NBER WORKING PAPER SERIES

REDUCING PARTISANSHIP IN JUDICIAL ELECTIONS CAN IMPROVE JUDGE QUALITY: EVIDENCE FROM U.S. STATE SUPREME COURTS

Elliott Ash
W. Bentley MacLeod

Working Paper 22071
http://www.nber.org/papers/w22071

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2016, Revised June 2021

We thank Yisehak Abraham, Ankeet Ball, Josh Brown, Josh Burton, Matthew Buck, David Cai, Eammonn Campbell, Zoey Chopra, Daniel Deibler, Seth Fromer, Gohar Harutyunyan, Archan Hazra, Montague Hung, Dong Hyeun, Mithun Kamath, James Kim, Michael Kurish, Jennifer Kutsunai, Steven Lau, Sharon Liao, Sarah MacDougall, Sam Meshoyrer, Justin McNamee, Sourabh Mishra, Brendan Moore, Arielle Napoli, Karen Orchansky, Bryn Paslawski, Olga Peshko, Quinton Robbins, Ricardo Rogriguez, Jerry Shi, Shawn Shi, Carol Shou, Alex Swift, Holly Toczko, Tom Verderame, Sam Waters, Sophie Wilkowske, John Yang, Geoffrey Zee, Fred Zhu, and Jon Zytnick for their meticulous help in assembling data and other research assistance. We thank Daniel Chen, Tom Clark, John Ferejohn, Sanford Gordon, Chris, Hanretty, Jon Kastellec, Lewis Kornhauser, Eric Posner and the participants at Princeton University Conference on Bureaucrats, SIOE meetings at Harvard Law School, NYU Law and Economics Workshop, and Conference on Empirical Legal Studies in Europe for helpful comments. Columbia University's Program for Economic Research, Columbia Law School, and the National Science Foundation Grant SES-1260875 provided financial support for this research. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Reducing Partisanship in Judicial Elections Can Improve Judge Quality: Evidence from U.S. State Supreme Courts

Elliott Ash and W. Bentley MacLeod

NBER Working Paper No. 22071
March 2016, Revised June 2021
JEL No. J24,K4

ABSTRACT

Should technocratic public officials be selected through politics or by merit? This paper explores how selection procedures influence the quality of selected officials in the context of U.S. state supreme courts for the years 1947-1994. In a unique set of natural experiments, state governments enacted a variety of reforms making judicial elections less partisan and establishing merit-based procedures that delegate selection to experts. We compare post-reform judges to pre-reform judges in their work quality, measured by forward citations to their opinions. In this setting we can hold constant contemporaneous incentives and the portfolio of cases, allowing us to produce causal estimates under an identification assumption of parallel trends in quality by judge starting year. We find that judges selected by nonpartisan processes (nonpartisan elections or technocratic merit commissions) produce higher-quality work than judges selected by partisan elections. These results are consistent with a representative voter model in which better technocrats are selected when the process has less partisan bias or better information regarding candidate ability.

Elliott Ash
ETH Zurich
IFW E47.1
Zurich 8044
Switzerland
ashe@ethz.ch

W. Bentley MacLeod
Department of Economics
Columbia University
420 West 118th Street, MC 3308
New York, NY 10027
and NBER
wbmacleod@wbmacleod.net

> The aim of every political constitution is, or ought to be, first to obtain for rulers men who possess most wisdom to discern, and most virtue to pursue, the common good of the society...
>
> – *Federalist No. 57*

1 Introduction

Our daily lives are shaped by the decisions of public officials, in particular those who make decisions in our courts (Djankov et al., 2003). In the case of common-law appellate cases, such as those handled by U.S. state supreme courts, the decisions have the power of law (Landes and Posner, 1980; Gennaioli and Shleifer, 2007; Baker and Mezzetti, 2012). Judges, like central bankers, are skilled technocrats tasked with making socially impactful decisions. Therefore, how they are selected is a crucial decision for constitutional designers (Maskin and Tirole, 2004; Alesina and Tabellini, 2007).

U.S. state supreme courts serve as the state judiciary's analogue to the U.S. Supreme Court.[1] State supreme court judges are some of the most powerful officials in state government, with the authority to review not just the decisions of lower courts but also the laws produced by state legislatures. These judges are the last appeal on important features of common law, including most rules on contract, property, employment, tort, and crimes. Their decisions serve as binding precedent for all courts in the state.

Given the stakes, it is interesting that U.S. state governments have experimented extensively with different procedures for selecting judges. Unlike the federal courts and the courts in most other countries, state courts have historically appointed appellate judges through elections. Originally these elections were essentially identical to partisan elections for any political office, including contested primaries. Many states continue to select judges through elections, with the stakes of the races reflected in large and growing campaign investments by business lobbying groups.[2] Over the last century, many states have tried to reduce the role of party politics by replacing partisan elections with nonpartisan elections, where judges do not have a stated party affiliation on the ballot. Many states have dropped elections altogether, replacing them with a merit-based process where judges are selected by experts (mostly senior judges). This paper analyzes what effect these two depoliticizing reforms have had on the relative quality of newly appointed judges.

Besides these natural-experiment reforms, the other attractive feature of this setting (from an empirical perspective) is that we can measure work quality in a high-skill environment. The job of a state supreme court judge is to review lower-court decisions and write opinions explaining

[1] The state court websites provide up-to-date information on how appellate courts work in each state. See `https://www.ohiobar.org/public-resources/commonly-asked-law-questions-results/appellate-judges-review-trial-court-decisions/` for a concise explanation. See also Hanssen (e.g. 1999); Helland and Tabarrok (e.g. 2002); Hanssen (e.g. 2004); Hall and Bonneau (e.g. 2006); Hall (e.g. 2007); Kritzer (e.g. 2011, 2015).

[2] See Bannon et al. (2013) for details and statistics about the increased campaign spending in state supreme court races. See Appendix Table 1 for a recent history of these procedures.

whether they should be upheld or reversed. The job of an appellate judge is very different from a trial court judge. They do not review evidence, nor see witnesses. Their job is to ensure that there were no legal errors. Hence, trial court decisions are overturned only in the event of a serious legal error. The judges work with a team of clerks to research the relevant case law, reason through the implications for present and future litigants, and articulating a precedent that future judges in the state must follow. Because these opinions represent essentially all of a judge's significant work product, they can be used to construct measures of judge performance.

We measure work quality by the frequency with which a judge's decisions are cited positively in future cases. These citations measure the usefulness of an opinion and its influence on the evolution of common law.[3] Importantly, citations in the law are not like citations in academia, where they are recognized as an imperfect measure of academic performance (see Ellison, 2013). As already mentioned, we can distinguish positive from negative judicial citations. Moreover, judges do not choose what cases they work on (unlike academics, who choose the papers they work on). While citations are an indirect measure of quality, at the moment there is no accepted way to directly evaluate the quality of a legal rule, nor how it affects social outcomes.[4] Hence, for the moment citations are the best available measure of judicial performance for appellate courts.[5]

In our data, we document that state supreme court judges systematically vary in their work quality even within the same court. We can capture persistent differences across judges using the percentile rank in positive citations per opinion. We also measure a number of other judicial behaviors, including the total output of text written and citations from out-of-state judges.

State supreme courts have a number of features that make them a nice natural laboratory for analyzing workplace productivity. First, judicial compensation is fixed and does not depend on the judges' work choices. Second, the job description for state judges hasn't changed for decades. Especially relative to other technical professions, such as medicine or management (Choudhry et al., 2005; Bloom and Reenen, 2007), the knowledge and skills relevant to good judging barely change over time. Due to norms of equal work across judges (and, most of the time, random or rotating assignment of judges to cases), more-experienced judges face the same workload as less-experienced judges.

The main empirical question is whether the less politicized, more technocratic, appointment systems select judges with higher or lower ability, as measured by forward citations. We approach

[3]See Posner (2008) for a seminal discussion of the work of judges and what we mean by quality. Choi et al. (2010) discuss in detail why citations are used to measure judicial performance. See Stephenson (2009) for a discussion of how common law citations work.

[4]See Niblett et al. (2010) for evidence on how legal rules do not necessarily evolve efficiently.

[5]It worth emphasizing that the work of an appellate court is very different from a trial court. At a trial, the outcome is a decision that determines an outcome for the defendant. There is a growing literature that studies how characteristics of the defendant, such as race, can affect judicial decision making, and the outcome for the defendant, both before and after trial (see Aizer and Doyle (2015), Arnold et al. (2018); Ash et al. (2021) , Dobbie et al. (2018) and Ash et al. (2021)). Appellate courts do not directly observe defendants, nor do they decide the outcome of a trial. It is an open question whether or not race can influence appellate court decisions, and or how one would measure such effects.

this question motivated by a simple information-theoretic framework, where a representative voter makes a choice based on a signal of the candidate judge's ability and knowledge of the judge's political affiliation. If the merit system observes a more precise ability signal, that should improve judge quality relative to elections. Because nonpartisan elections reduce political bias, that should improve quality relative to partisan elections.

The empirical approach is differences-in-differences, where we compare judges working at the same time in the same court, but who were selected under different procedures. Formally, the regressions include court-year fixed effects, which hold constant all current-year factors that affect all judges of a court, including retention-related pressures. Unlike a standard differences-in-differences regression, the treatment variation comes historically from judge starting year relative to the reform year, rather than from immediate impacts in the years before and after the reform. Therefore our parallel-trends assumption is in judge citation rates by starting year, rather than by the year that cases are written.

The baseline selection system, partisan elections, is the most highly politicized system in that each judge represents a political party that is clearly identified on the ballot. We analyze three types of reforms. The first type of reform replaces partisan elections with nonpartisan elections, which are competitive but they do not have primaries and party affiliations are not on the ballot. The second type of reform replaces partisan elections with merit selection, where judges are nominated by a commission of experts – senior attorneys and retired judges – and confirmed by the governor.[6] The third reform substitutes nonpartisan elections with merit selection. Table 1 lists how these reforms apply to the different states in a given year. Appendix C includes more detailed information on the systems and reforms.

The main empirical finding is that moving from partisan elections to merit selection increases the forward citation rate for the post-reform selected judges. We subject this finding to a number of robustness checks for specification and identification. We show that the effect is not driven by confounding trends in citations, by judge experience, or by selection of judges into different types of cases. This result is not sensitive to how work quality is specified and holds for a variety of alternative measures. Overall, this supports the view that moving from a partisan system to merit selection improves performance – consistent with better signals on candidate quality and perhaps a reduction in political bias.

Second, we find that moving from partisan to nonpartisan elections also increases quality as measured by citations. The estimates for this effect are driven partly by a differential caseload effect, where the nonpartisan-selected judges rule on more important cases than partisan-selected judges. While the estimates must be interpreted more carefully, they are consistent with improved judge quality via the reduction of political bias in the selection process.

[6] In the the merit system, also known as the "Missouri Plan", the retention process is also reformed. Instead of contested elections, judges stand for reelection in uncontested retention elections. The change to the retention process does not play a role in our empirical analysis because our research design compares judges sitting on the same court at the same time, thus holding constant retention-related factors.

Table 1: Judicial Selection Rules

State (Years)	Selection	State (Years)	Selection
Alaska	Merit	New Hampshire	Governor
Alabama	Partisan	North Carolina	Partisan
Arkansas	Partisan	North Dakota	Nonpartisan
Arizona (-1974)	**Nonpartisan**	**Nebraska (-1962)**	**Partisan**
Arizona (1975-)	**Merit**	**Nebraska (1963-)**	**Merit**
California	Governor	New Jersey	Governor
Colorado (-1966)	**Partisan**	New Mexico	Partisan
Colorado (1967-)	**Merit**	Nevada	Nonpartisan
Connecticut	Governor	New York (-1976)	Partisan
Delaware	Governor	New York (1977-)	Governor
Florida (-1971)	Partisan	Ohio	Partisan
Florida (1972-1976)	Nonpartisan	**Oklahoma (-1967)**	**Partisan**
Florida (1977-)	Merit	**Oklahoma (1968-)**	**Merit**
Georgia (-1984)	**Partisan**	Oregon	Nonpartisan
Georgia (1985-)	**Nonpartisan**	Pennsylvania	Partisan
Iowa (-1962)	**Partisan**	Rhode Island	Governor
Iowa (1963-)	**Merit**	South Carolina	Legislature
Idaho	Nonpartisan	**South Dakota (-1980)**	**Nonpartisan**
Illinois	Partisan	**South Dakota (1981-)**	**Merit**
Indiana (-1970)	**Partisan**	**Tennessee (-1971)**	**Partisan**
Indiana (1971-)	**Merit**	**Tennessee (1972-1977)**	**Merit**
Kansas (-1958)	**Partisan**	**Tennessee (1978-)**	**Partisan**
Kansas (1959-)	**Merit**	Texas	Partisan
Kentucky (-1975)	**Partisan**	**Utah (-1951)**	**Partisan**
Kentucky (1976-)	**Nonpartisan**	**Utah (1952-1985)**	**Nonpartisan**
Louisiana	Partisan	**Utah (1986-)**	**Merit**
Maine	Governor	Vermont (-1971)	Legislature
Maryland (-1976)	**Nonpartisan**	Vermont (1972-)	Governor
Maryland (1977-)	**Merit**	Virginia	Legislature
Massachusetts	Governor	Washington	Nonpartisan
Michigan	Partisan	Wisconsin	Nonpartisan
Minnesota	Nonpartisan	West Virginia	Partisan
Missouri	Merit	**Wyoming (-1972)**	**Nonpartisan**
Mississippi	Partisan	**Wyoming (1973-)**	**Merit**
Montana	Nonpartisan		

Notes. This table lists the elections systems for state supreme court judges observed in our data. Election-system reforms indicated by cell borders. Items in bold indicate reforms used in the analysis.

Third, we find that replacing nonpartisan elections with merit selection does not affect the measured citation rate of new judges (although the estimate is imprecise). This finding suggests that merit procedures by themselves are not sufficient for increasing citations. One interpretation is that there is a countervailing effect of merit reforms, relative to nonpartisan elections, due to increased bias. After all, the governor's pivotal role in the merit system invites a political dimension that could diminish any informational advantage.

These results provide empirical evidence showing that the process for selecting technocratic officials affects their performance. Should the selection procedures for judges and other professional offices be political or bureaucratic? Our evidence is consistent with the hypothesis that more bureaucratic, less politicized, institutions result in better work quality.[7] This evidence is supported by an unusual natural experiment where U.S. states tried out different systems, a rich longitudinal dataset on measurable work quality, and a unique setting where we can hold constant the workload between technocrats selected under different systems.

This research adds to the political economy literature on technocratic public officials, and in particular on judges. The closest paper is Choi et al. (2010), who look at a cross-section of state supreme court opinions. Using data from 1998-2000, they find that states with appointed judges also tend to have more citations per opinion.[8] This results still leaves open the question of whether there is a causal link between changes in a state's appointment system and the performance of the individuals selected, as measured by citations.

Lim and Snyder (2015) provide additional cross-sectional evidence relating the appointment system to judge quality, but they measure performance using peer evaluations of judges by bar associations (rather than citations). Lim and Snyder show that bar evaluations of judge quality affect voting in nonpartisan elections. In partisan elections, voters do not respond to bar evaluations because the party labels become the most important factor.[9] As a consequence, lower-quality judges are selected in the partisan system as measured by the bar evaluations.

We show that the citation-based measures of performance are correlated with the quality mea-

[7]The literature on bureaucrats and politicians is closely related to our institutional context and reforms. Maskin and Tirole (2004) consider the optimal choice of institution (removable "politician" or unaccountable "judge"). In that model, election pressure can cause officials to modify their decisions to reflect the interests of the electorate. In Alesina and Tabellini (2007, 2008), tenured "bureaucrats" are preferred for technical tasks, where organizations can evolve a mission, and for protecting the rights of minorities. In contrast, elected "politicians" are more sensitive to outcomes, and to the preferences of the median voter.

[8]Many more papers look at other aspects of how electoral institutions influence judges. In terms of the quality of decision-making, a noteworthy paper by Iaryczower et al. (2013) uses a structural model to interpret the harsher decision-making of elected judges as evidence of higher bias and error rates. In a different context (Canadian asylum courts), Norris (2019) finds that judges selected after a single partisanship-reducing reform make fewer mistakes (based on appellate reversals) than judges selected before the reform.

[9]Supporting evidence in this vein includes Hall and Bonneau (2006), who demonstrate that voters respond to challenger quality in judicial elections (see also Bonneau and Cann, 2015). In non-judicial elections, early work by Rahn (1993) shows that voters use partisan identifiers as a shortcut for decision-making. Recently, A. Kirkland and Coppock (2017) provide experimental evidence that if voters cannot observe partisan labels, then they put more weight upon observable judge characteristics that reflect quality. Outside of the judiciary, work on voter responsiveness includes Ferraz and Finan (2008), who find that voters in Brazil punish incumbents after observing information on corruption.

sures used in Lim and Snyder (2015). In addition, we exploit changes in the way judges are selected to directly measure the causal effect of the state appointment system upon performance. Hence, our results extend this literature by directly measuring the counterfactual question of what would happen to court quality (as measured by citations) if the appointment system changes from a partisan system to either a nonpartisan election system or a merit-based system.

Besides these papers that look at the quality of judges, a larger literature explores how electoral institutions shape the policy direction or partisan bias in judicial decisions. For example, a broad finding is that stronger electoral pressures induce harsher treatment of criminals (Huber and Gordon, 2004; Gordon and Huber, 2007; Lim, 2013; Berdejo and Yuchtman, 2013; Lim et al., 2015).[10] Significant policy and economic impacts of variation in such institutional rules have been demonstrated empirically in U.S. states (Besley and Case, 1995, 2003; Besley and Coate, 2003; Besley et al., 2010). In a broader global context, a rich literature reviewed in Dal Bó and Finan (2018) has explored the institutional and other factors contributing to the types of public officials selected for office. Pande (2011), for example, focuses on how the quality of information regarding candidates plays a central role; other work points to competition and/or wages (Dal Bo et al., 2013; Dal Bó et al., 2017).

The rest of the paper is organized as follows. Section 2 outlines a theoretical framework for guiding and interpreting the analysis. Sections 3, 4, and 5 describe the empirics – data, methods, and results, respectively. Section 6 discusses how the evidence relates to the model, while Section 7 concludes.

2 Conceptual Framework

The existing theory literature on the behavior of public officials has generated a rich set of predictions from a range of institutional, economic, and behavioral influences.[11] A challenge for empirical work in political economy is that the data is rarely sufficient to explore the finely detailed predictions that have been developed in the literature. Hence, we focus on a simple reduced-form model to organize our results. The formal details are presented in Appendix A.

[10] This tough-on-crime effect is also seen in state appellate courts (Iaryczower et al., 2013; Canes-Wrone et al., 2014). Meanwhile, Shepherd (2009a,b) uses a cross-sectional state courts dataset to show that both elected judges and appointed judges tend to vote in the direction favored by the entity tasked with re-appointing them (voters or the governor). Besley and Payne (2013) find that elected judges are more likely to support anti-discrimination law. In the case of federal judges, Epstein et al. (2013) review evidence that decisions tend to reflect the ideological leanings of the president who appointed them. For governor responses to voter preferences, see for example List and Sturm (2006).

