NBER WORKING PAPER SERIES

MANDATORY RETIREMENT FOR JUDGES IMPROVED PERFORMANCE ON U.S. STATE SUPREME COURTS

Elliott Ash
W. Bentley MacLeod

Working Paper 28025
http://www.nber.org/papers/w28025

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2020

We thank Yisehak Abraham, Ankeet Ball, Josh Brown, Josh Burton, Matthew Buck, 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, Claudia Marangon, Sarah MacDougall, Justin McNamee, Sourabh Mishra, Brendan Moore, Arielle Napoli, Karen Orchansky, Bryn Paslawski, Olga Peshko, Matteo Pinna, Quinton Robbins, Ricardo Rogriguez, Jerry Shi, Xiaofeng Shi, Carol Shou, Alex Swift, Holly Toczko, Tom Verderame, Sam Waters, Sophie Wilkowske, John Yang, Geoffrey Zee, Fred Zhu, and Jon Zytnick for their help in assembling data and other research assistance. We thank David Card who suggested that we look at the effect of aging on judges. We also thank Decio Coviello, Janet Currie, John J. Donohue III, Hanming Fang, Lewis Kornhauser, Michael Livermore, Jonathan Petkun, Jonathan Skinner, Megan Stevenson, and numerous workshop participants for helpful feedback. Columbia University's Program for Economic Research, Columbia Law School, Princeton University's Center for Health and Wellbeing, 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.

Mandatory Retirement for Judges Improved Performance on U.S. State Supreme Courts
Elliott Ash and W. Bentley MacLeod
NBER Working Paper No. 28025
October 2020
JEL No. D02,J26,J41,J44,K0,K4

ABSTRACT

Anecdotal evidence often points to aging as a cause for reduced work performance. This paper provides empirical evidence on this issue in a context where performance is measurable and there is variation in mandatory retirement policies: U.S. state supreme courts. We find that introducing mandatory retirement reduces the average age of working judges and improves court performance, as measured by output (number of published opinions) and legal impact (number of forward citations to those opinions). Consistent with aging effects as a contributing factor, we find that older judges do about the same amount of work as younger judges, but that work is lower-quality as measured by citations. However, the effect of mandatory retirement on performance is much larger than what would be expected from the change in the age distribution, suggesting that the presence of older judges reduces the performance of younger judges.

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

An appendix is available at: http://www.nber.org/data-appendix/w28018

1 Introduction

The goal of this paper is to provide empirical evidence on how mandatory retirement rules influence work performance in high-skill occupations. The empirical setting is state supreme courts, the state-level equivalent to the federal-level U.S. Supreme Court. State supreme court judges are tasked with reviewing lower-court decisions and setting new legal rules in a common-law milieu. Researching, deciding, and justifying the law is a high-skilled constellation of tasks requiring expertise and professionalism (see Posner, 2008), comparable in technicality to physicians and scientists. Moreover, judges work alongside peer judges and supervise teams of clerks and other staff, meaning the job also entails significant social and managerial skills.

We measure judge performance over the lifespan using a database of all decisions published by state supreme courts for the years 1947-1994, combined with citation measures from 1947-2012. Because these opinions comprise the near-entirety of a judge's work product, they can be used to produce cleaner measures of performance than is possible in other high-skill domains. More specifically, we measure work performance as the number of times a judge is cited (positively) by future judges in a given year (Choi et al., 2008). While forward citations do not identify a "correct" decision, they do index the degree to which a new interpretation or clarification of the law is helpful to future judges. Hence, citations provide a measure of performance that does not require a normative evaluation of the rule applied.

Beyond the performance measurements, state supreme courts have a number of desirable features for our research objectives. First, the job of a judge has not changed substantially over time, nor does it change over the course of the career. Relative to other high-skill professions, such as medicine or management (Choudhry et al., 2005; Bloom and Reenen, 2007; Bloom et al., 2012), the skills relevant for good judging are relatively constant. Second, the workload for judges is held constant; it cannot be influenced by judges themselves, and does not vary according to age or experience. Third, compensation does not vary between judges, nor is it contingent on work performance (see Landes and Posner, 2009). Fourth, state supreme court judges have relatively strong tenure protections with little chance of termination (Kritzer, 2011). Fifth and finally, they are at the top of their profession without further opportunities for promotion. Therefore, with these judges we can track long-run changes in work performance over full careers in a high-skill setting. The resulting within-judge comparisons over time are different from those in most previous studies (e.g. Tanaka and Higuchi, 1998; Ballesteros et al., 2009), which use physical and cognitive metrics taken at single points in time and compare across individuals.

