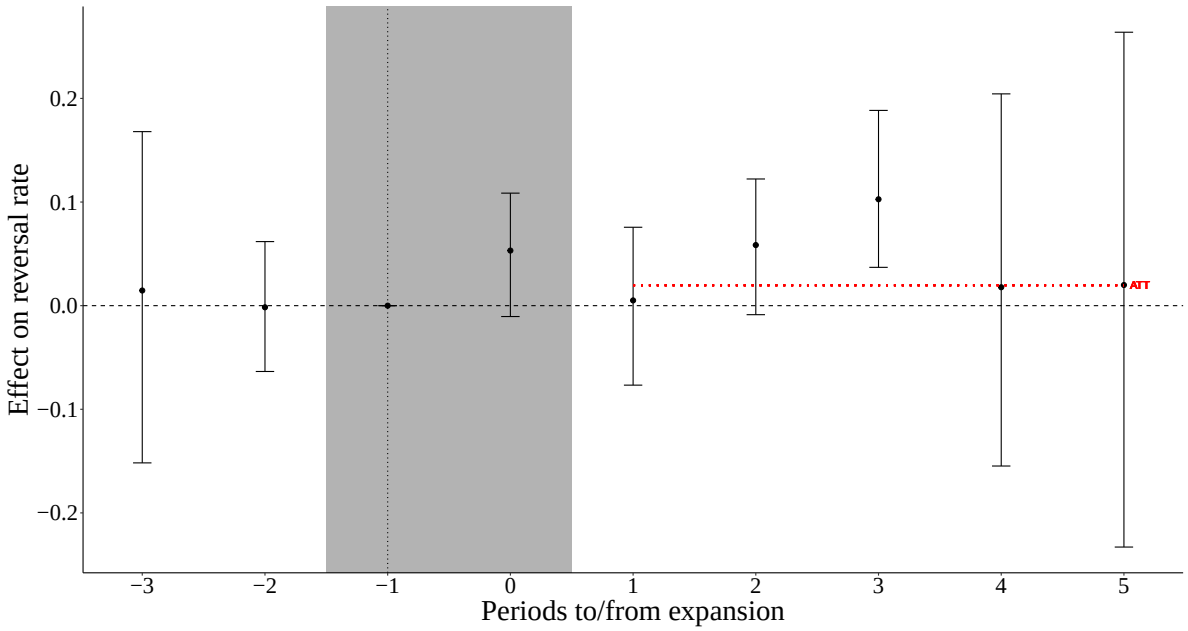
Figure A7: Seasonality in Opinion Publication



Notes: Total number of opinions issued in each calendar month across all courts and years in the sample. Courts appear to publish fewer opinions in summer months (July–August), consistent with reduced judicial activity during summer recesses.