[11] See Ashworth (2012), Ashworth et al. (2017), and Dal Bó and Finan (2018) for overviews of this literature. Building on the political-agency foundations provided by Ferejohn (1986), Banks and Sundaram (1998), and Dewatripont et al. (1999), this literature has extended in many promising directions. Canes-Wrone et al. (2001) and Canes-Wrone and Shotts (2007) highlight the role of voter information in the quality of officials and their decisions on whether or not to pander. Caselli and Morelli (2004) analyze the trade-off between private-sector and public-sector returns in determining candidate quality. A closely related model is Ashworth and de Mesquita (2008), who explore how partisanship interferes with quality signals in the context of the incumbency advantage.

To summarize, suppose that judges are selected by a representative voter or governor who cares about both judge quality and their political affiliation. This has two immediate implications. First, given two judges with the same political views, the voters or governor prefer the higher-quality judge. Second, given two judges of the same quality but different political parties, the voters or governor prefer judges that are closer to them in their political views. If we suppose these preferences are smooth and continuous, then it follows that there is a trade-off between politics and quality.

These observations provide predictions on the causal effect of reforms that change the appointment system: partisan elections, nonpartisan elections, or merit selection. The distinguishing feature of partisan elections is that voters observe the political affiliation of each judge candidate, biasing voter judgments in favor of the judge from the same party. This bias creates a trade-off between the quality of the judge and the voters' preference of judges whose ideology is closer to the voter. A stronger partisan bias means lower-quality judges on average.

These points lead to our first prediction on the partisan-to-nonpartisan reforms. Moving to nonpartisan elections switches off partisan preferences. Regardless of the strength of those preferences, then, the partisan-to-nonpartisan reform leads to the selection of judges who on average have higher quality.

In the case of merit selection states, the pool of candidates nominated are selected by a group of experts. Hence, the governor should be selecting from a group of higher quality candidates than those nominated by political primaries. If we suppose that the political bias of the governor is the same as the voters (after all he was voted into office), then we should observe that the quality of judges in the merit system is higher than in the case of a partisan election.

In the case of non-partisan versus merit selection there are two margins, judicial quality and partisan preferences. While merit systems have technocratic nominations, they also have a strong role for the politically biased governor. Hence the theory cannot make an unambiguous prediction. In this case, only an empirical analysis can determine which system on average delivers better judicial quality.

These informal claims are made more precise in Appendix A, where we explore a simple information-based model of voting in the spirit of Condorcet (1785). In the model, a representative voter selects between two candidate judges. The judges are drawn from the same underlying distribution of skill. The voter observes informative signals about each candidate and then chooses the judge that would provide the greatest expected quality of work. Judges also have a party affiliation, which (when observed) imposes a bias in the voter's preferences on judge ability. As shown empirically by Lim and Snyder (2015), voters decide about judges using both partisanship information and bar evaluations of technical ability.

3 Measuring Judge Performance

3.1 Data Overview

For the empirical analysis, we use an updated and significantly extended version of the data set from Ash and MacLeod (2015). To build our data set, we merge information on judge biographies, state-level court institutions, and published judicial opinions. We check and extend the data set by adding or updating with more information relevant to elections. These data allow panel estimates on the effects of selection institutions on judge performance.

There are 1,558 state supreme court judges in our data, and 1265 when limiting to the three selection systems of interest (partisan/nonpartisan elections and merit selection). Table 2 reports summary statistics on the characteristics of judges working in one of the three main systems. Unsurprisingly, there are more judges with public party affiliations in the partisan system. For most of the other variables, the systems are comparable. Merit judges are the most likely to have judicial experience, while partisan judges are the most likely to have political experience (although these differences are not that large). Merit judges are the most likely to retire from office, probably due to stronger tenure protections in merit systems relative to elections.

We construct the performance measures from published state supreme court opinions for the years 1947 through 1994, obtained (along with some annotated meta-data) from bloomberglaw.com. The full sample includes 1,024,261 cases. We drop opinions that do not have a named author (per curium decisions), resulting in a sample of 404,928 majority opinions. After dropping judge-year observations with fewer than five cases, we have that each judge decides an average of 25.3 cases per year.

3.2 Construction of Performance Measures

An important step in this research is to provide an effective measure of judge performance. We focus on two simple metrics for judge performance, work output and work quality. The measures build off of previous work by Choi et al. (2010), Epstein et al. (2013), and Ash and MacLeod (2015).

The baseline measure of *work output* is the total number of words written by a judge in opinions during a year on the job, which is a measure of the total volume of opinion-writing work that a judge is responsible for in that year. As alternatives to assess robustness, we look at the number of sentences written and at the number of characters written.

The measure for *work quality* for a judge in a given year is the average number of forward citations to her authored opinions published that year – that is, the number of citations received, divided by the number of opinions authored. Judges in a common-law system cite previous cases that are useful to their decision. Therefore citations can be seen as an expert evaluation of peer decision quality (Posner, 2008). A higher citation count means that a case (and the authoring judge) have a stronger influence on the path of the law.

The citation counts include only cites to a judge's majority opinions, as citation counts to

Table 2: Summary Statistics on Judge Characteristics by Selection System

	Partisan Elections		*Nonpartisan Elections*		*Merit Selection*	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Politics						
Has Partisan Affiliation	1.0	0.0	0.36	0.48	0.29	0.45
Democrat (if affiliated)	0.7	0.46	0.52	0.50	0.71	0.46
Background						
Start Age	52.92	8.81	52.3	8.41	51.44	7.59
Female	0.03	0.16	0.06	0.23	0.08	0.28
Top School	0.14	0.35	0.14	0.34	0.17	0.38
Previous Experience						
Private Practice	0.8	0.4	0.84	0.37	0.79	0.42
Judiciary	0.54	0.5	0.48	0.50	0.65	0.47
Politics	0.31	0.46	0.25	0.43	0.14	0.35
Academia	0.11	0.31	0.10	0.30	0.14	0.35
Career Length	14.85	10.2	13.28	8.74	11.29	7.60
How Ended						
Retired	0.54	0.50	0.52	0.50	0.71	0.45
Died in Office	0.10	0.30	0.12	0.32	0.11	0.31
Stopped for Other Reasons	0.37	0.48	0.36	0.48	0.16	0.37
Lost Election	0.12	0.32	0.05	0.22	0.04	0.19
Judges (N)	714		361		190	

Notes. Biographical information by judge election system. Observation is a judge. Has Partisan Affiliation means we could find a publicly documented party affiliation for the judge. Democrat is a dummy for being Democrat, conditional on having a documented affiliation. Start Age is judge age upon joining the court. Female is a dummy for being female. Top School means the judge attended law school at Yale, Harvard, Columbia, Stanford, or Chicago. The Previous Experience items equal one if the judge has previous experience in the respective area. Career Length is number of years working on the court, conditional on having left the court before 2014. The How Ended items equal one if the judge's state supreme court judgeship ended for this reason.

discretionary opinions are not available.[12] Our citations data were collected in 2012, so the counts are cumulative up to that year. In a robustness check, we show similar results when limiting to citations from within ten years of the original case. The citations measure is per case (divided by the number of cases), so it is workload-adjusted.

While using citations to measure judge quality is well-established in the literature (e.g. Choi et al., 2008; Posner, 2008; Choi et al., 2010; Epstein et al., 2013; Ash and MacLeod, 2015), the approach is not without controversy (e.g. Baker et al., 2009). It is a quite limited measure of what judges do, which includes not just providing a high-quality precedent but also finding justice in a particular case. Citations might reflect dimensions of decisions other than quality, such as the topic or the partisan bias of the decision. Judges might try to influence their citations through other ways besides quality, such as social networking. We can only partly address these concerns, such as by looking at multiple measures, controlling for topics, and validating against other quality signals. In the empirical analysis, we are implicitly assuming that it is the quality factor of citations that is most systematically varying in response to how judges are selected.

In our data we have information on the "flavor" of the citation – positive, negative, or distinguishing (as annotated by the data provider). For the baseline we look only at positive citations. As an alternative measure, we use all types of cites, including negative and distinguishing cites. We also employ discussion cites, where the citing court discusses a case at length, and quote cites, where a future court directly quotes language. The out-of-state cites measure (only citations in other jurisdictions) is informative because it means that there is no opportunity for judges to cites themselves, or to seek cites from colleague judges. Also, because state supreme court precedents have no bindingness in other states, out-of-state citations provide a stronger signal of legal usefulness or influence (Choi et al., 2010). A drawback of the measure is that out-of-state cites are highly concentrated among a few important courts, such as California and New York, and are quite sparse for most courts.

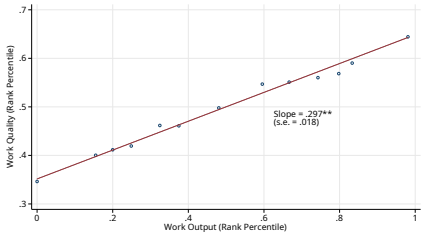
The goal of the analysis is to compare a judge's performance to other judges in the same court-year. We care about relative, rather than absolute, performance, mainly because relative performance is easier to measure. The volume of text a judge writes, or the number of citations a judge receives, cannot easily be translated into an interpretable quantity such as value in dollars. Relative differences between judges in these quantities can be measured and interpreted easily. In professions like judging, employment decisions are made based on relative performance (see Green and Stokey, 1983; Mookherjee, 1984). Therefore this is a relevant comparison from an economic-policy perspective.

In terms of getting at relative performance, a challenging feature of the data is that the distributions of the outcomes are extremely variable across courts and years, with significant outliers (see Appendix Figure OA.1). This high variability in the outcome variables means that the magnitudes of treatment effects are not comparable across states and over time. In particular, coefficients on

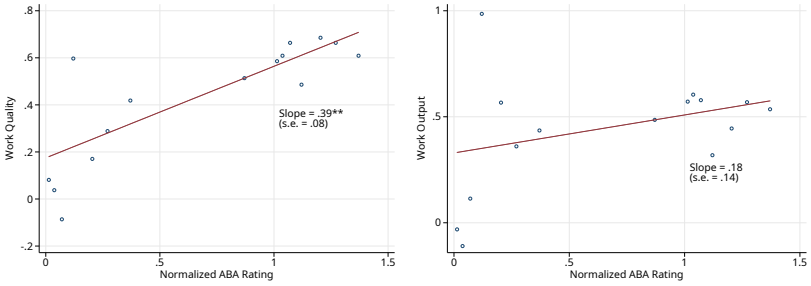
[12]From manual inspections, we found that citations to discretionary opinions are extremely rare in state supreme courts, much less than one out of a hundred citations.

Figure 1: Summary Statistics on Output and Quality

(a) Quality and Output are Positively Correlated



(b) Quality is Correlated with Bar Evaluations; Output is Not



Notes. Panel (a): Binscatter for Work Quality (percentile rank in positive citations per opinion, vertical axis) plotted conditional on Work Output (percentile rank in number of words written in opinions, horizontal axis). Panel (b): Binscatters for Work Quality (percentile rank in positive citations per opinion) and Work Output (percentile rank in number of words written in opinions) against bar association evaluation ratings. Estimated coefficients and standard errors from regression with court-year FE's and clustering by state and year).

treatments that affect different subsets of states are not comparable, as the different subsets have different outcome variance. Within-court variation in performance across judges could be lost in this across-court and across-year variation.

To address this issue, we follow recent work in empirical public finance on inequality and use a rank specification for the performance variables (e.g. Dahl and DeLeire, 2008; Chetty et al., 2014). Within a court-year group, we assign a value of one to the judge with the highest value, while the judge with the lowest value receives a zero. Each judge in between is uniformly distributed on that interval according to rank. This specification for judge performance nets out any level differences across the within-court-year distribution. By construction, the outcome distribution has a mean about one-half and a standard deviation about one-third, overall and in each court-year. As we will see below, rank percentiles do better than the citation counts in capturing a persistent judge factor in work quality.[13] Figure 1 Panel (a) shows that the rank percentiles of work output and work quality are positively correlated within court-year between judges.

To validate our outcome variables as judge-specific measures of performance, we explore the extent to which they are correlated with another performance measure previously used in the literature: the quality ratings issued by state bar associations. We are able to merge our data for a small number of judges with the data on evaluations provided by Lim and Snyder (2015). We

[13]The main empirical results are similar with the outcome variables in levels and logs (see Table 6).

then regressed our performance measures on the bar evaluations with state-year fixed effects, to see whether our quality and output measures are predictive of the bar association evaluations of a "good judge," as coded by the authors. Figure 1 Panel (b) shows that quality, but not output, is a strong predictor of judge qualifications as measured by bar association evaluations. The regressions reported in Appendix Table OA.3 provide statistical estimates showing the same.[14]

3.3 Case Characteristics

Our measure of work quality – relative citation count – is a joint outcome determined by both the judge's work input and the legal importance of the case. Some types of cases are more important than others – for example, cases that review the constitutionality of enacted statutes tend to be impactful, while routine cases denying habeas corpus tend not to be impactful. The former set of cases will get more citations than the latter set, regardless of the responsible authoring judge.

States vary in their rules for allocating cases to judges. Appendix Table OA.2 lists the states by the three official case assignment rules. Under random assignment cases are assigned by lottery. On the South Dakota Supreme Court, for example, "cases are randomly drawn and assigned from a cowboy hat".[15] Under rotating assignment, cases are assigned in a predetermined rotation which should be as good as random. In the analysis, we treat both "random" and "rotating" assignment as random assignment. Under discretionary assignment, finally, the chief judge chooses how cases are assigned.[16]

In the discretionary system, judges will likely rule on different types of cases as the chief judge tries to make assignments based on specialty and other factors. Under random or rotating assignment, there is still scope for selection, for example when judges recuse themselves due to a conflict of interest. As another example, when judges dissent with the majority they would not author the opinion. Christensen et al. (2012) show that there is some correlation between case types and judge types in all three systems, although it is strongest under discretionary assignment.

In all of the allocation systems, influence over the caseload is a potentially important factor in judge performance. This situation is not so different from that in other professions, where for example physicians usually have some choice over patients. In our empirical analysis we will examine how the selection-reform treatments affect the types of cases that judges decide.

To measure dimensions of case importance, we use rich information on the area of law and related industries of a case, annotated by Bloomberg staff attorneys. Each case has up to three legal areas and three related industrial sectors. Appendix Table OA.4 reports summary tabulations for the

[14]Note that the bar evaluations could not be used as an outcome in our empirical analysis because it is only available for recent years. Even if they were available, they might not be a good outcome because attorney evaluations of judges can reflect partisan political affiliations and ideological differences, rather than judicial performance (see Miles, 2015).

[15]The Virginia Supreme Court also draws cases out of a hat. See Christensen et al. (2012, pp. 18).

[16]There is another important dimension of case review in state supreme courts – that of mandatory or discretionary review. This rule says whether the high-court judges has to review appeals or not. Because review discretion does not vary across judges within a court, it does not play a part in our analysis.

most frequent legal areas and industrial sectors. The data set has a vector of case characteristics, in which each item is a dummy variable for each area and sector. Because there are so many of these characteristics, including separate covariates for every category would almost saturate the dataset. Instead, we represent case types as the first five principal components of this matrix of controls, which explains 65% of the variance of the matrix of case controls. We also assign legal topics into four broader categories: civil law, criminal law, administrative law, and constitutional law.

These case characteristics are important determinants of citations. Appendix Table OA.5 reports estimates from regressing citations on the case covariates: the four variables for broad legal topics, plus the five principal components for the legal topic and related industries. All of the four broad legal topics contribute significantly, the share of criminal cases most of all. Three of the first five principal components are also statistically predictive of citations.

3.4 Judges Vary in their Work Performance

In this paper we would like to compare judges in their work performance and relate that to selection procedures. Therefore an initial question is whether judges vary systematically in our outcome measures. This is not obvious. The previous work has shown that case quality measures vary across states (Choi et al., 2010) and over time (Ash and MacLeod, 2015). But these papers do not try to get at variation across judges within the same court. Judges work under many legal and institutional constraints, and clerks play an important role in writing opinions. Citations could be due to these factors rather than due to the judge's work input.

Appendix Figures OA.2 and OA.3 show some samples of the raw data for judge outcomes over time in select periods. The figures show that the performance ranking of the judges within the court is relatively stable across years. This trend is consistent with a judge-specific ability factor, as otherwise the judges would not maintain their relative performance level. The figures also show that the court-level distributions of the outcomes tends to shift significantly year-to-year, suggesting that rank percentiles could provide a more robust measure of judge quality than word counts or citation counts. Note, finally, that four of the six depicted states have random or rotating assignment (Georgia, Iowa, Nebraska, and Oklahoma), and they have just as clear separation of judges by citations over time as the discretionary-assignment states (Colorado and Kansas). The quality persistence in random-assignment states suggests that variation across judges is not driven just by differences in their caseload.

To show variation across judges statistically, we estimate the explanatory power of judge-specific fixed effects on citations per opinion y_{jct}. Formally, we compute the marginal R^2 of the judge fixed effects for y_{jct} (citations per opinion, in levels or in ranks) after residualizing out state-year fixed effects and case controls. These estimates are bootstrapped so we can interpret statistical differences in the explained variation (Ohtani, 2000).

The bootstrapped estimates for marginal R^2 of judge fixed effects for work quality are reported in Table 3. The outcome is the count of cites per opinion in Column 1, and the percentile rank of

Table 3: Explanatory Power of Judge Fixed Effects for Work Quality

	(1)	(2)	(3)	(4)	(5)
	Cites per Case	Percentile Rank in Cites per Case			
Marginal R^2 on	0.197	0.331	0.297	0.314	0.373
Judge FE's	(0.0384)	(0.0064)	(0.0065)	(0.0073)	(0.0122)
N	15004	14996	14996	10852	4144
Allocation Rule				Random	Non-Random
State-Year FE's	X	X	X	X	X
Case Controls			X		

Notes. Bootstrapped estimates of the R^2 of the judge fixed effects on the stated outcome after residualizing out fixed effects and controls. "Cites per Case" means citations per opinion in a year; "Percentile Rank in Cites per Case" means judges are uniformly distributed between zero and one based on rank within court-year (0 is lowest, 1 is highest). Standard error of bootstrapped estimate in parentheses. 128 bootstrap samples. Case Controls include controls for four major case types, five principal components of matrix of legal topics and related industries, and judge percentile in the number of cases seen, all fully interacted with both state fixed effects and year fixed effects. Allocation Rule means the sample is limited to Random or Non-Random states, as indicated (see lists in Appendix Table OA.2).

cites per opinion in Columns 2 through 5. In all columns, the outcome is residualized on state-year fixed effects.

The first observation is that judge fixed effects explain significantly more variation for the rank outcome (Column 2) than for the count outcome (Column 1). Especially for ranks (our preferred outcome specification), judge fixed effects have significant explanatory power: about one-third of the remaining variation after residualizing out state-year effects. The standard error is also much smaller for quality ranks relative to the count outcome, reflecting a much stabler measure of judge performance.