Appellate judging is a technically and professionally demanding career, at least on par with physicians, scientists, and managers. Our results will therefore be informative about how retirement policies would influence work productivity in these other high-skill professions. Beyond that, state supreme court judges are themselves an elite and powerful group. They have the authority to review not just the decisions of lower courts, but also legislation passed by state assemblies. Important common-law rules – such as contracts, real property, and torts – are made, applied, and distinguished in state supreme court decisions. These decisions have the force of binding precedent for all courts in a state and can even influence law in other states through a shared legal discourse and persuasive precedent.

Given the legal and social impacts, the quality of judge decisions is an important policy objective. In particular, the potential decline in performance due to aging could warrant a policy response. According to Posner (1995), "it is well-known within professional circles that some federal judges, including Supreme Court Justices, have continued to sit long after their judicial performance has become severely compromised by age-related disabilities" (p. 3). Recent news stories have highlighted anecdotal evidence of old age interfering with work performance in U.S. federal courts, where judges have life tenure and can stay on as long as they like. These accounts include some examples of older judges with dementia symptoms continuing with their work.[1] In addition, as the set of working judges gets older, the risk increases of a death on the court, a disruptive event exemplified by the 2020 death of Justice Ruth Bader Ginsburg.

In response to these concerns, many states have introduced maximum age constraints for full-time judges. Table 1 reports these rules and their reforms for the 1947-1994 period. In 1947 (the first year of our judge performance panel), 17 states had a mandatory retirement rule. By 1994 (the last year in the panel), an additional 14 states had adopted mandatory retirement. Besides the extensive-margin variation

[1]See the 2011 *ProPublica* article, "Life Tenure for Federal Judges Raises Issues of Senility, Dementia," available at `https://www.propublica.org/article/life-tenure-for-federal-judges-raises-issues-of-senility-dementia`. There is also the issue of uncertainty around health: e.g., "Justice Ginsurg in the Hospital Again", *New York Times*, 29 July 2020, available at `https://www.nytimes.com/2020/07/29/us/politics/justice-ginsburg-hospital.html`.

Table 1: Judge Retirement Rules and Reforms by State

A. Status Quo Rules at Period Start (1947)

Retirement Rule	List of States
No Mandatory Retirement	AR, CA, DE, GA, ID, KY, ME, MS, MT, ND, NE, NM, NV, OK, RI, TN, WI, WV, VT*
Retirement at Age 70	AK, HI, LA, MD, MA, MI,MO, NH, NJ, NY, OH
Retirement at Age 72	NC, SC
Retirement at Age 75	IL, IN, TX, UT

B. Retirement Rule Changes, 1948-1993

Mandatory Retirement Age		List of States (with Year Enacted)
Before	After	
None	70	AL (1973), AZ (1992), CT (1974), FL (1972), MN (1973), PA (1968), VA (1970), WI (1955), WY (1972)
None	72	CO (1962), IA (1965), WA (1952)
None	75	KS (1993), OR (1960)
70	None	WI (1984)

Notes. Initial retirement rules (in 1947, Panel A) and their reforms (Panel B), by state. * Vermont (VT) has mandatory retirement at age 90; we classify it as no mandatory retirement since there are just 2 judges in our entire sample (not in Vermont) who live that long.

of having a rule or not, there is additional intensive-margin variation in the maximum age: 70, 72, or 75.

Our empirical strategy uses the introduction of mandatory retirement rules for judges as a natural experiment. We ask how introducing mandatory retirement affects court performance in a differences-in-differences regression framework. Court fixed effects adjust for time-invariant characteristics by court, while year fixed effects adjust for nationwide trends affecting all courts. Our identification assumption is parallel trends, for which we provide evidence using event-study regressions.