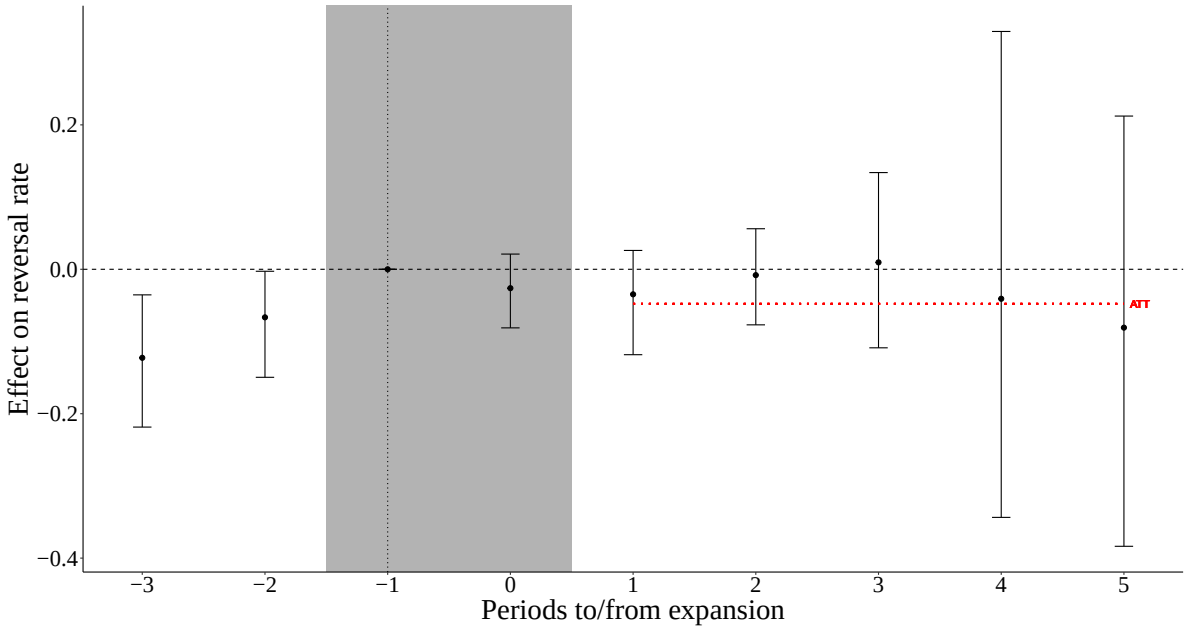
2 Sub-sample Event Studies

Figure A8: Event Study: Civil Cases (Sun–Abraham, Yearly)



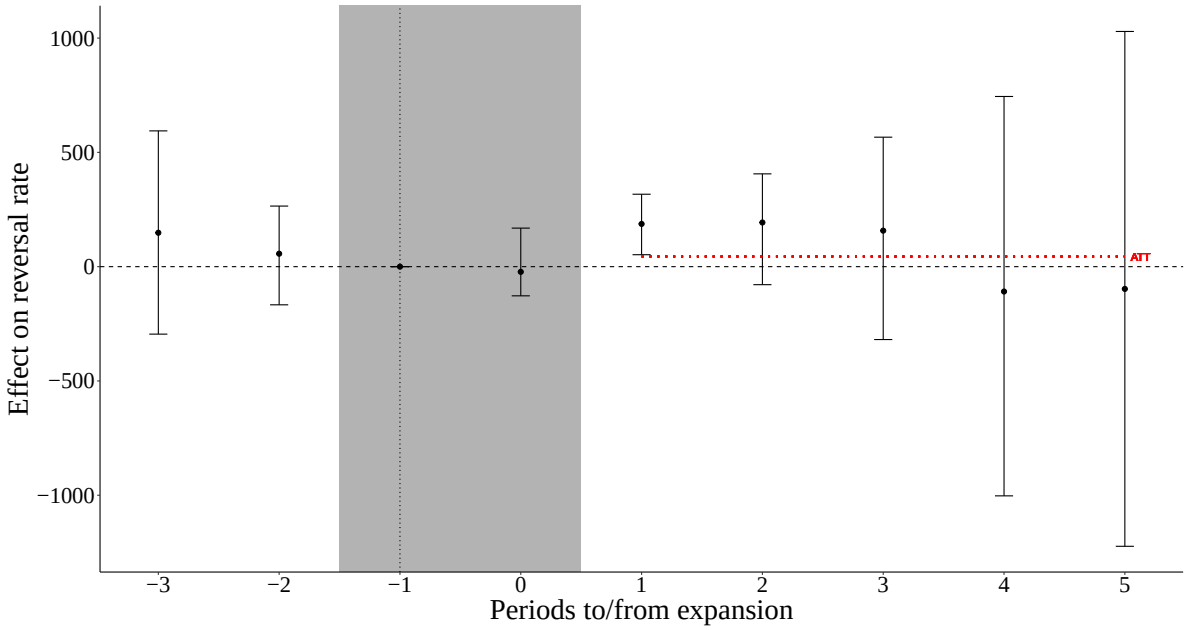
Notes: Same specification as Figure 3, restricted to opinions classified as civil cases.

Figure A9: Event Study: Criminal Cases (Sun–Abraham, Yearly)



Notes: Same specification as Figure 3, restricted to opinions classified as criminal cases.