In Column 3, the quality measure is also residualized on case characteristics – legal area and related industry, as discussed above in Section 3.3. We can see that the variation across judges is not substantially driven by case characteristics. While residualizing out case features first does reduce the marginal R^2 of judge fixed effects (from .33 to .30), it is a proportionally small difference.

Columns 4 and 5 limit the regressions to, respectively, states with random and non-random assignment of cases to judges (see Appendix Table OA.2). Judge fixed effects have more explanatory power under discretionary assignment (Column 5) than for random assignment (Column 4). This difference is intuitive because under discretionary assignment, judges can endogenously specialize in different types of cases. Specialization would be reflected in more distinctive portfolios by judge, so case characteristics would be partly picked up by the judge effects.

Appendix Table OA.6 shows an analogous table for judge work output. Again, judge fixed effects explain significant variation (about 40%). For output, the rank measure is not better-explained by judge fixed effects than the count measure.

Table 4: Persistence in Judge Quality

	(1)	(2)	(3)	(4)	(5)	(6)
	Cites per Case			Percentile Rank in Cites per Case		
	Panel OLS		Arellano-Bond	Panel OLS		Arellano-Bond
Lagged Outcome	0.533	-0.0143	0.0111	0.357	0.0259	0.132
(s.e.)	(0.053)	(0.012)	(0.035)	(0.020)	(0.013)	(0.015)
[p-value]	[0.00]	[0.25]	[0.75]	[0.00]	[0.057]	[0.00]
N	13446	13320	11773	13435	13310	11765
Year FE's	X	X	*n/a*	X	X	*n/a*
Judge FE's		X	*n/a*		X	*n/a*

Notes. Regression estimates for Equation (3.1). Observation is a judge working in a year. "Cites per Case" means citations per opinion in a year; "Percentile Rank in Cites per Case" means judges are uniformly distributed between zero and one based on rank within court-year (0 is lowest, 1 is highest). Estimates computed with Panel OLS and Arellano-Bond, as indicated. Standard errors in parentheses – for OLS, clustered by state and year, for AB, with robust. P-values in brackets.

To show judge persistence in work quality over time, we explicitly estimate

$$y_{jct} = \alpha_{jct} + \rho y_{jct-1} + \epsilon_{jct}, \tag{3.1}$$

a first-order auto-regressive model for judge j in court c at year t, where α_{jct} could include fixed effects (year and judge). The coefficient ρ summarizes the across-year within-judge correlation in performance measure y_{jct}. An estimate of ρ that is statistically greater than zero means that judges tend to maintain their within-court performance ranking over time.

In the OLS estimates for (3.1), the errors are correlated and ρ will be biased. The preferred specification is the Arellano-Bond (1991) GMM estimator, which instruments the lagged outcome with the twice-lagged outcome. This estimator is designed to net out mechanical effects due to serial correlation in the outcome.

Estimates for Equation (3.1) for judge work quality are reported in Table 4. Columns 1 through 3 have positive citations per opinion (counts) as y_{jct}, while Columns 4 through 6 have the percentile rank. For each outcome, we report OLS estimates with/without judge fixed effects, as well as the Arellano-Bond estimates. We see that there is indeed a within-judge persistence in citations, which is captured much more strongly by the the percentile rank outcome. This can be seen, especially, in the p-value for the Arellano-Bond estimates, which are highly significant with ranks (Column 6) but not significant for the count outcome (Column 3).[17] Adding case characteristics as covariates

[17]Part of this difference is mechanical, in that the distribution is normalized across courts and years in the percentile specification. To address this issue, we run the same regressions with the citation count standardized by court-year (divided by the standard deviation of judges in that year). As expected, the persistence is higher than the non-standardized measure. The t-statistic ($t = 2.67$) is significant, but much smaller than that for percentile ranks ($t = 8.75$).

to these regressions does not change the persistence estimates.

The analogous results for work output are reported in Appendix Table **??**. They show large and significant persistence for all specifications. Overall, these statistics support the hypothesis that our percentile-rank measures are capturing relative differences between judges in work quality and work output.

4 Econometrics

4.1 Empirical Approach

Our empirical strategy relies on reforms to the judge selection process, which we use as a set of natural experiments (see Appendix Table 1). There are five methods for selecting state appellate judges: partisan elections, nonpartisan elections, merit selection, governor selection, and legislative selection. There are an insufficient number of reforms for governor selection and legislative selection, so those systems are not a part of our analysis. We use reforms between these three systems: partisan elections, nonpartisan elections, and merit selection.[18]

In the time period of our data, three states change from partisan selection to nonpartisan selection: Georgia, Kentucky, and Utah. Seven states moved from partisan selection to merit selection: Colorado, Iowa, Indiana, Kansas, Nebraska, Oklahoma, and Tennessee.[19] Five states move from nonpartisan selection to merit selection: Arizona, Maryland, South Dakota, Utah, and Wyoming.[20]

The goal is to compare the performance of judges selected before these reforms to the performance of judges selected after these reforms. Because the number of cases per judge varies a lot across courts and over time, the data is collapsed to the judge-year level to make changes in performance more comparable across courts. We represent the treatment effect of these reforms as indicator variables for judges selected after the reform. Appendix Table OA.1 summarizes the relevant treatment variation by tabulating the court-years, judges, and judge-years for each reform.

Our specification flexibly adjusts for time-varying state-specific factors by including a full set of court-year (interacted) fixed effects. This specification effectively compares the performance of judges sitting on the same court at the same time, but selected under different regimes. Formally, we model outcome y_{jct} for judge j in court c at year t as

$$y_{jct} = \gamma_{ct} + X'_{jct}\beta + S'_j\rho + \epsilon_{jct} \tag{4.1}$$

[18] Admittedly, assigning the appointment systems into three categories ignores many (potentially important) institutional details (see, e.g., Posner, 2008). We have to focus on the basic features of the institutional environment for which we have sufficient variation.

[19] Tennessee moved to merit selection in 1972 but moved back to partisan selection in 1978.

[20] Florida moved from nonpartisan to merit five years after a partisan-to-nonpartisan reform. Only four judges were selected under the nonpartisan system. Therefore Florida is excluded from the baseline selection regressions, but including it does not change the results.

where γ_{ct} includes the court-year fixed effects and X_{jct} includes time-varying judge-level covariates (to be described further below). The vector S_j includes a set of treatment indicators for each electoral reform (partisan-to-nonpartisan, partisan-to-merit, and nonpartisan-to-merit), equaling one for judges selected after the reform in the respective states, and equaling zero otherwise.[21] Given the inclusion of the fixed effects, the coefficients ρ procure the average difference in performance between judges selected under the new system and judges selected under the old system, after adjusting for the other included covariates. Standard errors are clustered by state and year, which yields smaller standard errors than when clustering by state, clustering by state-year, or not clustering.

4.2 Identification

Consistent estimates of (4.1) depend on the standard identification assumptions for panel data. We require that, conditional on the fixed effects and controls, S_j is not correlated with ϵ_{jct}. That is, there are no unobserved outcome-relevant factors between colleague judges (on the same court at the same time) who are selected under different systems.

There are four major threats to identification. First, there is the problem of confounding trends in judge citation rates. Judge quality could be systematically changing over time in the aggregate, so judges selected after reforms could have mechanically different quality levels. Also, states that adopt new selection systems might already be increasing the quality of their judges for other reasons.

We address confounding trends in quality with judge cohort fixed effects and state-specific cohort trends. Formally, in X_{jct} we include fixed effects for judge cohort (decade of birth and decade joining the court). We also include state-specific linear trends in judge cohort. These covariates have a similar purpose to year fixed effects and state-specific trends in a standard differences-in-differences analysis.

A second threat to identification is differences in judge experience. Judges selected after a reform, by construction, are less experienced than judges selected before a reform. To control for judge experience, we include in X_{jct} a set of fixed effects for number of years on the court. With the inclusion of these experience fixed effects, our estimates are identified off of differences between judges of the same experience level.

A third threat is selective attrition in response to the reform. Pre-reform judges of higher or lower quality might tend to leave the court earlier or later in response to the reform. This would bias our resulting estimates of differences between pre-reform and post-reform judges. In Appendix Table A.1, we show that this type of bias is limited. While the election-system reforms tend to increase judge career length on average, the career-extending effect is not significantly different for judges with higher or lower pre-reform quality.

[21] Note that in the electoral selection systems, the judges may be initially appointed by the governor to fill a vacant seat, rather than being initially selected through a competitive electoral process. We still code the appointed judges as being selected under the electoral system – since the predecessor's choice whether to step down is endogenous to the system.

Fourth and finally, a threat to identification is the endogenous assignment of cases. Citations are a combined product of case importance and decision quality (see Section 3.3), and judges selected under different systems might be assigned different types of cases. Therefore, selection-reform effects on citations could be due in part to post-reform judges ruling on more or less important cases than pre-reform judges. Similarly, post-reform judges may get assigned more or fewer cases overall.

We seek to understand the role of case types in three ways. First, we use judge-level data on case type (described in Subsection 3.2) as an additional set of controls in X_{jct} when estimating (4.1). We include controls for case type, as well as controls for caseload size relative to colleagues, interacted with state fixed effects and year fixed effects. Second, we show whether there are statistically significant effects of our reforms on the types of cases that judges decide on. Third, we use the fact that a majority of state supreme courts have random or rotating assignment. We check whether our results hold when limiting estimates to random-assignment states.

5 Results

5.1 Main Results

Table 5 reports the regression results for effects of the selection system reforms on work quality and work output. The coefficient estimate $\hat{\rho}$ from Equation (4.1) gives the statistical difference between post-reform and pre-reform judges, on the same court at the same time, in terms of their percentile rank for the outcome. The three treatments contained in S_j, listed from top to bottom in Table 5, are the partisan-to-merit selection reform, the partisan-to-nonpartisan selection reform, and the nonpartisan-to-merit selection reform. In Columns 1 through 3, the outcome is work quality (percentile rank in positive forward citations per case). In Columns 4 through 6, the outcome is work output (percentile rank in the total number of words written).

Columns 1 and 4 provide the baseline with court-year fixed effects. Relative to their pre-reform partisan-selected colleagues, post-reform merit-selected judges (top row) and post-reform nonpartisan-selected judges (middle row) have higher work quality and output. Meanwhile, there is no difference (and if anything a negative coefficient) in the performance measures between nonpartisan-selected judges and merit-selected judges (bottom row).

In the subsequent columns (2, 3, 5, and 6) we modify the regression specification to tackle two identification issues described in Subsection 4.2 above. First, we address the issue of confounding trends in quality nationally or at the state level in states that undertook selection reforms. Columns 2 and 5 add fixed effects for nationwide judge cohort, as well as state-specific trends in judge cohort. This modification is specially relevant for the partisan-to-merit reform. The partisan-to-merit effects decrease to about half, and the effect on work output is no longer significant. The cohort controls do not change the partisan-to-nonpartisan estimates much.

Second, Columns 3 and 6 add fixed effects for the judge's years of tenure on the court to deal

Table 5: Effects of Selection System Reforms on Judge Performance

	(1)	(2)	(3)	(4)	(5)	(6)
	Work Quality			*Work Output*		
Partisan	0.178**	0.0982*	0.0831*	0.0632^{+}	0.0379	0.0280
to Merit	(0.0430)	(0.0402)	(0.0409)	(0.0348)	(0.0394)	(0.0421)
Partisan to	0.101*	0.0956^{+}	0.0866^{+}	0.0886**	0.0812^{+}	0.100*
Nonpartisan	(0.0477)	(0.0530)	(0.0473)	(0.0139)	(0.0463)	(0.0477)
Nonpartisan	-0.0686	-0.0887	-0.107	-0.0629	-0.0602	-0.0717
to Merit	(0.136)	(0.153)	(0.155)	(0.0732)	(0.128)	(0.131)
N	14996	14894	14890	14996	14894	14890
R^2	0.004	0.046	0.051	0.001	0.024	0.058
Court-Year FE	X	X	X	X	X	X
Cohort FE / Trends		X	X		X	X
Experience FE			X			X

Notes. Estimates of ρ from Equation (4.1): the average difference between judges selected under a new system, relative to judges selected under the old system – e.g., "Partisan to Merit" means the merit-judge outcome relative to the partisan-judge outcome. "Work Quality" is citations per opinion in a year. "Work Output" is the number of words written in majority opinions in a year. Outcomes are rank percentiles, where the lowest-value in a court-year group gets a zero, the highest-value gets a one, and other judges are distributed uniformly in between according to rank. Court-Year FE refers to court-year interacted fixed effects. Cohort FE's / Trends includes fixed effects for decade that a judge started on a court and judge birth, plus state-specific trends in judge starting cohort. Experience FE includes fixed effects for years of experience on the court. Standard errors, adjusted for two-way clustering by state and year, in parentheses. +p<.1, * p<0.05, ** p<0.01.

Table 6: Effects on Alternative Work Quality Measures

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Positive Cites Per Case			*Other Specs for Cites Per Case (Rank)*			
	in Levels	*in Logs*	*Rank (10 yr)*	*Pos+Neg*	*Discuss*	*Quote*	*Out-State*
Partisan	0.887^{+}	0.0654^{+}	0.0767^{+}	0.0943*	0.0953**	0.131**	0.0842*
to Merit	(0.499)	(0.0371)	(0.0444)	(0.0356)	(0.0280)	(0.0325)	(0.0320)
Partisan to	0.650	0.0814^{+}	0.0574	0.0572	0.0968*	0.0822	0.0531
Nonpartisan	(1.018)	(0.0452)	(0.0424)	(0.0537)	(0.0445)	(0.0705)	(0.0325)
Nonpartisan	-0.650	-0.0519	-0.103	-0.0838	-0.0895	-0.0454	-0.0333
to Merit	(1.910)	(0.101)	(0.115)	(0.141)	(0.115)	(0.105)	(0.0751)
N	14894	14894	14894	14894	14894	14894	14894
R^2	0.618	0.807	0.043	0.045	0.039	0.040	0.027
Court-Year FE's	X	X	X	X	X	X	X
Cohort FE / Trends	X	X	X	X	X	X	X

Notes. Estimates of ρ from Equation (4.1): the average difference between judges selected under a new system, relative to judges selected under the old system – e.g., "Partisan to Merit" means the merit-judge outcome relative to the partisan-judge outcome. The first three columns have positive citations per case as the outcome: "in Levels" means the regression uses the unadjusted judge-year level outcome; " in Logs" means the outcome is transformed by the log of 1 plus the unadjusted value; "Rank (10 yr)" refers to the percentile outcome, but only including citations within the next ten years after a case is published. "Pos + Neg" includes all citations (not just positive, but also negative and distinguishing). "Discuss" only includes cites in which the case is discussed and applied. "Quote" only includes cites where the case's language is directly quoted. "Out-State" includes only citations from courts in other states. For Columns 3 through 7, outcomes are rank percentiles, in which the lowest-value in a court-year group gets a zero, the highest-value gets a one, and other judges are distributed uniformly in between according to rank. Court-Year FE refers to court-year interacted fixed effects. Cohort FE's / Trends includes fixed effects for the decade that a judge started on a court and judge birth, plus state-specific trends in judge starting cohort. Standard errors, adjusted for two-way clustering by state and year, in parentheses. +p<.1, * p<0.05, ** p<0.01.

with the issue that post-reform judges have less experience mechanically than pre-reform judges. These fixed effects do not change the estimates. Therefore the estimated treatment effects are not driven by differences in experience levels.

Focusing on Columns 2 and 3, we can see that the post-reform judges (merit or nonpartisan judges, relative to partisan judges) score about .08 to .1 higher in rank percentiles than their pre-reform colleagues. This effect size is interpretable as about one ranking spot on average on a nine-judge panel. It is about one-fourth to one-third of a standard deviation in the outcome.

Table 6 shows additional regressions with alternative work quality outcomes. The specification includes court-year fixed effects, cohort fixed effects, and state cohort trends (equivalent to Table 5 Column 2).[22] We alter the outcome in two ways: changing how the measure of positive citations per opinion is constructed (Columns 1 through 3), and using alternative citation counts (besides just counting positive cites) (Columns 4 through 7).

First, Column 1 has the raw data on the outcome: the count of positive citations per opinion in levels. Second, Column 2 has the log of positive citations per opinion – these coefficients can be interpreted as proportional changes, showing that partisan-selected judges get 6 to 8 percent fewer citations per opinion than their merit-selected or nonpartisan-selected colleagues. Column 3 uses the baseline measure (rank percentile for positive cites per opinion) but limiting only to citations that occur within ten years of a case (rather than up through 2012 when we collected the data). This specification adjusts for differences across years due to more chances for a case to be cited.

[22] Appendix Table A.3 reports all of these estimates, including experience fixed effects.

Column 4 has the rank percentile for all citations (not just positive ones) per opinion, meaning that the count includes positive and negative citations. This modification results in a more inclusive measure that does not rely on Bloomberg staff attorneys distinguishing the tone of cites. Column 5 is a more exclusive measure of citation – whether a case is discussed and specifically applied in a subsequent decision. Column 6 counts the number of times a previous case is directly quoted, indicating that the writing style is worth repeating. Finally, Column 7 includes positive out-of-state citations by courts in other states. This measure is important because it excludes judges citing themselves or seeking citations from colleagues, and because state supreme court precedents are not binding on other states.

The top row of estimates shows that overall, the specification for the outcome variable does not matter for the partisan-to-merit reform. Along all of these quality measures, merit judges provide higher quality work than their partisan-selected colleagues. On the other hand, the specification does matter for the partisan-to-nonpartisan reform (middle row of estimates). While the coefficients are consistently positive in magnitude, the estimated effect is statistically significant in only two of the seven specifications.

Appendix Table A.2 reports analogous results for other specifications for work output. It shows that for any other specifications for output, there are no effects for any of the reforms. Therefore there is no effect of a judge selection procedure on work output.

To summarize, the judicial selection reforms that changed partisan to merit produced judges that provided higher quality work than their pre-reform colleagues. The effect is not driven by correlated changes in the judge cohort, nor is it driven by mechanical differences in judge experience. The effect is robust to an array of alternative specifications for work quality. The partisan-to-nonpartisan reform also increased work quality, but that effect is less robust to other definitions of work quality. The nonpartisan-to-merit reform effect is consistently zero (yet noisy) across all regression models.

5.2 Relevance of Case Characteristics

As discussed in Subsection 4.2 above, a potential mechanism for the effects of reforms on citation rates is differential case assignment across judges. The effects of the selection systems estimated in Subsection 5.1 could be driven by post-reform judges getting different types of cases as pre-reform judges. For example, if merit judges have higher intrinsic motivation for influencing the law than partisan judges, they might tend to rule on more important cases that get more cites due to their importance rather than due to the judge's work quality.