We first look for a first stage effect. Do mandatory retirement rules affect the judge age distribution? We find that introducing a retirement age decreases the age of working judges by 2-4 years. The effect is observed across the age distribution, with the oldest incumbent judges being replaced by new younger judges.

Next, we estimate the treatment effect of retirement rules on court performance (as measured by citations to the court). We find a positive and statistically significant effect of mandatory retirement on court performance. The effect is quantitatively large, at about a 25% proportional increase in positive citations to the court. The effect is robust to a number of alternative specifications, and we find similar effects using a less restrictive citation measure (all citations, not just positive ones) and a more restrictive one (just citations from other states, where a ruling is persuasive rather than binding precedent). We find that the effect is mostly driven by quantity (an increase in opinions produced), with smaller positive effects on quality (citations per opinion).

Why does mandatory retirement improve court performance? A salient possibility is that older judges do lower-quality work, so replacing them with younger judges would mechanically increase quality. In the second part of the analysis, we measure changes in judge performance over the lifespan in order to see if the mechanism at work is the removal of lower-performing judges. Our empirical strategy exploits the panel variation in our data across courts and across time. We first compare colleague judges by age within the same court, and second, looking at changes in performance within judge across the lifespan. We also compare judges based on starting age and ending age.

These regressions produce systematic evidence that performance falls with age. Judges improve in performance only at the youngest ages, below 50. The effect of age is quadratic, consistent with an accelerating decline in work quality. The effect is robust to an array of specifications for performance, including in rank percentiles within court-year. There are only minimal effects on the workload of older judges (number of cases decided), but quality (cites per case) decreases substantially.

The effects of aging are not driven by the types of cases that judges review. We can rule out experience as an important factor. Selective attrition (exit timing) is important, but works against our effect because the judges who end up working longer tend to perform better at earlier ages. In an auxiliary dataset, we show that clerk quality (ax proxied by law school ranking) is uncorrelated with judge age.

Using these life-cycle estimates, we can show that the aging effect does not explain the mandatory retirement effect. That is, the retirement reform's effect on performance is greater than that expected from the shift in the age distribution. In particular, while the reform effect is driven mostly by quantity, the aging effect is driven mostly by quality. Moreover, a significant component of the retirement reform effect comes from a within-judge shift, where the younger judges on the court increase their output after the reform. These results point to a quantitatively important *team effect* of aging, in which the presence of older judges slows down the pace of work on the court.

To summarize, we find that the introduction of mandatory retirement reduces judges' age and increases court-level work quality -- total citations per year. We document a connection between these effects in that judge-level work quality is decreasing with age. Overall, these results support the view that the effects of mandatory retirement on quality are due to the replacement of older judges with younger ones, who tend to provide higher-quality work on average.

These results add to the significant and active literature on the economics of aging and retirement (e.g. Lumsdaine and Mitchell, 1999). First and in particular, we add to the empirical literature on mandatory retirement. One of the closest papers is Ashenfelter and Card (2002), who find that banning a mandatory retirement age for university faculty significantly increased the share of older academics. Frederiksen and Flaherty Manchester (2019) look at the private sector and find that banning mandatory retirement reduced the share of long-term employment contracts and increased the prevalence of performance pay. Our setting is different in that we observe the *implementation* (rather than removal) of mandatory retirement. Unlike Ashenfelter and Card, we directly observe changes in productivity in addition to changes in worker age. Moreover, our judicial context holds fixed other aspects of the contract (e.g. performance pay), allowing us to isolate a productivity effect.

A second related literature is the structural work on retirement choice, which has analyzed worker responses to pensions and other retirement incentives (Gustman and

Steinmeier, 1986; Stock and Wise, 1990; Gustman and Steinmeier, 1991, 2005). In political economy, Diermeier et al. (2005) and Keane and Merlo (2010) derive structural estimates of the parameters underlying retirement choices of U.S. Congressmen. The structural approach allows rich counterfactual exercises but relies on strong assumptions of the underlying choice model. Our approach is different in that we focus on producing causal estimates using natural experiments changing the retirement rule.