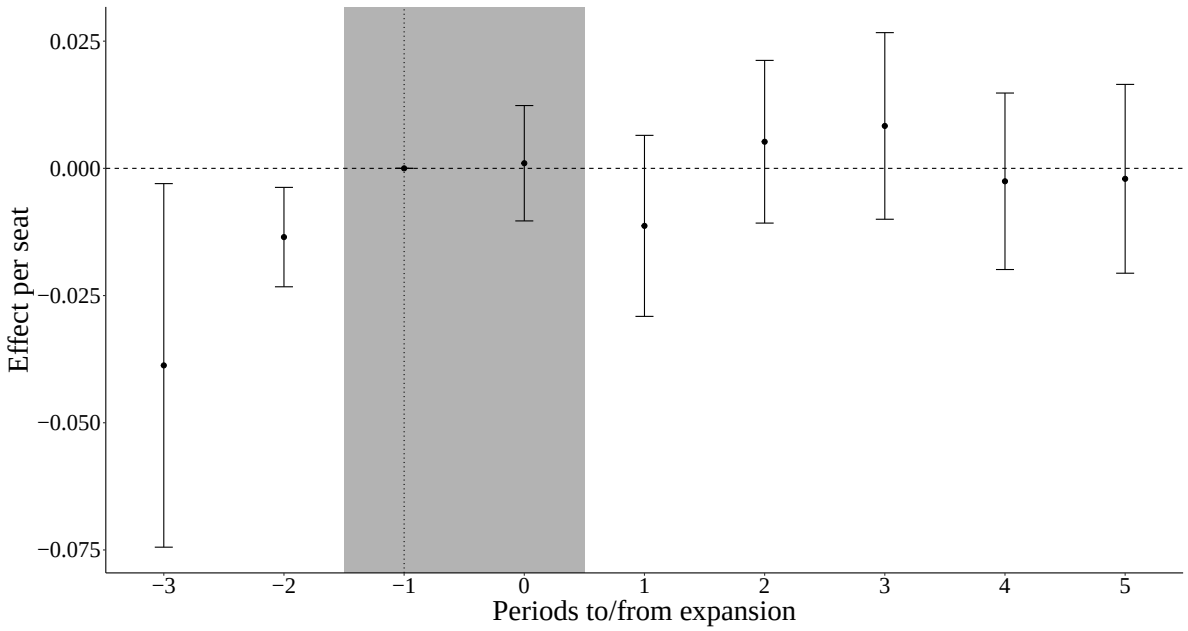
Figure A10: Event Study: Cases Published per Court-Year (Sun–Abraham)



Notes: Event-study coefficients estimating the effect of court expansion ($\geq$ 4 seats added) on the number of published opinions per court-year, using the Sun–Abraham estimator with wild cluster bootstrap confidence intervals.

3 Continuous Event Study

Figure A11: Continuous Event Study: Effect per Seat Added on Win Rate



Notes: Coefficients from an event-study regression interacting event-time indicators with the continuous capacity-change measure (observed Δ judges), with court and year fixed effects. Reference period is $t = -1$. Collinearity drops several event-time interactions because the capacity change is zero in most court-years.

4 Quarterly Event Studies

Figure A12 presents the Sun–Abraham event study at the quarterly frequency, using a 12-quarter pre-treatment and 12-quarter post-treatment window. The higher-frequency specification provides more granular evidence on the timing of effects but introduces additional noise from quarter-to-quarter variation in case composition.

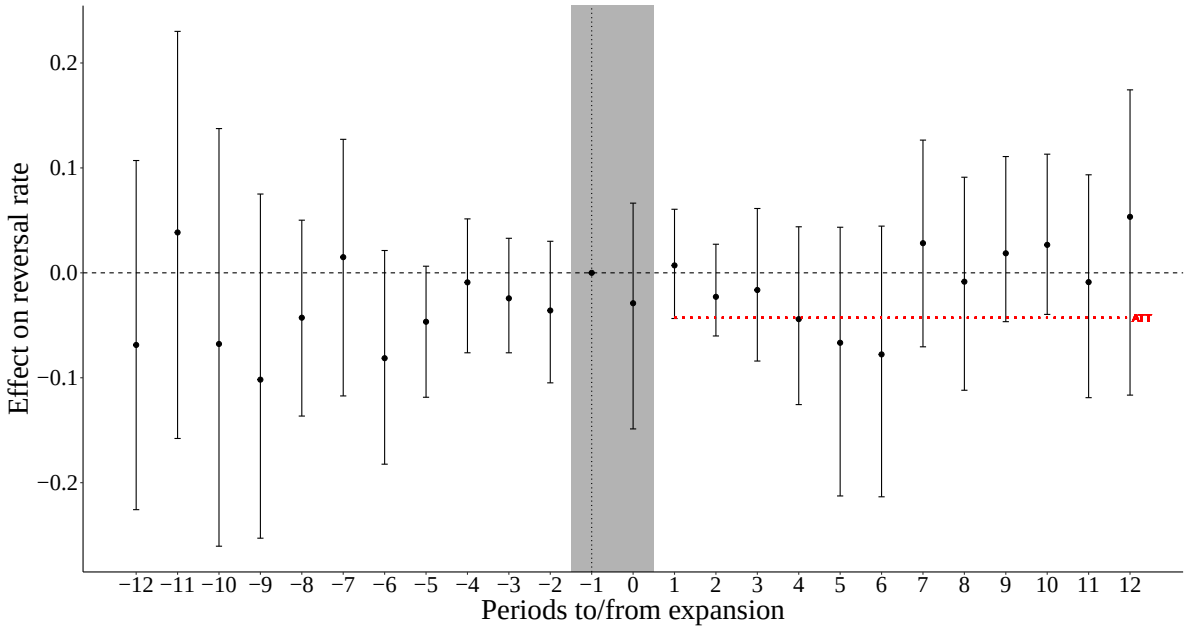


Figure A12: Event Study: All Cases, Quarterly (Sun–Abraham)

Notes: Quarterly analog of Figure 3. Event-study coefficients and 95% wild cluster bootstrap confidence intervals from the Sun–Abraham estimator. Treatment is defined as adding ≥ 4 judges. Reference period is $q = -1$. The quarterly specification uses a 12-quarter pre-treatment and 12-quarter post-treatment window.