Table 7 provides checks along these lines. First, Column 1 provides the main regression specification for the effect of the selection reforms on rank percentile in positive citations per opinion with court-year fixed effects and cohort effects/trends (Table 5 Column 2). The difference from the baseline specification is that we add controls for caseload characteristics. These include the share of cases across broad legal topics, the principal components for detailed legal topic and related industries, and the number of opinions written (rank percentile). All of these covariates are fully

Table 7: Relevance of Case Characteristics

	(1)	(2)	(3)	(4)	(5)
	Work Quality		*Caseload*	*Share*	*Case*
	(with Case Controls)	*(Rand Assign States)*	*Size*	*Criminal Cases*	*Importance*
Partisan	0.0902*	0.122**	0.00507	-0.0304	-0.00293
to Merit	(0.0413)	(0.0306)	(0.0401)	(0.0239)	(0.00406)
Partisan to	0.0870+	0.0514	0.0681**	-0.0508	0.00352**
Nonpartisan	(0.0478)	(0.0538)	(0.0195)	(0.0720)	(0.00126)
Nonpartisan	-0.0479	-0.0563	0.0465	-0.0345	-0.0146
to Merit	(0.148)	(0.0507)	(0.128)	(0.0619)	(0.0158)
N	14894	10833	14894	14894	14894
R^2	0.163	0.046	0.028	0.061	0.665
Court-Year FE	X	X	X	X	X
Cohort FE / Trends	X	X	X	X	X
Case Controls	X				

Notes. Estimates of ρ from Equation (4.1): the average difference between judges selected under a new system, relative to judges selected under the old system – e.g., "Partisan to Merit" means the merit-judge outcome relative to the partisan-judge outcome. Outcomes are rank percentiles, in which the lowest-value in a court-year group gets a zero, the highest-value gets a one, and other judges are distributed uniformly in between according to rank. Column 2 regressions are limited to random assignment states: Georgia (P-to-NP), Iowa (P-to-M), Nebraska (P-to-M), Oklahoma (P-to-M), Tennessee (P-to-M), South Dakota (NP-to-M), and Utah (P-to-NP). Work quality is positive citations per opinion. Caseload size is the number of authored opinions. Share criminal cases is the proportion of cases on the broad criminal law topic. Case importance is the linear prediction for positive cites per opinion based on case features, as discussed in the text. Court-Year FE refers to court-year interacted fixed effects. Cohort FE's / Trends includes fixed effects for the decade that a judge started on a court and judge birth, plus state-specific trends in judge starting cohort. Case Controls include controls for four major case types, five principal components of matrix of legal topics and related industries, and judge percentile in the number of cases seen, all fully interacted with both state fixed effects and year fixed effects. Standard errors, adjusted for two-way clustering by state and year, in parentheses. $^{+}$ p<.1, * p<0.05, ** p<0.01.

interacted with the state fixed effects and year fixed effects, allowing for different effects on quality by state and over time. With the inclusion of these controls, the effects of the reforms on citation ranks are still positive and significant.

Next, in Column 2 we limit the analysis to the states with random or rotating assignment of judges. By official court rules, in these states, judges should not have systematically different case portfolios. While it is a smaller sample, the effect of the partisan-to-merit reform (top row) is robust and a bit stronger in the random-assignment states. The partisan-to-nonpartisan reform (middle row), however, is smaller and no longer significant. There is no difference in the nonpartisan-to-merit reform (bottom row).

Column 3 goes back to the full sample and looks at the effect of the reforms on caseload size. The outcome, in this regression, is the within-court-year rank percentile of the number of majority opinions on which the judge has authoring responsibility. Reassuringly, the top row shows that the post-reform merit judges and pre-reform partisan judges have identically sized caseloads. However, in the partisan-to-nonpartisan reform (middle row), there is a large and significant difference between judges selected pre- and post-reform. Post-reform judges get more cases to rule on than pre-reform judges, so an important factor in observed effects of this reform is the caseload.

As discussed in Section 3.3, citations depend in part on case type features, especially the share of criminal cases. We use this in our empirical analysis, first, by putting the share of criminal cases as an outcome for the selection-reform regression. Table 7 Column 4 reports these results. None

of the reforms affects the share of criminal cases in a selected judge's case portfolio (there are no effects on other case types either).

Finally, we form a measure of case importance as the linear prediction for citation rates based on these case characteristics. We then use that as an outcome in Table 7 Column 5. This regression evaluates whether the cases predicted to be important based on exogenous features are more or less likely to be assigned to post-reform judges. The top row in Column 5 shows that there is no effect of the partisan-to-merit reform on case importance, consistent with non-selected portfolios.

In the middle row, however, we see a positive and significant effect for the partisan-to-nonpartisan reform. This estimate means that, similar to the results obtained with caseload size (Column 3), we do not have equivalent caseloads in the case of partisan-to-nonpartisan reforms. Post-reform judges are ruling on more important cases than their pre-reform colleagues. An interpretation of this result is that the chief judge responds to different ability or motivation of non-partisan judges by assigning them legally more important cases.

Summing up, the estimates for the effect on forward citations of partisan states moving to merit selections are not driven by post-reform judges selecting into different types of cases. However, the effect of partisan-to-nonpartisan reforms is driven at least in part by case characteristics. All results for the partisan-to-nonpartisan reform should be interpreted with this mechanism in mind.

5.3 Additional Results and Robustness Checks

Appendix Section B reports additional results and robustness checks. We show that there is no evidence of selective attrition (changes in career length) among control judges according to pre-reform work quality (Appendix Table A.1). Appendix Table A.2 explores different specifications for work output, while Appendix Table A.3 shows the results on alternative quality measures (same as Table 6 above) with experience fixed effects.

Appendix Table A.4 reports a collection of robustness checks. The regressions show that the main results on the partisan-to-merit reform and work quality hold while dropping the first and last year of a judge's career. The main results do not change when adding more principal components to the vector of case controls, controlling for state-government expenditures on the judicial branch in a judge's starting year, controlling for election-cycle timing, or allowing for cohort-specific judge-experience trends. The main results remain significant when weighting regressions by caseload size, weighting by inverse career length (to treat judges equally), subsetting to a window of years before and after the selection reform, alternative clustering of standard errors. Appendix Table A.5 shows robustness to dropping each treated state individually.[23]

Overall, the partisan-to-merit effect on quality is robust, while the partisan-to-nonpartisan effect often becomes insignificant (but is consistently positive in magnitude). If we pool the treatments

[23]In addition, we implemented the statistical test from Oster (2019) on the baseline specification for the partisan-to-merit reform using the cohort and experience controls as observables. On the standard parameters, the test implies that selection on unobservables would have to be 1.8 times as large as selection on unobservables to reduce the effect to zero.

and treat partisan-to-merit and partisan-to-nonpartisan as a single reform, we estimate robust positive effects on judge quality that are qualitatively similar to the partisan-to-merit results. Results are also robust to including New York's 1976 reform (moving from partisan elections to governor appointment) as part of this pooled treatment.

In the appendix, we report event study estimates for the effect of the partisan-to-merit reform on judge quality. That is, we estimate the differences in judge quality by judge starting year, relative to starting in the year the reform was enacted. We include a window of three years before the reform, up until ten years after the reform. Formally, we have

$$y_{jct} = \gamma_{ct} + X'_{jct}\beta + \sum_{k \in K} S_j^k \rho_k + \epsilon_{jct} \tag{5.1}$$

where y_{jct} is rank percentile in total out-of-state citations for judge j in court c at year t, γ_{ct} include court-year fixed effects (for the current year where performance is measured), and X_{jct} includes starting year fixed effects and state-specific starting year trends. The main difference from the earlier specification (Equation 4.1) is the summation term, which includes a set of indicator variables S_j^k equaling one for judge j starting k years after the partisan-to-merit selection reform. The set of years K includes $k \in \{-3, -2, -1, 1, ..., 10\}$, as $k = 0$ (the reform enactment year) is the left-out estimation baseline. For example, if a judge started two years after the reform, $S_j^2 = 1$ and all other items in $S_j^k = 0$. The estimate $\hat{\rho}_2$ summarizes the difference in the outcome for set of judges starting two years after these reforms, relative to judges starting the year the reform is enacted (the last cohort under the old system).

Appendix Figure A.1 visualizes the event study estimates in a coefficient plot. There is no evidence of a pre-trend, and a clear positive jump starting the second year after the reform. The pre-reform coefficients are not jointly significant ($p = .29$), while the post-reform coefficients are highly jointly significant ($p < .0001$).

The event-study results are in the appendix because they are not that robust, relative to the previously reported results. In particular, while the main results are robust to the choice of the outcome variable, for the event study we only see consistent positive estimates when using total out-of-state citations per judge-year. When using the other outcome variables (total positive citations, positive citations per opinion, or out-of-state citations per opinion), the coefficients bounce around a lot more and we see both negative and positive coefficients in the post period. Still, the estimates are jointly significant in post periods and jointly insignificant in the pre-periods. Meanwhile, the event-study estimates for the partisan-to-nonpartisan reform are too noisy to be interpretable, although the pooled treatment effect (treating partisan-to-merit and partisan-to-nonpartisan as a single reform) generates qualitatively similar results to those in Appendix Figure A.1.

Finally, Appendix Table A.6 reports the main regression specification with other outcomes on the left-hand side. These include the rate at which a judge affirms a lower-court case, average opinion length, the intensity of caselaw research, dissent rate, total citations (not just per opinion),

the rate at which a judge is overruled by a later court, and the rate at which cases are superseded by statute. The partisan-to-merit reform does not have a statistically significant effect on any of these dimensions of judicial behavior. For the partisan-to-nonpartisan reform, there is a significant positive effect on total cites (reflecting that post-reform judges have larger caseloads than pre-reform judges) and a negative effect on being overruled (although this is a very sparse outcome).

5.4 How Reforms Changed Judge Characteristics

A lingering question is: Can the changes in work quality due to selection reforms be explained by changes in the observable characteristics of judges? For example, are merit-selected judges more likely to attend top-ranked Ivy League law schools? We use our detailed data on judge biographical characteristics to examine this issue.

The regression specification is slightly different, as we are interested in seeing which judge biographical characteristics are predictive of the judge selection system. We estimate

$$S_{jct} = \gamma_{ct} + X'_{jct}\beta + \epsilon_{jct} \tag{5.2}$$

where S_{jct} is an indicator for how judge j was selected (a separate outcome for each system) and γ_{ct} includes the court-year fixed effects. X_{jct} includes observable judge characteristics which may be relevant to work quality: whether the judge has a partisan affiliation, the starting age, gender, whether the judge attended a top-ranked law school, and whether the judge has experience in a previous judgeship, private practice, or academia. We estimate the elements of $\hat{\beta}$ to see which judge characteristics are selected for by the different appointment systems. In addition, we re-weight estimates to adjust for career length and treat judges equally.

Table 8 reports the estimates for equation (5.2) . Columns 1 through 3 present these estimates with the respective reform dummies as the dependent variable: partisan-to-merit (Column 1), partisan-to-nonpartisan (Column 2), and nonpartisan-to-merit (Column 3). As expected, the partisan-to-merit and partisan-to-nonpartisan reforms reduce the share of judges with a public party affiliation. For the nonpartisan-to-merit reform, unsurprisingly, there is no effect. The merit reforms (Columns 1 and 3) both increase starting age, suggesting they select for judges with more advanced, mature careers. The merit systems select for more female judges and more judges from top-ranked law schools. The partisan-to-nonpartisan reform, in contrast, selects for younger judges and fewer females.

In terms of previous work experience, there is no significant correlation for the partisan-to-merit reform. The post-reform partisan-to-nonpartisan judges are less likely to have judging experience or experience in private practice. The post-reform nonpartisan-to-merit judges were more likely to be in private practice.

To help interpret these differences, in Column 4, we run the same regression with out-of-state

Table 8: Relevance of Judge Characteristics

	(1 *Partisan to Merit*	(2) *Partisan to Nonpartisan*	(3) *Nonpartisan to Merit*	(4) *Out-of-State Cites Per Case*
Party-Affiliated	-0.0199**	-0.0052**	0.0019	0.0065
	(0.0024)	(0.0015)	(0.0019)	(0.0086)
Starting Age	0.0011**	-0.0003**	0.0006**	-0.0017+
	(0.0001)	(0.0001)	(0.0001)	(0.0009)
Female	0.0093**	-0.0016**	0.0132**	0.0829**
	(0.0021)	(0.0005)	(0.0047)	(0.0232)
Top School	0.0037*	0.0004	0.0026^{+}	0.0208+
	(0.0016)	(0.0005)	(0.0014)	(0.0113)
Previous Experience:				
– as Judge	0.0001	-0.0004^{+}	-0.0003	0.0161+
	(0.0023)	(0.0002)	(0.0018)	(0.0093)
– in Private Practice	0.0001	-0.0015*	0.0098**	0.0314**
	(0.002)	(0.0007)	(0.002)	(0.0102)
– in Academia	0.0006	-0.0011	0.0023	0.0165
	(0.0016)	(0.0011)	(0.0019)	(0.0144)
N	14983	14983	14983	13471
R^2	0.884	0.917	0.841	0.055
Court-Year FE's	X	X	X	X

Notes. Estimates of β from Equation (5.2): the OLS correlations of the indicated dependent variable (columns) with the indicated judge characteristics (rows) as the right-side variables. Party-Affiliated is a dummy for whether public partisan information could be found; Starting Age is the age at which the judge starts on the supreme court; Female is a dummy for female gender; Top School is a dummy for attending a top-ten law school; Previous Experience includes indicators for previous experience in another judgeship, at a law firm, or teaching at a law school, respectively. Outcomes include post-reform dummies for the selection procedure reforms (Columns 1 through 3) and the percentile rank for out-of-state citations per opinion (Column 4). In Columns 1 through 3, regressions include missing dummies in case any judge biographical characteristics are missing, and they are weighted to adjust for judge career length. Column 4 only includes untreated judges and additionally controls for judge cohort and case importance. Standard errors clustered by state-year in parentheses. + p<.1, * p<0.05, ** p<0.01.

citations per opinion as the dependent variable.[24] The results indicate that female, top-school, former-judge, and former-private-practice are associated with more citations. Higher starting age is associated with lower work quality.

These correlations can partly explain the estimated selection-reform effects. In particular, the partisan-to-merit reform selects for more female, top-school judges, both characteristics which are associated with more citations. On the other hand, the effect on starting age goes in the wrong direction.

So overall, the correlations of characteristics with work quality do not match with the correlations with selection system reforms. There seems to be a significant unobserved factor for quality that is not captured by these characteristics. Consistent with that, our main results of the effect of selection reforms on quality are robust to including these judge characteristics as controls (Appendix Table A.4 Column 7).

6 Discussion

Now we can relate out empirical results to the information-theoretic approach outlined in Section 2. That model provides two relevant predictions. First, procedures that use more precise information should select for higher-quality judges on average. Second, procedures that have lower partisan bias should select for higher-quality judges on average.

Evidence for the first prediction is in the effects of the partisan-to-merit reform. We have robust evidence that post-reform merit-selected judges provide higher work quality than their pre-reform partisan-selected colleagues. The merit-selected judges review the same types of cases and do about the same amount of work (similar caseload and output) as their partisan-elected colleagues, but their opinions are of higher quality along several metrics. In terms of the information-theoretic approach, this is consistent with the merit system using better information about judge ability than partisan voting, without increasing bias.[25]

Evidence for the second prediction comes from the reforms that replace partisan elections with nonpartisan elections. While the results are not as strong as those for the partisan-to-merit reform, overall the evidence suggests that judges selected under nonpartisan elections are of higher quality than those selected under partisan elections. In terms of the information model, one can interpret this effect as a reduction in the bias component of candidate selection as party affiliation is no

[24] The baseline outcome of positive cites per opinion does not produce any revealing correlations, perhaps because they include many cites from lower-court judges in the same state. These cites reflect the importance of a binding precedent, rather than the quality of a persuasive precedent (which is captured better by the out-of-state citations metric).

[25] A separate possible mechanism is that a merit-system judgeship is a more attractive job than an election-system judgeship. After all, merit judges have stronger tenure and do not face as many political pressures. The reforms, therefore, could attract higher-quality candidates due to the higher perceived income from the position. While this mechanism would not be easy to distinguish empirically, a reason to discount it (relative to the information mechanism) is that merit candidates are nominated by an independent commission, which limits the scope for prospective judges to "campaign" for a position.

longer on the ballot. A related possible mechanism is that the decisions of these judges are less politically biased, which itself increases quality as measured by citations.

Meanwhile, moving from nonpartisan elections to merit selection has no effect on the relative performance of judges. In terms of the model, this null effect could be interpreted as two countervailing effects: any improvements in the quality of information due to the merit process are canceled out by increases in bias. After all, the partisan governor plays an important role in the merit process, so bias increases (along with information) when moving from nonpartisan elections to the merit system. It could also be that bias matters more than information in the selection of state supreme court judges.

Overall, the strongest evidence is for an increase in judge quality when moving from partisan elections to merit selection. In terms of the theoretical framework, this can be interpreted as the combination of two self-reinforcing effects. First, there is an increase in information, as merit commissions can observe judge candidate quality with more precision than voters. Second, there is a reduction in bias, as merit commissions put less importance on ideology than political primaries. The information-theoretic approach predicts that both effects contribute to higher average quality in the selected judges. However, the null effect of moving from nonpartisan elections to merit selection casts some doubt on the relative importance of the information mechanism and reinforces confidence in the relative importance of the political bias mechanism.

7 Conclusion

The goal of this paper has been to produce evidence regarding the hypothesis that the choice between a "politician" and "bureaucrat" entails a tradeoff between a sensitivity to the desires of the electorate and the execution of the mission to make high-quality legal decisions (Maskin and Tirole, 2004; Alesina and Tabellini, 2007, 2008). In our case, judges selected by a technocratic merit commission are of higher quality. These results are consistent with a selection model where better-informed experts can choose higher-quality officials than voters on average. Also, nonpartisan elections select better judges than partisan elections, which suggests that bias can interfere with voter selection for quality.

Our evidence is broadly in line with the early rational-choice approaches of Downs (1957) and Ferejohn (1986), in which voters use their information to make the best decisions they can, conditional upon their policy preferences. However, more information is not always better; more information on candidate quality can improve performance (see Pande, 2011), but more information on political affiliation can reduce performance.

Should all states immediately move to a merit system? Our evidence suggests that doing so would increase the work quality of appellate judges. But there are other criteria besides judicial citations for ranking courts, and ballot referenda for the merit plan have failed many times. There may be many other social impacts of these courts, but at present, we don't have data-driven tools to measure them. There is an ongoing debate on which is the superior system (e.g. Pozen, 2010);

the fact that states continue to experiment with different systems suggests that it is not clear which system is optimal. If a single system were clearly optimal, then we would have expected the market to have moved in that direction quickly, consistent with Posner's (1987) view that legal institutions move in the direction of efficient exchange.

Another caveat is that our findings are restricted to appellate court decision making. These are a highly selected group of judges, whose task is quite different from the work of most judges who adjudicate trials with a defendant who is present. One would need a different set of performance criteria for that case. Some relevant evidence in this regard, from Ash and Poyker (2019), is that elected trial-court judges respond more to tough-on-crime media in their sentencing decisions than appointed trial-court judges.