A third strand of this literature has focused on diminishing productivity in the later years of life. Although cognitive abilities begin to fall early in the career (Desjardins and Warnke, 2012), workers also learn and enhance their skills.[2] Thus, wages and employment continue to rise after the start of physical and cognitive decline and only begin to fall for individuals older than fifty years (Medoff and Abraham, 1980; Abraham and Farber, 1987). These stylized facts underlie the standard economic model, where younger individuals invest in human capital that depreciates over the lifespan (e.g., Blundell and Macurdy, 1999). Empirically, a concave relationship between age and productivity has been found for scientists (Levin and Stephan, 1991), economists (Oster and Hamermesh, 1998), and physicians (Choudhry et al., 2005).[3] Like these previous papers, we measure work product (opinions) and quality (citations) directly. However, our estimates of the age-performance trend are somewhat easier to interpret due to the empirically desirable features of state supreme courts. Unlike scientists or physicians, the workload, compensation, and other contract features are held relatively constant over the lifespan. State supreme court judges have relatively strong tenure protections with little chance of termination nor of promotion. Moreover, the team performance of judges on the court is also directly and jointly observed – that, combined with our mandatory-retirement natural experiments, can be leveraged to disentangle individual

[2]In the words of cognitive psychology: While pattern recognition and logic skills (fluid intelligence) begin diminishing at a young age, verbal skills (i.e. writing skills) and knowledge (crystallized intelligence) improve into relatively advanced ages (Desjardins and Warnke, 2012). Another way of saying this, from Ramscar et al. (2014), is that as people age they have a larger data set in their mind. This in turn leads to slower processing speeds as they search larger data sets.

[3]Choudhry et al. (2005) conclude that "older physicians possess less factual knowledge, are less likely to adhere to appropriate standards of care, and may also have poorer patient outcomes." In the case of assembly-line workers, Borsch-Supan and Weiss (2016) find no evidence of a performance decline before the age of 60. Using a database of personal credit choices, Agarwal et al. (2009) find a U-shape pattern in the frequency of financial errors (such as suboptimal balance transfers) with optimal choicemaking at age 53. Small et al. (2011) report a within-person lab study where episodic/semantic memory demonstrated no decline before the age of 75. Galenson (2011) explores the different life cycles in productivity across professions, focusing on "young geniuses" who peak early and"old masters" with sustained increases in performance over the lifespan.

from team effects of aging. We find that both are significant.

The previous evidence on judge productivity over the lifespan is mixed. In an early and detailed study of U.S. federal judges, Posner (1995, ch. 8) finds that opinion quality (citations per opinion) is maintained into advanced age (into the 80s) before decreasing. In a sample of twenty judges on the Australia High Court, Smyth and Bhattacharya (2003) find that quality (as measured by citations) peaks at age 64, which is somewhat later than in our setting. Teitelbaum (2006) provides descriptive time-series evidence that the U.S. Supreme Court produces about the same number of cases regardless of the average age of the justices. Dimitrova-Grajzl et al. (2012) compared performance by age in three Slovenian trial courts; in one court, there was a concave relationship between productivity (number of cases resolved) and age, and in the other two courts, there was no relationship. Our study contributes to this literature by using a larger sample size of judges across many courts, measuring a broader array of performance measures, and analyzing panel variation in mandatory retirement rules. These features allow us to look across courts for team effects, which have not been found in the previous literature.

More generally, these results adds to the emerging literature on how organizational practices influence performance of public-sector professionals. This literature includes Prendergast (2001) and Shi (2009) on police performance, Bloom et al. (2015) on education, and Coviello et al. (2015, 2014) on work organization in the Italian judiciary. We add to this literature by showing that retirement rules work not just at the level of individuals, but also at the level of organizations.

The rest of the paper is organized into the following parts. Section 2 provides some background and describes the data. Section 3 provides the empirical analysis of how mandatory retirement reforms affect judge performance. In Section 4, we analyze variation in performance over the judge life cycle. Section 5 reviews these patterns of evidence and explores team effects of aging. Section 6 concludes.