5 Synthetic Control

Figure A13 presents pooled event-study estimates from the stacked synthetic control design described in Section 4.

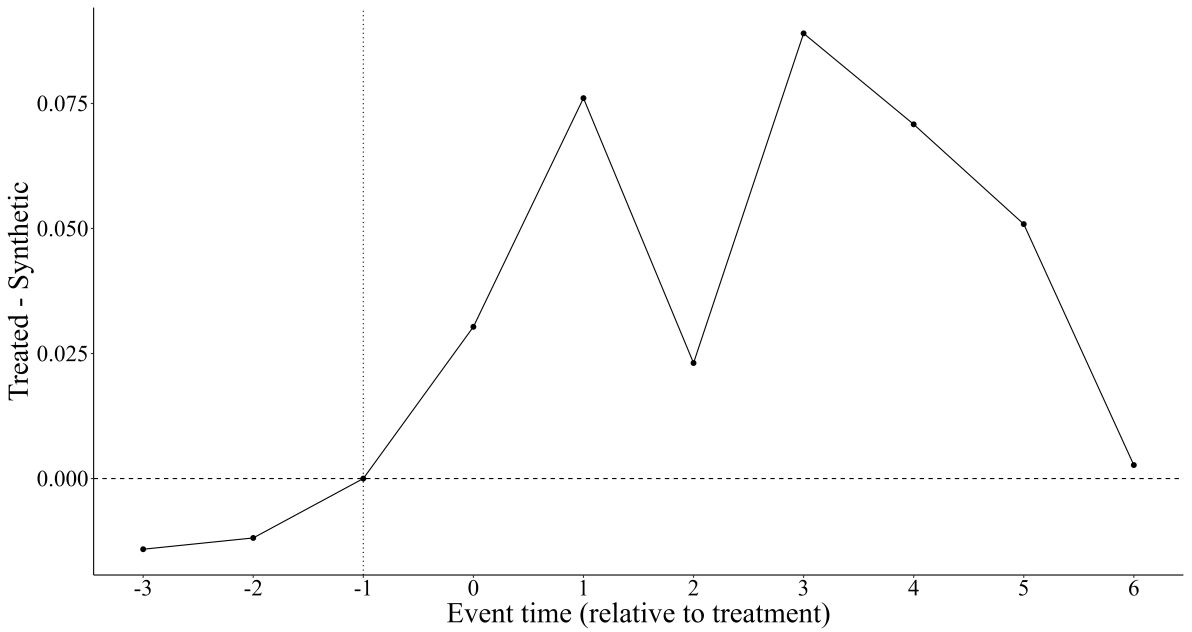


Figure A13: Pooled Synthetic Control Event Study: All Cases

Notes: Pooled event-study estimates from a stacked synthetic control design. For each treated court, a synthetic control is constructed from never-treated and not-yet-treated courts to match the pre-treatment reversal rate trajectory. The pooled estimate averages across treated courts.

6 Ninth Circuit Bankruptcy Appellate Panel

Figure A14: Reversal Rates: Ninth Circuit vs. Other Federal Circuits

Notes: Yearly reversal rates for the Ninth Circuit and the average across all other federal circuits. The dashed vertical line marks 1979, the year the BAP was established. Sample: 1970–1988.

Table A9: Diff-in-Diff: Ninth Circuit Bankruptcy Appellate Panel (1979)

	Court-Year Panel	Opinion-Level
Ninth × Post-1979	0.011	0.015**
	(0.007)	(0.006)
Observations	217	219 091
R^2	0.691	0.010
Adj. R^2	0.641	0.010
Court FE	Yes	Yes
Year FE	Yes	Yes
Cluster	Circuit	Circuit

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Sample: all federal circuit court opinions, 1970–1988. The Eleventh Circuit (created 1981) enters the sample in 1981. The dependent variable is a reversal indicator (1 = reversed/remanded/vacated, 0 = affirmed/dismissed). Standard errors clustered by circuit.