Still, given the importance of appellate judges – indicated, for example, by the large and growing literature using judicial decisions as a source of policy variation (e.g. Belloni et al., 2012) – these results have direct policy relevance. In developing countries, where courts play an important rule in establishing inclusive institutions (Djankov et al., 2003), policymakers could use our results as inputs in designing the selection system for judges and other officials (see Dal Bó and Finan, 2018). On the other hand, our evidence is from the latter decades of last century, and one should be cautious in extrapolating our results to the present.

The fact that we do find a pattern of effects predicted by our simple model helps explain why there is experimentation. While the results are consistent with merit commissions selecting better judges, judging is not a purely technical activity. The political views of judges color the ideological content of their decisions (see Epstein et al., 2013), which may explain why many jurisdictions prefer to give voters a clear signal of the political views of judges. Optimizing states would change systems only if it led to an improvement; hence at any point in time, there should be only small variation across states (as Choi et al. (2010) find).

A promising avenue for future research is to look at the retention effects (rather than selection effects) of electoral rules. As mentioned, while our empirical strategy holds retention incentives constant, those incentives are indeed changing with the reforms and one could look at within-court or within-judge effects on performance. For example, the evidence in Shepherd (2009b) suggests that retention incentives are an important factor in the decision-making of state supreme court judges. In addition, the selection-process could have an independent (perhaps behavioral) incentive effect, for example due to fairness considerations or reciprocal relationships with groups who supported a judge's candidacy (Dal Bó et al., 2010).

A broader issue is how to measure judge quality – one might argue that citations are not the relevant measure and that instead research should focus on how decisions affect social outcomes. The challenge is that there is neither an accepted measure of law quality nor any causal evidence on how law quality affects outcomes. While there is a large literature (too large to cite here) on how how rule changes in (for example) tort law, environmental law, or tax law affect economic performance, these results do not directly speak to the enforcement of rules by judges. What we

can say is that citations are correlated with bar evaluations, and that voters prefer judges with better evaluations (Lim and Snyder, 2015); and therefore that citations are correlated with voters' perception of performance. An open research question is whether one can design better judicial performance metrics that can be directly connected to economic performance.

The fact that voters – and even more so merit commissions – prefer more highly cited judges does not imply that one should implement a reward system based upon citations. Such rewards are likely to have unanticipated and harmful consequences. Hence, the best we can do is recommend more research on this important issue.

Regardless of how legal work quality is measured, it is noteworthy that a "judge" is not a single individual, but a team of individuals that includes clerks and secretarial staff. Judges select the clerks that are working for them. Hence our measures can be seen as composites that depend upon both the judge's legal skill when researching, reasoning, and writing, as well as managerial skill when selecting and directing clerks. As we know from Bloom et al. (2012), management quality varies across firms, and there are systematic relationships between management quality and firm performance.

Finally, our results highlight the fact that the American legal system is neither simple nor static. It is a complex, dynamic system consisting of many interlocking ingredients. Our study focuses upon one of the most important and influential ingredients of this system: U.S. state supreme court judges, who rule on all aspects of private law, including contract, tort, and property law. Our evidence is consistent with the hypothesis that these judges are professionals who are interested in enhancing the quality of the law. Hence, we have observed many states moving away from partisan political processes for selection toward nonpartisan and merit-based processes. These more "bureaucratic" systems have selected better judges and imposed incentives more aligned with the mission of legal quality.

References

A. Kirkland, P. and A. Coppock (2017, 06). Candidate choice without party labels:: New insights from conjoint survey experiments. *Political Behavior 40*, 1–21.

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

Alesina, A. and G. Tabellini (2007). Bureaucrats or politicians? part i: A single policy task. *The American Economic Review 97*(1), 169–179.

Alesina, A. and G. Tabellini (2008). Bureaucrats or politicians? part ii: Multiple policy tasks. *Journal of Public Economics 92*(3-4), 426–447.

Arellano, M. and S. Bond (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The review of economic studies 58*(2), 277–297.

Arnold, D., W. Dobbie, and C. S. Yang (2018). Racial bias in bail decisions. *The Quarterly Journal of Economics 133*(4), 1885–1932.

Ash, E., S. Asher, A. Bhowmick, D. L. Chen, T. Devi, C. Goessmann, P. Novosad, and B. Siddiqi (2021). Are indian judges biased? evidence from 8 million criminal court cases. *Center for Law & Economics Working Paper Series 2021*(03).

Ash, E. and W. B. MacLeod (2015). Intrinsic motivation in public service: Theory and evidence from state supreme courts. *Journal of Law and Economics 58*(4).

Ash, E. and M. Poyker (2019). Conservative news media and criminal justice: Evidence from exposure to fox news channel. *Columbia Business School Research Paper*.

Ashworth, S. (2012). Electoral accountability: Recent theoretical and empirical work. *Annual Review of Political Science 15*(1), 183–201.

Ashworth, S., E. Bueno de Mesquita, and A. Friedenberg (2017, May). Accountability and information in elections. *American Economic Journal: Microeconomics 9*(2), 95–138.

Ashworth, S. and E. B. de Mesquita (2008). Electoral selection, strategic challenger entry, and the incumbency advantage. *Journal of Politics 70*(4), 1006–1025.

Baker, S., A. Feibelman, and W. P. Marshall (2009). The continuing search for a meaningful model of judicial rankings and why it (unfortunately) matters. *Duke Law Journal 58*(7), 1645–1666.

Baker, S. and C. Mezzetti (2012). A theory of rational jurisprudence. *The Journal of Political Economy 120*(3), 513–551.

Banks, J. and R. Sundaram (1998, OCT). Optimal retention in agency problems. *Journal of Economic Theory 82*(2), 293–323.

Bannon, A., C. Lisk, and P. Hardin (2013). *Who Pays for Judicial Races?* Brennan Center for Justice at New York University School of Law.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica 80*(6), 2369–2429.

Berdejo, C. and N. Yuchtman (2013). Crime, punishment, and politics: an analysis of political cycles in criminal sentencing. *Review of Economics and Statistics 95*(3), 741–756.

Besley, T. and A. Case (1995). Does electoral accountability affect economic-policy choices - evidence from gubernatorial term limits. *Quarterly Journal of Economics 110*(3), 769–798.

Besley, T. and A. Case (2003). Political institutions and policy choices: evidence from the united states. *Journal of Economic Literature 41*(1), 7–73.

Besley, T. and S. Coate (2003). Elected versus appointed regulators: Theory and evidence. *Journal of the European Economic Association 1*(5), 1176–1206.

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

Besley, T., T. Persson, and D. M. Sturm (2010). Political competition, policy and growth: theory and evidence from the us. *The Review of Economic Studies 77*(4), 1329–1352.

Bloom, N., C. Genakos, R. Sadun, and J. V. Reenen (2012, February). Management practices across firms and countries. NBER Working Papers 17850, National Bureau of Economic Research, Inc.

Bloom, N. and J. V. Reenen (2007, November). Measuring and explaining management practices

across firms and countries. *The Quarterly Journal of Economics 122*(4), 1351–1408.

Bonneau, C. W. and D. M. Cann (2015, MAR). Party identification and vote choice in partisan and nonpartisan elections. *Political Behavior 37*(1), 43–66.

Canes-Wrone, B., T. S. Clark, and J. P. Kelly (2014, 2). Judicial selection and death penalty decisions. *American Political Science Review 108*, 23–39.

Canes-Wrone, B., M. Herron, and K. Shotts (2001, JUL). Leadership and pandering: A theory of executive policymaking. *American Journal of Political Science 45*(3), 532–550.

Canes-Wrone, B. and K. W. Shotts (2007, MAY). When do elections encourage ideological rigidity? *American Political Science Review 101*(2), 273–288.

Caselli, F. and M. Morelli (2004). Bad politicians. *Journal of Public Economics 88*(3-4), 759 – 782.

Chetty, R., N. Hendren, P. Kline, and E. Saez (2014). Where is the land of opportunity? the geography of intergenerational mobility in the united states. *The Quarterly Journal of Economics 129*(4), 1553–1623.

Choi, S. J., G. M. Gulati, and E. A. Posner (2008). Judicial evaluations and information forcing: Ranking state high courts and their judges. *Duke LJ 58*, 1313.

Choi, S. J., G. M. Gulati, and E. A. Posner (2010). Professionals or politicians: The uncertain empirical case for an elected rather than appointed judiciary. *Journal of Law, Economics, and Organization 26*(2), 290.

Choudhry, N. K., R. H. Fletcher, and S. B. Soumerai (2005). Systematic review: the relationship between clinical experience and quality of health care. *Annals of Internal medicine 142*(4), 260–273.

Christensen, R. K., J. Szmer, and J. M. Stritch (2012). Race and gender bias in three administrative contexts: Impact on work assignments in state supreme courts. *Journal of Public Administration Research and Theory*.

Condorcet, M. d. (1785). Essay on the application of analysis to the probability of majority decisions.

Dahl, M. W. and T. DeLeire (2008). *The association between children's earnings and fathers' lifetime earnings: estimates using administrative data*. University of Wisconsin-Madison, Institute for Research on Poverty.

Dal Bó, E. and F. Finan (2018). Progress and perspectives in the study of political selection. *Annual Review of Economics 10*, 541–575.

Dal Bó, E., F. Finan, O. Folke, T. Persson, and J. Rickne (2017). Who becomes a politician? *The Quarterly Journal of Economics 132*(4), 1877–1914.

Dal Bo, E., F. Finan, and M. A. Rossi (2013). Strengthening state capabilities: The role of financial incentives in the call to public service. *Quarterly Journal of Economics 128*(3), 1169–1218.

Dal Bó, P., A. Foster, and L. Putterman (2010). Institutions and behavior: Experimental evidence on the effects of democracy. *American Economic Review 100*(5), 2205–29.

DeGroot, M. H. (1972). *Optimal Statistical Decisions*. New York, NY: McGraw-Hill Book C.

Dewatripont, M., I. Jewitt, and J. Tirole (1999). The economics of career concerns, part ii: Application to missions and accountability of government agencies. *Review of Economic Studies 66*(1), pp.199–217.

Djankov, S., R. La Porta, F. Lopez-de Silanes, and A. Shleifer (2003). Courts. *The Quarterly Journal of Economics 118*(2), 453–517.

Dobbie, W., J. Cioldin, and C. S. Yang (2018, FEB). The effects of pretrial detention on conviction, future crime, and employment: Evidence from randomly assigned judges. *American Economic Review 108*(2), 201–240.

Downs, A. (1957). An economic theory of political action in a democracy. *Journal of Political Economy 65*(2), pp. 135–150.

Ellison, G. (2013, July). How does the market use citation data? the hirsch index in economics. *American Economic Journal: Applied Economics 5*(3), 63–90.

Epstein, L., W. M. Landes, and R. A. Posner (2013). *The Behavior of Federal Judges*. Harvard University Press.

Ferejohn, J. (1986). Incumbent performance and electoral control. *Public Choice 50*(1-3), 5–25.

Ferraz, C. and F. Finan (2008). Exposing corrupt politicians: The effects of brazil's publicly released audits on electoral outcomes. *Quarterly Journal of Economics 123*(2), 703–745.

Gennaioli, N. and A. Shleifer (2007). The evolution of common law. *The Journal of Political Economy 115*(1), 43–68.

Gordon, S. and G. Huber (2007). The effect of electoral competitiveness on incumbent behavior. *Quarterly Journal of Political Science 2*(2), 107–138.

Green, J. R. and N. L. Stokey (1983, June). A comparison of tournaments and contracts. *Journal of Political Economy 91*(3), 349–364.

Hall, M. (2007). Voting in state supreme court elections: Competition and context as democratic incentives. *Journal of Politics 69*(4), 1147–1159.

Hall, M. and C. Bonneau (2006, January). Does quality matter? challengers in state supreme court elections. *American Journal of Political Science 50*(1), 20–33.

Hanssen, F. (2004). Is there a politically optimal level of judicial independence? *The American Economic Review 94*(3), 712–729.

Hanssen, F. A. (1999). The effect of judicial institutions on uncertainty and the rate of litigation: The election versus appointment of state judges. *The Journal of Legal Studies 28*(1), pp.205–232.

Helland, E. and A. Tabarrok (2002). The effect of electoral institutions on tort awards. *American Law and Economics Review 4*(2), 341–370.

Huber, G. A. and S. C. Gordon (2004). Accountability and coercion: Is justice blind when it runs for office? *American Journal of Political Science 48*(2), 247–263. Times Cited: 79 Huber, GA Gordon, SC Huber, Gregory/A-5950-2012 Huber, Gregory/0000-0001-6804-8148 79.

Iaryczower, M., G. Lewis, and M. Shum (2013). To elect or to appoint? bias, information, and

responsiveness of bureaucrats and politicians. *Journal of Public Economics 97*, 230–244.

Kritzer, H. M. (2011). Competitiveness in state supreme court elections, 1946–2009. *Journal of Empirical Legal Studies 8*(2), 237–259.

Kritzer, H. M. (2015). *Justices on the Ballot: Continuity and Change in State Supreme Court Elections*. Cambridge University Press.

Landes, W. M. and R. A. Posner (1980). Legal change, judicial behavior, and the diversity jurisdiction. *The Journal of Legal Studies 9*(2), pp.367–386.

Lim, C. H. S. (2013). Preferences and incentives of appointed and elected public officials: Evidence from state trial court judges. *American Economics Review*.

Lim, C. H. S. and J. M. Snyder (2015, April). Is more information always better? party cues and candidate quality in u.s. judicial elections. *Journal of Public Economics*.

Lim, C. S. H., J. M. Snyder, and D. StrÃ¶mberg (2015). The judge, the politician, and the press: Newspaper coverage and criminal sentencing across electoral systems. *American Economic Journal: Applied Economics 7*(4), 103–135.

List, J. A. and D. M. Sturm (2006). Elections matter: Theory and evidence from environmental policy. *Quarterly Journal of Economics 121*(4), 1249–1281. Times Cited: 50 50.

Maskin, E. and J. Tirole (2004). The politician and the judge: Accountability in government. *The American Economic Review 94*(4), 1034–1054.

Miles, T. J. (2015). Do attorney surveys measure judicial performance or respondent ideology? evidence from online evaluations. *The Journal of Legal Studies 44*(S1), S231–S267.

Mookherjee, D. (1984). Optimal incentive schemes with many agents. *The Review of Economic Studies 51*(3), 433–446.

Nelson, M. J., R. P. Caufield, and A. D. Martin (2013). Oh, mi: A note on empirical examinations of judicial elections. *State Politics & Policy Quarterly*, 1532440013503838.

Niblett, A., R. A. Posner, and A. Shleifer (2010, June). The evolution of a legal rule. *The Journal of Legal Studies 39*(2), 325–358.

Norris, S. (2019). Examiner inconsistency: Evidence from refugee appeals. (2018-75).

Ohtani, K. (2000). Bootstrapping r2 and adjusted r2 in regression analysis. *Economic Modelling 17*(4), 473–483.

Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics 37*(2), 187–204.

Pande, R. (2011). Can informed voters enforce better governance? experiments in low-income democracies. *Annual Review of Economics 3*(1), 215–237.

Posner, R. (2008). *How Judges Think*. Harvard University Press.

Posner, R. A. (1987). The law and economics movement. *The American Economic Review 77*(2), pp.1–13.

Pozen, D. (2010). Judicial elections as popular constitutionalism. *Columbia Law Review 110*, 2047–2134.

Rahn, W. (1993, May). The role of partisan stereotypes in information-processing about political candidates. *American Journal of Political Science 37*(2), 472–496.

Savage, L. J. (1972 (first published 1954)). *The Foundations of Statistics*. New York, N.Y.: Dover Publications.

Shepherd, J. M. (2009a). Are appointed judges strategic too? *Duke Law Journal 58*(7), 1589–1626.

Shepherd, J. M. (2009b). The influence of retention politics on judges' voting. *The Journal of Legal Studies 38*(1), 169–206.

Stephenson, M. C. (2009). Legal realism for economists. *The Journal of Economic Perspectives 23*(2), pp.191–211.

A Model Appendix

The purpose of this appendix is to make precise the claims made in section 2 regarding the effects of changing the election system. The model supposes that there is a representative voter who has a noisy signal of judge performance. The voter prefers (is biased for) the candidate from their political party. In general, the exact level of bias is likely to be uncertain, and to vary from voter to voter. Hence, bias is modeled as the mean of the quality signal. It is then assumed that the voter chooses the judge with the highest net quality, conditional upon their information regarding party affiliation. In our empirical setting, this is indistinguishable from having utility for a particular party versus having incorrect beliefs.

A.1 Setup

Consider two judges, denoted by $j \in \{I, E\}$.[26] The judges have ability, indexed by $q_j \in \Re$, which is not directly observed by the voters. We assume that all judges come from the same underlying distribution of log skill:

$$q_j \sim N(0, 1). \tag{A.1}$$

Where the mean is normalized to zero and variance is normalized to one as we do not have data on variation in the pool of judges.

Politics creates a *bias* in voter beliefs, denoted by a number $b \in \Re$. Prior beliefs of the voters are given by:

$$q_I \sim N(b, 1),$$
$$q_E \sim N(-b, 1)$$

[26] We can think of I as the incumbent and E as the challenger, or simply two judges competing for the same position. We note that this is a simplification for all systems. The nonpartisan system sometimes has one judge running, or more than two. In the partisan system, there is a primary process which we abstract away from in the model. In the merit system, there is a multi-agent process, including a governor and a nominating commission, which we also abstract from.

which differ from the true prior (A.1) due to the bias. Under nonpartisan elections and merit selection, $b = 0$, while under partisan elections $b \neq 0$. When $b > 0$, the incumbent comes from the same party as the voter, while the entrant is from the other party. If $b < 0$ the roles are reversed. These priors are "biased" in the sense of representing a preference for a judge based upon their politics. As we know from Savage (1954), probability assessments are part of one's beliefs, and hence in a one-shot choice problem it may be impossible to distinguish between tastes and beliefs. Modeling bias in terms of beliefs allows us to have uncertainty in preferences combined with closed-form solutions for expected quality. It is assumed that all judges have the same ability, and hence $\mathbb{E}(q_j) = 0$.

The representative voter or governor chooses a judge based upon an *unbiased* noisy signal of judge ability given by:

$$s_j = q_j + \sigma_j \gamma_j, j \in \{I, E\}$$

where q_j is the realized quality of the judge, $\gamma_j \sim N(0,1)$ is drawn from a standard normal distribution, and σ_j^2 is the variance of the total error term $\sigma_j \gamma_j$. Let the precision of the error be defined by $\rho_j = \frac{1}{\sigma_j^2}$. We can allow for an incumbency information effect for judge I by letting the precision of the signal for the incumbent to be higher ($\rho_I \geq \rho_E$).

The representative voter observes the signals, updates beliefs, and then selects the judge with the highest expected value $\hat{q}_j = \mathbb{E}(q_j | s_j)$. In the data, we have an unbiased measure of the true quality of the judge, as measured in their work on judicial opinions. Our objective is to understand how the various parameters in the appointment system contribute to variation in work quality. For this purpose, let the exogenous parameters of the model be $\omega = \{b, \rho_I, \rho_E\} \in \Omega = \Re \times \Re_{++}^2$, the bias of the voter and the precision of the signals for each judge.