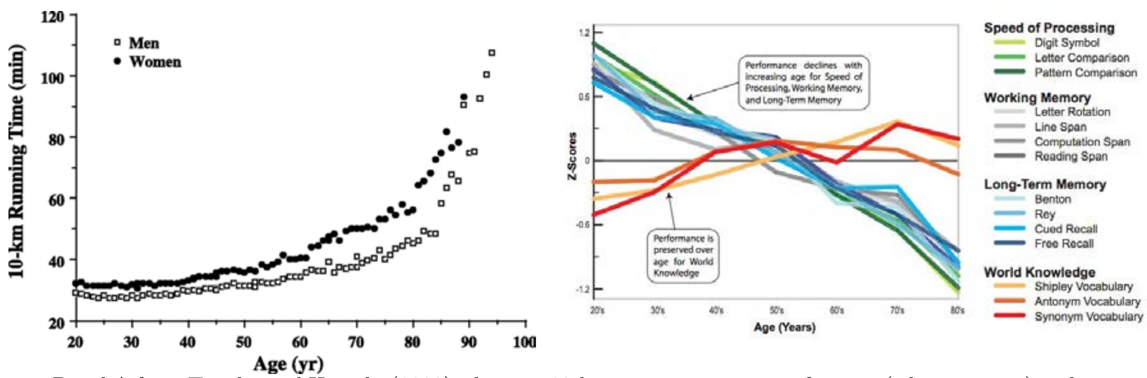
2 Background and Data

2.1 Overview

At some point all good things must come to an end, including our careers. A key factor in this decision to stop working is the inevitable depreciation in skills with age. Take

Figure 1: Performance vs. Age for Physical and Cognitive Tasks

(A) 10-K Running Time **(B) Cognitive Performance**



Notes. Panel A from Tanaka and Higuchi (1998), showing 10-km race running times for men (white squares) and women (black squares) by age. Panel B from Ballesteros et al. (2009) showing how measures of different factors of intelligence or cognitive performance from psychological tests vary by age. The green lines measure processing speed, the gray lines working memory, blue lines long-term memory, and red lines world knowledge. All are decreasing into old age except knowledge.

the case of professional runners: As shown in Figure 1 Panel (A), the ability to run 10 kilometers falls continuously from about age 40, and there is a very steep decrease around age 85.[4] For any given professional runner, the decision to end his/her career is simple because running speed can be easily measured. The same goes for any job that depends on skills that can be cheaply and accurately measured.

Most jobs in a knowledge economy are not based on physical speed or other singular performance dimension. But aging still takes its toll: As illustrated in Figure 1 Panel (B), cognitive ability declines continuously starting at age 20. The long-run decline in processing power is partly compensated by improvements in memory and knowledge. The presence of multiple countervailing factors reflects that the decision on when to end a cognitive career is more complicated than one based on running speed.

More generally, one cannot usually measure the productivity of skilled professionals. Age-related performance decline takes several years, and hence it may not be clear when to step down. Moreover, if a professional is permanently employed, then dismissal must be with cause. In such a context, the employer must provide documentary evidence that performance is not sufficient, and naturally the employee might disagree, or may even bring an age discrimination claim.

[4]This decline is in spite of a great deal of positive selection: individuals who are still able to run a 10-K at age 85 are a very selected group!

The process of ending careers is further complicated by nominal rigidity in wages. For many reasons, wage cuts are not usually an option for older workers. Part of this is personal: Few managers would relish telling a valued employee that his/her performance has declined to the point that the wage must go down or else regular employment must cease. Given the aforementioned difficulty in measuring performance, employers will find it difficult to produce decisive evidence that a wage cut is warranted. Meanwhile, a predetermined age-based wage cut – even if it would be accepted by workers – would require an accurate estimation of age-related productivity decline (perhaps many years in the future), and would require (inefficiently) identical treatment of workers with highly heterogeneous performance trajectories. For these reasons, firms cannot bring the compensation of older workers in line with performance when performance declines.[5] As health choices and technologies improve, one could expect even more dramatic variation across workers in longevity and late-life productivity.[6]

A potential solution to these end-of-career conflicts is the use of mandatory retirement provisions (Lazear, 1979). When parties have agreed in advance to a specified retirement date, then the employee's performance is not being called into question during the termination process. Hence, the employer does not need to provide evidence that an individual's performance is no longer acceptable. Beyond avoiding such conflicts, another benefit of such a rule is that it helps employees plan in advance their savings, so that they may best enjoy the fruits of a long working life. A potential cost is that many individuals are still productive into advanced age and would prefer to continue working in some capacity. Yet under mandatory retirement the employer can, at her discretion, continue to employ the retiree in a similar job, or in a job for which she is better suited.