A.2 Analysis

The judicial reforms in our data set represent natural experiments that shift these parameters. A change from partisan to nonpartisan elections corresponds to a change in the bias from $b \neq 0$ to zero. A change from elections to merit selection corresponds to an increase in the precision of signals, ρ_I and ρ_E. After solving the model, we revisit the associated predictions for how the reforms affect work quality.

In a linear model with normally distributed errors, Bayes' Rule implies that after observing the quality signal, the optimal belief-updating rule is a linear function of the prior and signal (DeGroot, 1972), giving posteriors:

$$q_I(s_I) \equiv \frac{b + \rho_I s_I}{1 + \rho_I},$$
$$q_E(s_E) \equiv \frac{-b + \rho_E s_E}{1 + \rho_E}.$$

The existence of bias in beliefs leads the voter to increase her assessment of the judge she believes

is better *ex-ante.* Notice that if the quality of information ρ_j is sufficiently large, the effect of bias becomes negligible.

The optimal decision rule is to choose the judge with the highest expected ability:

$$q\left(s_I, s_E\right) = \max\left\{q_I\left(s_I\right), q_E\left(s_E\right)\right\}. \tag{A.2}$$

We want to compute the expected quality of the judge, given the true prior (A.1) and the parameters ω:

$$q^*\left(\omega\right) = \mathbb{E}\left\{q\left(s_I, s_E\right)|\omega\right\} \tag{A.3}$$

and use this expectation to analyze variation across appointment systems.

To work out the expected quality, we need to take expectations over the true distribution of possible qualities. We begin by observing that the choice of judge depends upon the sign of the difference in posteriors:

$$\begin{aligned} t\left(s_I, s_E, \omega\right) &= q_I\left(s_I\right) - q_E\left(s_E\right), \\ &= b\left(\frac{1}{1+\rho_I} + \frac{1}{1+\rho_E}\right) \\ &+ \left(q_I + \sigma_I\gamma_I\right)\frac{\rho_I}{1+\rho_I} \\ &- \left(q_E + \sigma_E\gamma_E\right)\frac{\rho_E}{1+\rho_E}. \end{aligned} \tag{A.4}$$

When the $t\left(\right)$ statistic is positive, judge I is chosen; judge E is chosen otherwise. As the statistic $t\left(\right)$ is a linear combination of normally distributed, independent random variables it is itself normally distributed. It is also an informative signal of judge quality. Hence we can use Bayes' Rule to compute $\mathbb{E}\left(q_j|t,\omega\right)$, the expected value of judge quality conditional upon t.

Let $h\left(t,\omega\right)$ denote the unconditional distribution of $t\left(\right)$. The expected quality of an elected judge is given by

$$q^*\left(\omega\right) = \int_{-\infty}^{0} \mathbb{E}\left\{q_E|t,\omega\right\} h\left(t,\omega\right) dt + \int_{0}^{\infty} \mathbb{E}\left(q_I|t,\omega\right) h\left(t,\omega\right) dt. \tag{A.5}$$

This expression can be explicitly computed using Bayes' Rule. The solution to (A.5) is given in the next proposition:

Proposition 1. *Given the parameters $\omega \in \Omega$, the expected quality of the selected judge is given by:*

$$q^*\left(\omega\right) = A\left(\omega\right)\phi\left(B\left(\omega\right)\right),$$

where $\phi(x) = \int_x^\infty x f(x)\, dx$, $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ *is the normal pdf, and*

$$A(\omega) = \sigma_t \left(\frac{\rho_{SI}}{(1+\rho_{SI})} \left(\frac{1+\rho_I}{\rho_I} \right) + \frac{\rho_{SE}}{(1+\rho_{SE})} \left(\frac{1+\rho_E}{\rho_E} \right) \right),$$
$$B(\omega) = -\frac{b}{\sigma_t} \left(\frac{1}{1+\rho_I} + \frac{1}{1+\rho_E} \right),$$

where

$$\sigma_t^2 = \frac{\rho_I}{(1+\rho_I)} + \frac{\rho_E}{(1+\rho_E)}$$

is the variance of t *in (A.4). The precision of the judge performance signals for judges* I *and* E *based upon* $t = q_I - q_E$ *are:*

$$\rho_{SI} = \left(\frac{\rho_I}{1 + \frac{\rho_E (1+\rho_I)^2}{\rho_I (1+\rho_E)}} \right),$$
$$\rho_{SE} = \left(\frac{\rho_E}{1 + \frac{\rho_I (1+\rho_E)^2}{\rho_E (1+\rho_I)}} \right).$$

Proof. We begin by observing that the signal t is the sum of independent normally distributed random variables, and hence by (A.4) we get:

$$\mathbb{E}\{t\} = b \left(\frac{1}{1+\rho_I} + \frac{1}{1+\rho_E} \right), \tag{A.6}$$
$$\begin{aligned} \text{var}(t) &= \left(1 + \frac{1}{\rho_I}\right) \left(\frac{\rho_I}{1+\rho_I} \right)^2 \\ &\quad + \left(1 + \frac{1}{\rho_E}\right) \left(\frac{\rho_E}{1+\rho_E} \right)^2 \\ &= \frac{\rho_I}{(1+\rho_I)} + \frac{\rho_E}{(1+\rho_E)} \end{aligned} \tag{A.7}$$

Next we need to work out $\mathbb{E}\{q_j | t, \omega\}$. We can write:

$$\begin{aligned} SI &= \left(\frac{1+\rho_I}{\rho_I} \right) \left(t - b \times \left(\frac{1}{1+\rho_I} + \frac{1}{1+\rho_E} \right) \right), \\ &= q_I + \left(\frac{1}{\sqrt{\rho_I}} \gamma_I - \frac{\rho_E (1+\rho_I)}{\rho_I (1+\rho_E)} \left(q_E + \frac{1}{\sqrt{\rho_E}} \gamma_E \right) \right). \end{aligned}$$

This is the signal equation for judge I based upon t. Notice that the last three random variables

on the right are *ex ante* i.i.d. $N(0,1)$. Hence S_I is an unbiased signal of q_I with precision:

$$\begin{aligned}
\rho_{SI} &= \left(\frac{1}{\rho_I} + \frac{\left(\rho_E^2 + \rho_E\right)\left(1+\rho_I\right)^2}{\rho_I^2\left(1+\rho_E\right)^2}\right)^{-1} \\
&= \left(\frac{1}{\rho_I} + \frac{\rho_E\left(1+\rho_I\right)^2}{\rho_I^2\left(1+\rho_E\right)}\right)^{-1} \\
&= \left(\frac{\rho_I\left(1+\rho_E\right) + \rho_E\left(1+\rho_I\right)^2}{\rho_I^2\left(1+\rho_E\right)}\right)^{-1} \\
&= \left(\frac{\rho_I^2\left(1+\rho_E\right)}{\rho_I\left(1+\rho_E\right) + \rho_E\left(1+\rho_I\right)^2}\right). \\
&= \left(\frac{\rho_I}{1 + \frac{\rho_E\left(1+\rho_I\right)^2}{\rho_I\left(1+\rho_E\right)}}\right).
\end{aligned}$$

Observe that $\lim_{\rho_I \to 0} \rho_{SI} = 0$ and $\lim_{\rho_I \to \infty} \rho_{SI} = \left(1 + \frac{1}{\rho_E}\right)$.

From this we get:

$$\begin{aligned}
\mathbb{E}\left\{q_I | t, \omega\right\} &= \frac{\rho_{SI} SI}{1+\rho_{SI}} \\
&= \frac{\rho_{SI}\left(\frac{1+\rho_I}{\rho_I}\right)\left(t - b \times \left(\frac{1}{1+\rho_I} + \frac{1}{1+\rho_E}\right)\right)}{1+\rho_{SI}}
\end{aligned}$$

Now, we can do a similar calculation for judge E who enters with the opposite sign. Let:

$$\begin{aligned}
SE &= \left(\frac{1+\rho_E}{\rho_E}\right)\left(-t + b \times \left(\frac{1}{1+\rho_I} + \frac{1}{1+\rho_E}\right)\right), \\
&= q_E + \left(\frac{1}{\sqrt{\rho_E}}\gamma_E - \frac{\rho_I\left(1+\rho_E\right)}{\rho_E\left(1+\rho_I\right)}\left(q_I + \frac{1}{\sqrt{\rho_I}}\gamma_I\right)\right).
\end{aligned}$$

Then we have:

$$\rho_{SE} = \left(\frac{\rho_E^2\left(1+\rho_I\right)}{\rho_E\left(1+\rho_I\right) + \rho_I\left(1+\rho_E\right)^2}\right),$$

and

$$\begin{aligned}
\mathbb{E}\left\{q_E | t, \omega\right\} &= \frac{\rho_{SE} SE}{1+\rho_{SI}} \\
&= -\frac{\rho_{SE}\left(\frac{1+\rho_E}{\rho_E}\right)\left(t - b \times \left(\frac{1}{1+\rho_I} + \frac{1}{1+\rho_E}\right)\right)}{1+\rho_{SE}}
\end{aligned}$$

Let us define a change of variable:

$$v = \sqrt{\rho_t}\left(t - b \times \left(\frac{1}{1+\rho_I} + \frac{1}{1+\rho_E}\right)\right),$$

where $\rho_t = \frac{1}{\text{var}(t)}$ as defined above. Thus $v \sim N(0,1)$. Moreover, judge I is chosen whenever $t \geq 0$ or $v \geq -\sqrt{\rho_t} b \left(\frac{1}{1+\rho_I} + \frac{1}{1+\rho_E}\right)$. Conversely, judge E is chosen when $t < 0$. Using this change in variables we have:

$$\mathbb{E}\{q_I | v, \omega\} = \frac{\rho_{SI}\left(\frac{1+\rho_I}{\rho_I}\right)}{\sqrt{\rho_t}\,(1+\rho_{SI})} v,$$
$$\mathbb{E}\{q_E | v, \omega\} = -\frac{\rho_{SE}\left(\frac{1+\rho_E}{\rho_E}\right)}{\sqrt{\rho_t}\,(1+\rho_{SE})} v.$$

From these expressions and (A.5) we have:

$$q^*(\omega) = -\frac{\rho_{SE}\left(\frac{1+\rho_E}{\rho_E}\right)}{\sqrt{\rho_t}\,(1+\rho_{SE})} \int_{-\infty}^{-b\sqrt{\rho_t}\left(\frac{1}{1+\rho_I}+\frac{1}{1+\rho_E}\right)} v f(v)\, dv \tag{A.8}$$
$$+ \frac{\rho_{SI}\left(\frac{1+\rho_I}{\rho_I}\right)}{\sqrt{\rho_t}\,(1+\rho_{SI})} \int_{-b\sqrt{\rho_t}\left(\frac{1}{1+\rho_I}+\frac{1}{1+\rho_E}\right)}^{\infty} v f(v)\, dv.$$

Observe that since $v \sim N(0,1)$ we have

$$\int_{-\infty}^{-b\sqrt{\rho_t}\left(\frac{1}{1+\rho_I}+\frac{1}{1+\rho_E}\right)} v f(v)\, dv + \int_{-b\sqrt{\rho_t}\left(\frac{1}{1+\rho_I}+\frac{1}{1+\rho_E}\right)}^{\infty} v f(v)\, dv = 0.$$

Using this fact in (A.8) implies $B(\omega)$ and $A(\omega)$ in the statement of the proposition. □

While Proposition (1) illustrates the relative complexity of the voter's decision, we can derive some generic features of the optimal choice. First, the bias b affects quality through $\phi(B(\omega))$. The function $\phi(x)$ (the expected value of the truncated normal) is positive, takes its maximum at $x = 0$, and has $\lim_{x\to\pm\infty} \phi(x) = 0$. Therefore, holding the precision of information fixed, an increase in the magnitude of bias away from zero decreases the $\phi(B(\omega))$ term, reducing the expected quality of the selected official.

Notice that bias is not affecting the quality of the signal in this model; it is only affecting the *interpretation of the signal.* In other words, suppose we have two voters who have the same-magnitude bias but of opposite signs – one is a Democrat, while the other is a Republican. Further suppose that $\rho_E = \rho_I$, and that the incumbent is a Republican and the entrant is a Democrat. The voters receive the same signal. Under the assumption that the pool of candidates from both parties have the same quality distribution, then Proposition 1 implies that the expected ability of

the elected official is the same regardless of the political affiliation of the representative voter. Thus we have:

Corollary 2. *If the distribution of judge quality is the same for both parties, then an increase in bias reduces the expected quality of selected judges.*

Thus, in a state where the pool of judges from each party is similar, we can expect that moving from a partisan selection system to a nonpartisan system results in an increase in quality.

Next consider changes in the information, ρ_I, ρ_E. We can decompose the effects of information into two parts. First, there is a direct effect of information due to selecting a better candidate more frequently. This direct effect of information works through the $A(\cdot)$ term in Proposition (1), where:

$$\lim_{\rho_I,\rho_E\to 0} A(\omega) = 0$$

with $A(\omega) > 0$ for non-zero information. Thus we have:

Corollary 3. *Holding fixed the level of bias and the pool of prospective judges, then increasing the quality of information regarding judge ability increases the expected quality of the selected judge.*

A goal of a merit committee to select judges is to increase the quality of the information available for each candidate. This result suggests that merit system should enhance the quality of selected judges relative to a partisan system.

A second channel for the effect of information on judge quality is through its interaction with the political bias. Information can influence quality by mitigating the effect of bias. Observe:

$$\lim_{\rho_I,\rho_E\to\infty} B(\omega) = 0,$$

and hence there is no bias when parties are well-informed. Suppose that one has perfect information regarding the incumbent ($\rho_I \to \infty$), then :

$$A\left(b, \sigma_I^2 = 0, \rho_E\right) = \left(\frac{1+\rho_E}{1+2\rho_E}\right)^{1/2}\left(1+\frac{\rho_E}{(1+\rho_E)}\right),$$
$$B\left(b, \sigma_I^2 = 0, \rho_E\right) = -b\frac{1}{(1+2\rho_E)^{1/2}(1+\rho_E)^{1/2}}.$$

Note that even with perfect information regarding one candidate, if there is uncertainty regarding the other candidate, then any bias in favor of a candidate increases the likelihood of that person winning. As the precision of information regarding the entrant increases, this bias decreases. Since $A\left(b, \sigma_I^2 = 0, \rho_E\right)$ is increasing with ρ_E then we can see that increasing information regarding the entrant increases expected quality, q^*, even when the quality of the incumbent is perfectly observed. We have

$$\lim_{\rho_I,\rho_E\to\infty} q^*(\omega) = 2\sqrt{2}\phi(0)$$

in the limit.

To see this, observe that $0 = \arg\max_{x \in \Re} \phi\left(\mathrm{x}\right)$ and $\lim_{x \to \pm\infty} = 0$. Hence no bias, $b = 0$, provides the highest $q^{*}\left(\omega\right)$.

Next consider the consequences of no information, in which case we let $\rho_I, \rho_E \to 0$.

$$\begin{aligned}\lim_{\rho_I, \rho_E \to 0} \sigma_t^2 &= \lim_{\rho_I, \rho_E \to 0} \frac{\rho_I}{\left(1+\rho_I\right)} + \frac{\rho_E}{\left(1+\rho_E\right)} \\ &= 0.\end{aligned}$$

Since $A\left(\omega\right)$ includes the term σ_t, we have $\lim_{\rho_I, \rho_E \to 0} A\left(\omega\right) = 0$, and hence $\lim_{\rho_I, \rho_E \to 0} q^{*}\left(\omega\right) = 0$.

Next consider what happens when parties perfectly observe the quality of the incumbent. $\rho_I \to \infty$. In that case we have:

$$\begin{aligned}\lim_{\rho_I \to \infty} \sigma_t^2 &= \lim_{\rho_I \to \infty} \frac{\rho_I}{\left(1+\rho_I\right)} + \frac{\rho_E}{\left(1+\rho_E\right)} \\ &= 1 + \frac{\rho_E}{\left(1+\rho_E\right)} \\ &= \frac{1+2\rho_E}{1+\rho_E}\end{aligned}$$

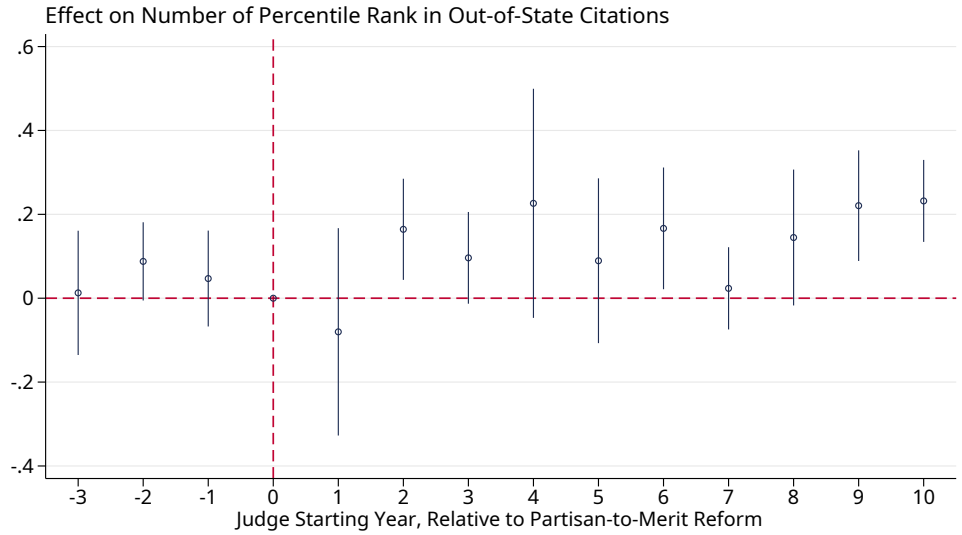
$$\begin{aligned}\lim_{\rho_I \to \infty} \rho_{SI} &= \lim_{\rho_I \to \infty} \left(\frac{\rho_I^2 \left(1+\rho_E\right)}{\rho_I \left(1+\rho_E\right) + \rho_E \left(1+\rho_I\right)^2} \right) \\ &= \left(\frac{1+\rho_E}{\rho_E} \right)\end{aligned}$$

$$\begin{aligned}\lim_{\rho_I \to \infty} \rho_{SE} &= \lim_{\rho_I \to \infty} \left(\frac{\rho_E^2 \left(1+\rho_I\right)}{\rho_E \left(1+\rho_I\right) + \rho_I \left(1+\rho_E\right)^2} \right) \\ &= \frac{\rho_E^2}{\rho_E + \left(1+\rho_E\right)^2}\end{aligned}$$

From these expressions we get:

$$\begin{aligned}\lim_{\rho_I \to \infty} A\left(\omega\right) &= \lim_{\rho_I \to \infty} \sigma_t \left(\frac{\rho_{SI}}{\left(1+\rho_{SI}\right)} \left(\frac{1+\rho_I}{\rho_I} \right) + \frac{\rho_{SE}}{\left(1+\rho_{SE}\right)} \left(\frac{1+\rho_E}{\rho_E} \right) \right), \\ &= \left(\frac{1+2\rho_E}{1+\rho_E} \right)^{1/2} \left(\left(\frac{1+\rho_E}{1+2\rho_E} \right) + \frac{\rho_E^2}{\left(1+\rho_E\right)\left(1+2\rho_E\right)} \left(\frac{1+\rho_E}{\rho_E} \right) \right) \\ &= \left(\frac{1+\rho_E}{1+2\rho_E} \right)^{-1/2} \left(\frac{1+\rho_E}{1+2\rho_E} \right) \left(1 + \frac{\rho_E}{\left(1+\rho_E\right)} \right) \\ &= \left(\frac{1+\rho_E}{1+2\rho_E} \right)^{1/2} \left(1 + \frac{\rho_E}{\left(1+\rho_E\right)} \right)\end{aligned}$$

Figure A.1: Event Study Visualization for Partisan-to-Merit Reform



Notes. Coefficient plot for the event study specification (5.1). Coefficients are effects of judge starting year relative to the year of the partisan-to-merit selection reform. Error spikes give 95% confidence intervals. The pre-reform coefficients are not jointly significant ($p = .29$), while the post-reform coefficients are highly jointly significant ($p < .0001$). Includes court-year fixed effects, starting year fixed effects, and court-specific starting-year trends. Standard errors clustered by state and year.

and

$$\begin{aligned}\lim_{\rho_I \to \infty} B(\omega) &= \lim_{\rho_I \to \infty} -b\sigma_t \left(\frac{1}{1+\rho_I} + \frac{1}{1+\rho_E} \right), \\ &= -b \left(\frac{1+2\rho_E}{1+\rho_E} \right)^{1/2} \left(\frac{1}{1+\rho_E} \right) \\ &= -b \frac{(1+2\rho_E)^{1/2}}{(1+\rho_E)^{3/2}}\end{aligned}$$

B Additional Empirical Results

This section includes a number of table and figures with additional results and robustness checks, discussed throughout the main text.

Table A.1: No Evidence of Selective Attrition due to Electoral Reforms

	(1)	(2)
	Judge Career Length (Years)	
Pre-Reform Judge Quality	-1.468	0.193
	(3.792)	(0.273)
Nonpartisan to Merit	9.003*	10.67
	(3.197)	(6.611)
Nonpartisan to Merit ×	-4.693	-0.337
Pre-Reform Judge Quality	(6.720)	(0.577)
Partisan to Merit	19.18**	19.51**
	(5.369)	(5.726)
Partisan to Merit ×	-4.325	-0.375
Pre-Reform Judge Quality	(7.266)	(0.538)
Partisan to Nonpartisan	7.025**	10.03**
	(2.395)	(2.685)
Partisan to Nonpartisan ×	-0.464	-0.383
Pre-Reform Judge Quality	(4.125)	(0.272)
Quality Measure	Ranks	Logs
N	175	175
adj. R^2	0.475	0.468
Court FE's	X	X
Cohort FE's	X	X

This table reports estimates for a regression at the judge level, where the outcome is career length (number of years on the state supreme court. The explanatory variables are indicator variables for judges that leave the court after the election reforms. These variables are interacted with average judge quality (positive citations per opinion) from years before the reforms (in Column 1, that is measured in rank percentiles, while in Column 2, that is measured in logs). Court and judge cohort fixed effects are absorbed. The reform indicators show that judges that last until after a reform tend to stay longer, reflecting the stronger tenure. The interaction terms show that this career extension effect does not vary according to pre-reform judge quality. Standard errors clustered by state in parentheses. $^{+}$ p<.1, * p<0.05, ** p<0.01.

Table A.2: Effects on Alternative Work Output Measures

	(1)	(2)	(3)	(4)
	Total Words Written		*Other Output Metrics (Rank)*	
	in Levels	*in Logs*	*Characters*	*Sentences*
Partisan	4179.9	-0.00648	0.0408	0.0487
to Merit	(3196.3)	(0.0500)	(0.0393)	(0.0416)
Partisan to	-426.0	-0.0162	0.0732	0.0576
Nonpartisan	(1274.1)	(0.0232)	(0.0503)	(0.0640)
Nonpartisan	-2831.7	-0.0226	-0.0614	-0.0246
to Merit	(8478.9)	(0.136)	(0.130)	(0.126)
N	14894	14894	14894	14894
R^2	0.575	0.559	0.024	0.025
Court-Year FE's	X	X	X	X
Cohort FE / Trends	X	X	X	X

Notes. Estimates of ρ from Equation (4.1): the average difference between judges selected under a new system, relative to to judges selected under the old system – e.g., "Partisan to Merit" means the merit-judge outcome relative to the partisan-judge outcome. The first three columns have positive citations per case as the outcome: "in Levels" means the regression uses the unadjusted judge-year level outcome; " in logs" means the outcome is transformed by the log of 1 plus the unadjusted value/ For Columns 3 and 4, outcomes are rank percentiles, where the lowest-value in a court-year group gets a zero, the highest-value gets a one, and other judges are distributed uniformly in between according to rank. Characters is the total number of characters (letters) written. Sentences is the total number of sentences written. Court-Year FE refers to court-year interacted fixed effects. Cohort FE's / Trends includes fixed effects for decade that a judge started on a court and judge birth, plus state-specific trends in judge starting cohort. Standard errors, adjusted for two-way clustering by state and year, in parentheses. +p<.1, * p<0.05, ** p<0.01.

Table A.3: Effects on Alternative Work Quality Measures (Experience FE)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Positive Cites Per Case			*Cites Per Case (Rank)*			
	in Levels	*in Logs*	*Ranks (10 yr)*	*Pos+Neg*	*Discuss*	*Quote*	*Out-State*
Partisan	0.751	0.0532	0.0568	0.0803*	0.0863**	0.116**	0.0758*
to Merit	(0.510)	(0.0371)	(0.0451)	(0.0350)	(0.0265)	(0.0321)	(0.0304)
Partisan to	0.464	0.0758	0.0508	0.0467	0.0973*	0.0818	0.0637*
Nonpartisan	(1.082)	(0.0471)	(0.0439)	(0.0548)	(0.0377)	(0.0670)	(0.0289)
Nonpartisan	-0.873	-0.0657	-0.125	-0.101	-0.0956	-0.0581	-0.0345
to Merit	(1.917)	(0.103)	(0.116)	(0.143)	(0.116)	(0.105)	(0.0755)
N	14890	14890	14890	14890	14890	14890	14890
R^2	0.619	0.808	0.049	0.051	0.043	0.045	0.032
Court-Year FE's	X	X	X	X	X	X	X
Cohort FE / Trends	X	X	X	X	X	X	X
Experience FE	X	X	X	X	X	X	X

Notes. Estimates of ρ from Equation (4.1): the average difference between judges selected under a new system, relative to to judges selected under the old system – e.g., "Partisan to Merit" means the merit-judge outcome relative to partisan-judge outcome. The first three columns have positive citations per case as the outcome: "in Levels" means the regression uses the unadjusted judge-year level outcome; " in logs" means the outcome is transformed by the log of 1 plus the unadjusted value; "Rank (10 yr)" refers to the percentile outcome, but only including citations within the next ten years after a case is published. "Pos + Neg" includes all citations (not just positive, but also negative and distinguishing). "Discuss" only includes cites where the case is discussed and applied. "Quote" only includes cites where the case's language is directly quoted. "Out-State" includes only citations from courts in other states. For Columns 3 through 7, outcomes are rank percentiles, where the lowest-value in a court-year group gets a zero, the highest-value gets a one, and other judges are distributed uniformly in between according to rank. Court-Year FE refers to court-year interacted fixed effects. Cohort FE's / Trends includes fixed effects for decade that a judge started on a court and judge birth, plus state-specific trends in judge starting cohort. Experience FE means fixed effects for years of tenure on the court. Standard errors, adjusted for two-way clustering by state and year, in parentheses. +p<.1, * p<0.05, ** p<0.01.

Table A.4: Additional Robustness Checks for Main Results

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
					Effect on Work Quality					
	Drop First/Last	*More Case Controls*	*Caseload Weights*	*Career Weights*	*Effect Window*	*State-Year Cluster*	*Bio Controls*	*Expend Controls*	*Elect-Cycle Controls*	*Cohort × Experience*
Partisan	0.0879*	0.0980*	0.122**	0.136**	0.136+	0.0982**	0.104*	0.0916*	0.101*	0.0792+
to Merit	(0.0402)	(0.0422)	(0.0426)	(0.0472)	(0.0780)	(0.0256)	(0.0398)	(0.0405)	(0.0409)	(0.0459)
Partisan to	0.130*	0.0964+	0.0457	0.0316	0.140+	0.0956	0.105*	0.092	0.0827	0.0926+
Nonpartisan	(0.0640)	(0.0491)	(0.0499)	(0.0276)	(0.0762)	(0.0650)	(0.0518)	(0.0550)	(0.0680)	(0.0531)
Nonpartisan	-0.0706	-0.0418	-0.0707	-0.130	-0.121	-0.0887+	-0.0962	-0.0872	-0.0976	-0.123
to Merit	(0.156)	(0.146)	(0.165)	(0.132)	(0.182)	(0.0535)	(0.152)	(0.150)	(0.153)	(0.157)
N	13169	14894	14894	14881	2604	14894	14894	14894	14894	14894
R^2	0.064	0.184	0.080	0.126	0.119	0.046	0.049	0.046	0.05	0.049
Court-Year FE's	X	X	X	X	X	X	X	X	X	X
Cohort FE / Trends	X	X	X	X	X	X	X	X	X	X

Notes. Estimates of ρ from Equation (4.1): the average difference between judges selected under a new system, relative to to judges selected under the old system – e.g., "Partisan to Merit" means the merit-judge outcome relative to the partisan-judge outcome. The outcomes is rank percentile in positive citations per opinion, where the lowest-value in a court-year group gets a zero, the highest-value gets a one, and other judges are distributed uniformly in between according to rank. Court-Year FE refers to court-year interacted fixed effects. Cohort FE's / Trends includes fixed effects for decade that a judge started on a court and judge birth, plus state-specific trends in judge starting cohort. Column 1 drops the first and last year of each judge's career; Column 2 includes 10 principal components of the vector of case characteristics; Column 3 weights the regressions by a judge's number of cases; Column 4 weights the regressions by the inverse of the number of years that a judge is in the dataset; Column 5 limits the regression to 12 years before and after the rule change; Column 6 is the main specification with clustering by state-year; Column 7 includes judge biographical controls (and associated missing dummies): starting age, partisan, female, top school, previous judge, previous private practice, and previous academia experience. Column 8 includes a control for log government expenditures on the judicial branch during a judge's starting year. Column 9 includes a fixed effect for the years until the next election process for a judge. Column 10 includes judge-cohort-specific (starting decade specific) trends in experience (years-on-the court). Standard errors, adjusted for two-way clustering by state and year, in parentheses. +p<.1, * p<0.05, ** p<0.01.

Table A.5: Robustness to Dropping Individual Treated States

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Effect on Work Quality									
Drop State	*GA*	*KY*	*UT*	*CO*	*IA*	*IN*	*KS*	*NE*	*OK*	*TN*
Partisan	0.0984*	0.0947*	0.0985*	0.103*	0.0822+	0.0767+	0.132**	0.0943+	0.0963+	0.103*
to Merit	(0.0403)	(0.0399)	(0.0402)	(0.0458)	(0.0450)	(0.0384)	(0.0328)	(0.0494)	(0.0502)	(0.0464)
Partisan to	0.127	0.0448	0.128*	0.0971+	0.0987+	0.0939+	0.0989+	0.0954+	0.0954+	0.0941+
Nonpartisan	(0.0819)	(0.0334)	(0.0479)	(0.0535)	(0.0524)	(0.0524)	(0.0530)	(0.0547)	(0.0529)	(0.0537)
Nonpartisan	-0.0912	-0.0869	-0.0897	-0.0899	-0.0869	-0.0871	-0.0846	-0.0857	-0.0892	-0.0869
to Merit	(0.154)	(0.153)	(0.154)	(0.154)	(0.153)	(0.154)	(0.153)	(0.154)	(0.154)	(0.154)
N	14548	14633	14654	14552	14455	14651	14561	14555	14362	14651
R^2	0.045	0.046	0.046	0.042	0.046	0.045	0.047	0.045	0.046	0.047
Court-Yr FE	X	X	X	X	X	X	X	X	X	X
Cohort FE	X	X	X	X	X	X	X	X	X	X

Notes. Estimates of ρ from Equation (4.1): the average difference between judges selected under a new system, relative to to judges selected under the old system – e.g., "Partisan to Merit" means the merit-judge outcome relative to the partisan-judge outcome. The outcomes is rank percentile in positive citations per opinion, where the lowest-value in a court-year group gets a zero, the highest-value gets a one, and other judges are distributed uniformly in between according to rank. Court-Year FE refers to court-year interacted fixed effects. Cohort FE's / Trends includes fixed effects for decade that a judge started on a court and judge birth, plus state-specific trends in judge starting cohort. Columns drop an individual state (selected from the set of partisan-to-merit and partisan-to-nonpartisan states), indicated in the Drop State row. Standard errors, adjusted for two-way clustering by state and year, in parentheses. +p<.1, * p<0.05, ** p<0.01.

Table A.6: Effects on other Dimensions of Judge Behavior

	(1) *Affirm Rate*	(2) *Case Length*	(3) *Research*	(4) *Dissent*	(5) *Total Cites*	(6) *Overruled*	(7) *Superseded*
Partisan	0.0202	0.0665	0.0841	-0.0321	0.0399	0.0419	0.0174
to Merit	(0.0469)	(0.0696)	(0.0801)	(0.0492)	(0.0334)	(0.0389)	(0.0361)
Partisan to	0.0474	-0.0555	0.0156	0.0849	0.120**	-0.0233**	0.0632
Nonpartisan	(0.126)	(0.0411)	(0.0745)	(0.0726)	(0.0143)	(0.00793)	(0.0562)
Nonpartisan	-0.0214	-0.104	-0.106	-0.0536	-0.0150	-0.0412	-0.0701*
to Merit	(0.0533)	(0.129)	(0.158)	(0.0890)	(0.118)	(0.0437)	(0.0337)
N	14894	14894	14894	14894	14894	14894	14894
R^2	0.020	0.034	0.036	0.078	0.031	0.421	0.282
Court-Year FE's	X	X	X	X	X	X	X
Cohort FE's / Trends	X	X	X	X	X	X	X

Notes. Estimates of ρ from Equation (4.1): the average difference between judges selected under a new system, relative to to judges selected under the old system – e.g., "Partisan to Merit" means the merit-judge outcome relative to the partisan-judge outcome. Outcomes are rank percentiles, where the lowest-value in a court-year group gets a zero, the highest-value gets a one, and other judges are distributed uniformly in between according to rank. Affirm Rate is the rate at which a judge affirms (rather than reverses) a lower-court decision. Case Length is the average number of words per published opinion. Research is the average number of previous cases cited in each opinion. Dissent is the number of dissents published. Total Cites is the total number of positive citations to a judge in a year (not per opinion, as in the main text). Overruled Rate is the rate that a judge is over-ruled, either by a future state supreme court case, or a U.S. Supreme Court case. Superseded Rate is the rate that a judge decision is reversed by state legislation. Court-Year FE refers to court-year interacted fixed effects. Cohort FE's / Trends includes fixed effects for decade that a judge started on a court and judge birth, plus state-specific trends in judge starting cohort. Standard errors, adjusted for two-way clustering by state and year, in parentheses. +p<.1, * p<0.05, ** p<0.01.

C [Online] Notes on Institutional Reforms

Table 1 provides the list of just selection systems. This data were collected from a range of primary and secondary sources. This appendix includes some further notes on the data and the institutional reforms, as well as further regression specifications.

First, some notes on how the systems were categorized. Besides the three systems mentioned, there are governor appointment and legislative appointment. There is a fourth hybrid system, where judges are initially selected through partisan elections but thereafter face uncontested retention elections. California has governor appointment but uncontested retention elections. The other states either have some combination of governor or legislative appointment, both for initial selection and for period retention. In Massachusetts, New Hampshire, New Jersey, and Rhode Island, judges have lifelong tenure. In Ohio and Michigan, judicial elections are difficult to classify within the partisan/nonpartisan dichotomy because they have partisan primaries and nomination processes, but the political party is not on the ballot in general elections. Following Nelson et al. (2013), we classify these states as partisan selection and nonpartisan retention. Alternative codings, or leaving them out of the analysis, does not change our results.

A second issue is that of other non-electoral reforms, such as the introduction of an intermediate appellate court. Colorado instituted an intermediate appellate court in 1971, four years after the election reform. Changing Colorado to a four year window does not change the results. Florida moved from partisan to nonpartisan elections in 1972, then moved from nonpartisan to merit-uncontested in 1977. Kentucky instituted an intermediate appellate court at the same time that it moved from partisan to nonpartisan elections. The Maryland governor began selecting new appointees by merit commission beginning in 1971.

Note that when we look at selection effects, we are holding court-specific incentives constant. The other reforms were more likely to have an incentive rather than selection effect. Still, to deal with these issues, we ran all the regressions while leaving one state out. None of the results were substantially changed in these checks.

Appendix Table OA.1 provides tabulations on the relevant treatment variation in the data. The first column of numbers gives the number of court-years where at least one treated (selected post-reform) and one control judge (selected pre-reform) is on the court that year. The second set of columns gives the number of judge-year observations in the control and treatment groups (and total). The third list of columns gives the number of distinct judges in these respective groups.

Appendix Table OA.2 lists the states by their official case assignment rule. In the analysis, we treat "random" and "rotating" assignment as random assignment.

D [Online] Additional Summary Statistics

Tables OA.3, OA.4, and OA.5 provide additional summary statistics, as referenced in the main text. We have the regression showing a relationship between our measures and bar association

Table OA.1: Tabulations on Treatment and Control Judges

Reform	Number of Court-Years with Treatment Variation	Number of Obs (Judge-Year) Pre	Post	Total	Number of Judges Pre	Post-	Total
Partisan to Merit	164	423	628	1051	45	67	112
Partisan to Nonpartisan	33	91	104	195	17	21	38
Nonpartisan to Merit	63	154	200	354	19	27	46

Notes. Summary tabulations on Pre-Reform ("Control") and Post-Reform ("Treatment") Judges. The column gives the number of court-years that are used in the estimates – that is, where at least one treated and one control judge is on the court that year. The second set of columns gives the number of judge-year observations in the control and treatment groups (and total). The third list of columns gives the number of judges in these respective groups.

Table OA.2: Case Assignment Rules on State Supreme Courts

Discretionary	Random	Rotating
Arizona	Idaho	Alaska
California	Louisiana	Alabama
Colorado	Mississippi	Arkansas
Connecticut	New Hampshire	Florida
Delaware	New York	Georgia
Hawaii	Ohio	Iowa
Indiana	South Dakota	Illinois
Kansas	Tennessee	Maine
Kentucky	Texas	Minnesota
Massachusetts	Virginia	Missouri
Maryland	Washington	Montana
New Jersey	Wisconsin	North Carolina
Oregon		North Dakota
Pennsylvania		Nebraska
Wyoming		New Mexico
		Nevada
		Oklahoma
		Rhode Island
		South Carolina
		Utah
		Vermont
		West Virginia

List of states by rules for case assignment in state supreme courts. Rules collected by Christensen et al. (2012).

Table OA.3: Quality and Output as Predictors of Bar Association Evaluations

	(1)	(2)	(3)
	Bar Evaluation: "Good Judge"		
Work Output		1.722^{+}	0.501
		(1.016)	(0.553)
Work Quality	3.553**		3.331**
	(1.209)		(0.901)
N	61	61	61
Court-Year FE's	X	X	X

Notes. $N = 61$ judge-years for the set of judges in Pennsylvania, Texas, and Washington for the years 1987 through 1994. Outcome is an indicator for being a "good" judge as defined in Lim and Snyder (2015), with mean 0.86. Coefficients estimated by conditional logit. Standard errors clustered by state in parentheses. + p<.1, * p<0.05, ** p<0.01.

evaluations, summary stats on the areas of law and related industries in the cases, and the regression showing determinants of case quality.

Figure OA.1 provides visualizations of relevant summary statistics on judge quality. We show the variation left over in work output and work quality after residualizing out state-year fixed effects. As can be seen in the figure, there is significant variation in output and quality across judges, even after controlling for institutional characteristics. We also compare the counts to rank measures. Especially for citations (panel a), we can see that the large variance and major outliers demonstrate the importance of the rank-percentile approach.

To get a look at the raw data, in Figure OA.2 and Figure OA.3 we show the average measures of output and quality respectively, for six state-specific snapshots of nine-year periods of our data. These periods were selected because they are states with reforms, and these periods have the least turnover of judges. The low turnover allows us to look at the same cohort of judges for a relatively extended period. As can be seen in the snapshots of raw data, the ranking of judges compared to their colleagues tends to persist across years.

Table OA.4: Summary Statistics on Area of Law and Related Industries

Area of Law	Freq.	Percent	Related Industrial Sector	Freq.	Percent
Criminal Law	191810	21.85	Real Estate	28527	13.64
Civil Procedure	74757	8.52	Law Enforcement	10758	5.14
Evidence	66377	7.56	Automobiles	10206	4.88
Torts	57915	6.6	Insurance	9158	4.38
Damages & Remedies	45073	5.14	Tax	8509	4.07
Contracts	40888	4.66	Construction & Engineering	6332	3.03
Real Property	36408	4.15	Workers' Compensation	5397	2.58
Constitutional Law	34038	3.88	Banking	4917	2.35
Family Law	32191	3.67	Legal & Compliance Services	4682	2.24
Workers' Compensation	22955	2.62	Automobile Insurance	4124	1.97
Insurance Law	19375	2.21	Property Management	4108	1.96
Administrative Law	18264	2.08	Transportation	3890	1.86
Wills, Trusts & Estates	18179	2.07	Child Welfare	3689	1.76
Tax & Accounting	16978	1.93	Employment Services	3679	1.76
Employment Law	14601	1.66	Health & Medical	3478	1.66
Habeas Corpus	13426	1.53	Oil & Gas	3189	1.52
Appellate Procedure	13140	1.5	Railroads	2777	1.33
Professional Responsibility	12052	1.37	Hospitals	2719	1.3
Motor Vehicles & Traffic Law	9644	1.1	Education	2586	1.24
Land Use Planning & Zoning	9122	1.04	Trucking	2097	1
Government	8942	1.02	Bridges & Roads	1751	0.84
Mortgages & Liens	7531	0.86	Agriculture & Farming	1729	0.83
Landlord & Tenant	5499	0.63	Mortgage Lending	1680	0.8
Construction Law	4997	0.57	Manufacturing	1612	0.77
Elections & Politics	4972	0.57	Real Estate Agents & Brokers	1573	0.75
Eminent Domain	4943	0.56	Unions	1485	0.71
Labor Law	4790	0.55	Financial Services	1469	0.7
Government Employees	4773	0.54	Judiciary	1448	0.69
Debtor Creditor	4260	0.49	Politics	1336	0.64
Employee Benefits	4208	0.48	Teachers	1300	0.62
Medical Malpractice	4113	0.47	Medical Procedures	1273	0.61
Personal Property	3994	0.46	Public Works	1223	0.58
Corporate Law	3958	0.45	Life Insurance & Annuities	1155	0.55
Negotiable Instruments	3843	0.44	Apartment Leasing	1127	0.54
Education Law	3803	0.43	Mining & Natural Resources	1115	0.53
Banking & Finance	3380	0.39	Drug Trafficking	1105	0.53
Alcohol & Beverage	3213	0.37	Sewer & Water	990	0.47
Civil Rights	3138	0.36	Electric	985	0.47
Health Law	2950	0.34	Water & Sewer	972	0.46
Transportation Law	2839	0.32	Physicians	966	0.46
Partnerships	2333	0.27	Firearms & Weapons	962	0.46
Natural Resources	2301	0.26	Motorcycles	919	0.44
Legal Malpractice	2285	0.26	Water	904	0.43
Products Liability	2280	0.26	Food & Beverage	888	0.42
Alternative Dispute Resolution	2144	0.24	Commercial Real Estate	883	0.42
Communications & Media	2048	0.23	Property & Casualty Insurance	854	0.41
Environmental Law	1857	0.21	Administration	837	0.4

Table OA.5: Case-Level Determinants of Positive Citations

	(1)	(2)	(3)
	Positive Cites Per Opinion (Rank)		
Administrative Cases	0.0759	0.150*	0.119*
	(0.0462)	(0.0668)	(0.0588)
Civil Cases	0.194**	0.299**	0.276**
	(0.0326)	(0.0390)	(0.0399)
Constitutional Cases	0.216**	0.324**	0.312**
	(0.0430)	(0.0671)	(0.0640)
Criminal Cases	0.192**	0.494**	0.500**
	(0.0343)	(0.0557)	(0.0562)
PCA 1	0.00222	0.00860+	0.0100**
	(0.00163)	(0.00459)	(0.00366)
PCA 2	0.00333*	0.0107**	0.00933**
	(0.00150)	(0.00242)	(0.00249)
PCA 3	0.00147	0.00193	0.00201
	(0.00179)	(0.00405)	(0.00387)
PCA 4	-0.00112	-0.00494	-0.00477+
	(0.00204)	(0.00310)	(0.00278)
PCA 5	-0.00708**	-0.0108**	-0.0108**
	(0.00188)	(0.00263)	(0.00265)
N	14996	14996	14894
R-sq	0.009	0.021	0.065
Court-Year FE's		X	X
Cohort FE /Trends			X

Standard errors clustered by state and year in parentheses. + p<.1, * p<0.05, ** p<0.01.

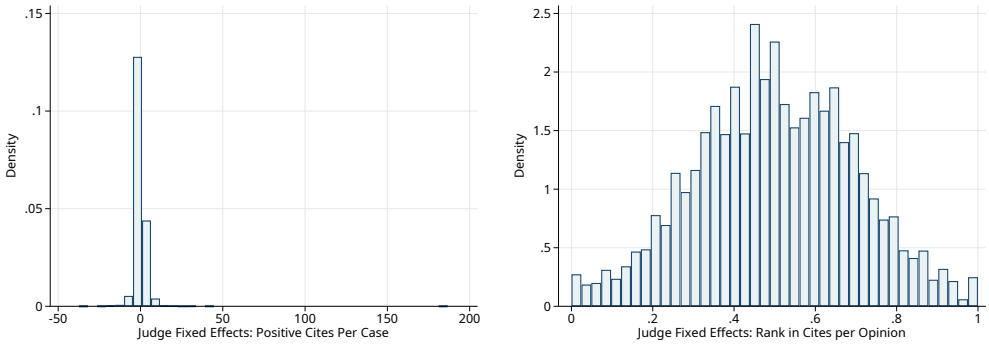
Table OA.6: Explanatory Power of Judge Fixed Effects for Work Output

	(1)	(2)	(3)	(4)	(5)
	Total Words Written	Percentile Rank in Words Written			
Marginal R^2 on	0.443	0.404	0.352	0.396	0.425
Judge FE's	(0.00855)	(0.00666)	(0.00639)	(0.00776)	(0.0119)
N	15004	14996	14996	10852	4144
Allocation Rule				Random	Non-Random
State-Year FE's	X	X	X	X	X
Case Controls			X		

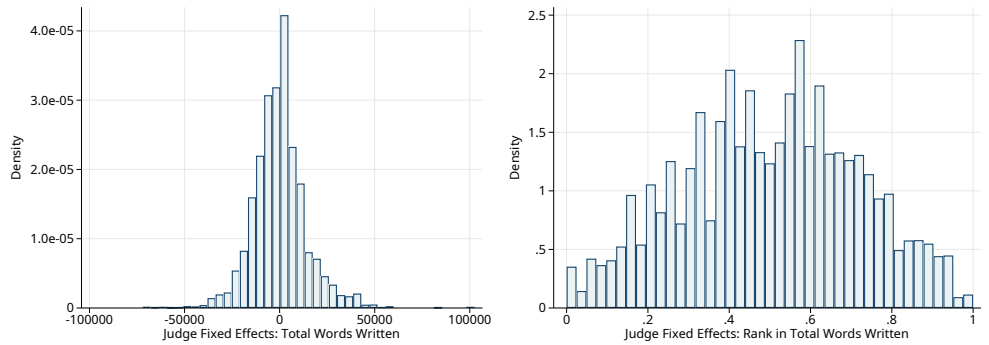
Notes. This table shows the explantory power of the judge fixed effects for the output measure. Bootstrapped estimates of the R^2 of the judge fixed effects on the stated outcome after residualizing out fixed effects and controls. "Total words written" means total words written in a year; "Percentile Rank in word written" means judges are uniformly distributed between zero and one based on rank within court-year (0 is lowest, 1 is highest). Standard error of bootstrapped estimate in parentheses. 128 bootstrap samples. Case Controls include controls for four major case types, five principal components of matrix of legal topics and related industries, and judge percentile in the number of cases seen, all fully interacted with both state fixed effects and year fixed effects. Allocation Rule means the sample is limited to Random or Non-Random states, as indicated (see lists in Appendix Table OA.2).

Figure OA.1: Distribution of Judge Fixed Effects

(a) Work Quality: Cite Counts and Rank Percentiles

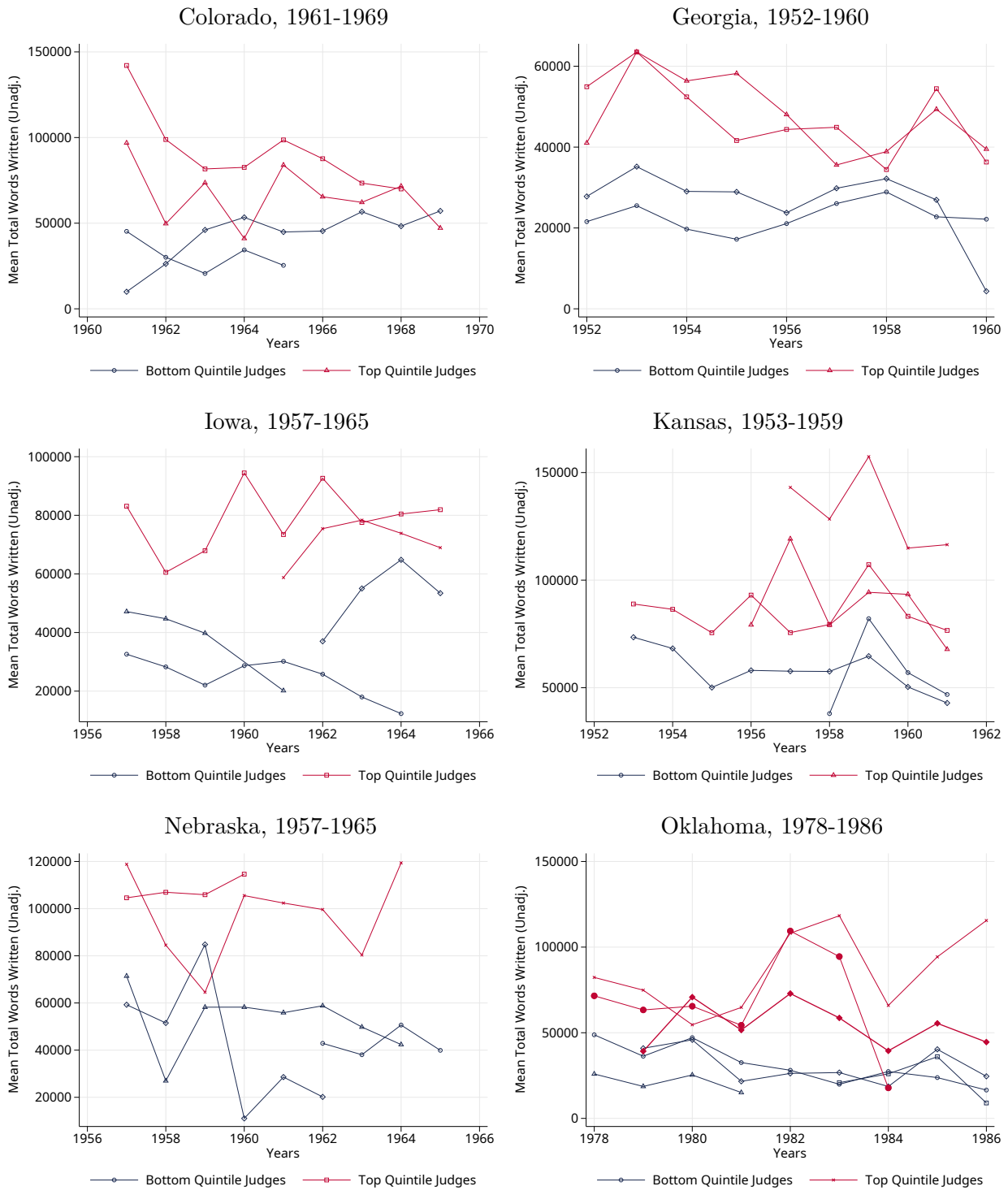


(b) Work Output: Word Counts and Rank Percentiles



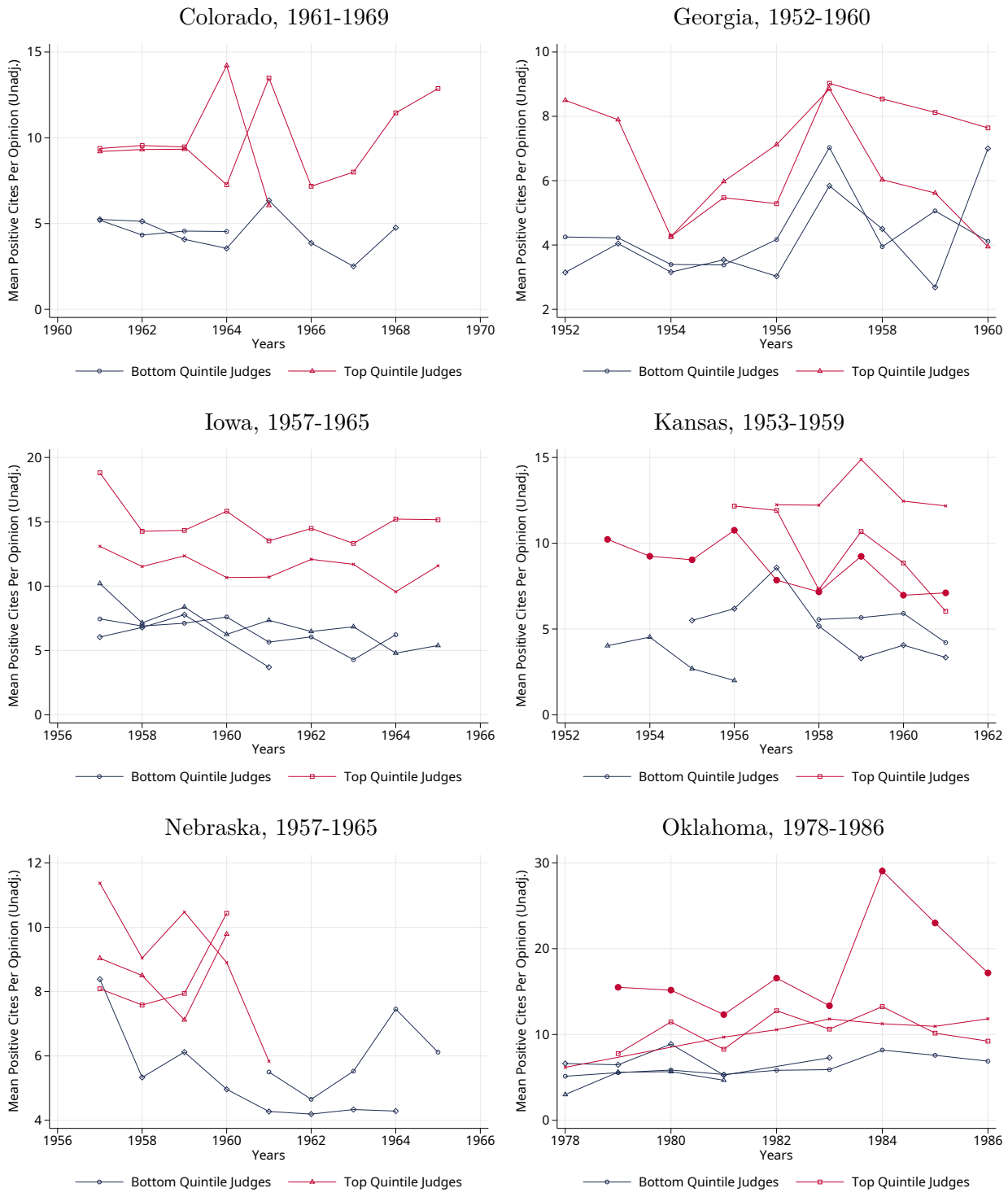
Notes. Panel (a): Judge-specific means for Work Quality: counts of positive citations per opinion, left panel, and ranks in positive cites per opinion (right panel)). Panel (b): Judge means in Work Output (number of words written in opinions, left panel, with ranks in right panel. Count measures are residualized on court-year fixed effects.

Figure OA.2: Work Output Distinctions Between Judges



Judge-year averages on work output for selection of judges in the indicated states and time periods. Top quintile and bottom quintile judges on the court indicated in red and blue, respectively. Georgia, Iowa, Nebraska, and Oklahoma have official case allocation rules requiring rotating or random assignment.

Figure OA.3: Work Quality Distinctions Between Judges



Judge-year averages on work quality (citations per opinion) for selection of judges in the indicated states and time periods. Top quintile and bottom quintile judges on the court indicated in red and blue, respectively. Georgia, Iowa, Nebraska, and Oklahoma have official case allocation rules requiring rotating or random assignment.

,ã,