Regardless of the net benefits, mandatory retirement is prohibited (with some exceptions) in the United States under the Age Discrimination Act of 1967.[7] The justification for this prohibition is that mandatory retirement by age is facially discrim-

[5]Frederiksen and Flaherty Manchester (2019) provide evidence that firms have historically addressed the problem of varying performance trajectories by keeping base wages low and using more performance pay. This is also consistent with the evidence in Lemieux et al. (2009) that find an increased use of performance pay over time for highly skilled professionals.

[6]In the United States in 2010, one could expect about 19 more years of life conditional upon reaching 65; this number was up from about 14 years in 1960. Appendix Figure A.2 Panel (a) shows the average retirement age by year, which has been stable for men but increasing significantly for women.

[7]The Age Discrimination Act of 1967 is designed "to promote employment of older persons based on their ability rather than age". The text of the law is excerpted in Appendix E.

inatory toward workers above that age. Instead, employers are required to evaluate the performance of the worker so that potentially high-productivity workers are retained. Economic theory provides little guidance on whether mandatory retirement dates should be allowed: with costless renegotiation mandatory retirement would have no effect or, if it does have an effect, that private parties would themselves choose the optimal rule (see Lazear, 1979; Frederiksen and Flaherty Manchester, 2019).

Separate from the question of legality, it is an open empirical question whether mandatory retirement improves work productivity. If performance declines with age, then reducing the average age of workers to some degree would improve productivity. On the other hand, if high-productivity older workers are removed during their prime due to the mandate, that could reduce productivity. In addition, with a looming mandatory retirement date workers might reduce investment in job-specific skills. To see which effects dominate, empirical evidence is needed.

The profession of judging has some special features which should be mentioned. Posner (1995, ch.8) discusses some reasons that judges, relative to professionals in other domains, might be able to maintain a high level of work quality into advanced age. In Posner's words, judging is a "late peak, sustained" career, where productivity increases into advanced age and then sustained until near death. In particular, many of the tasks facing judges draw on crystallized intelligence, which is maintained into advanced age. Similarly, Grossmann et al. (2010) show that reasoning about social conflicts – the raison d'être of judges - improves into old age, as indicated by higher-order reasoning emphasizing the need for multiple perspectives and compromise. Meanwhile, Lindenberger (2014) suggests that frequent participation in intellectual challenges and social engagement – both salient features of appellate judging – may mitigate cognitive decline.

These factors of job performance as well as professional engagement are both important determinants of the judge retirement decision, which tends to come later than that for the average U.S. worker (see Appendix Figure A.2). The lengthy career reflects in part a high intrinsic professional motivation on the part of the judges (Ash and MacLeod, 2015), indicated for example by the fact that a small minority of federal judges take senior status (a reduced caseload a full salary) as soon as it becomes available (Posner 1995, pg. 186; see also Choi et al. 2013). Meanwhile, a related literature in political science has shown evidence of strategic retirement – judges timing their retirement based on the party of the appointing governor or president to influence the political ideology of the successor judge (e.g. Nixon and Haskin, 2000).[8]

2.2 Institutional Context

Our empirical setting is state supreme courts. While judicial systems do vary from state to state, they share major characteristics and structures. The fundamental role of a state supreme court judge is to rule on questions of state law (rather than federal law). These questions arise in cases appealed from lower state courts. A case begins when a plaintiff files a lawsuit or a prosecutor indicts a criminal. At trial, facts are litigated and a judge/jury gives a verdict, which the losing party can appeal. If the state has an intermediate appeals court (as most do), that court will then take the case and may affirm, reverse, or modify the trial verdict. After this intermediate court's decision (or after the trial decision when the state does not have an intermediate appellate court), the ruling can be appealed to the state supreme court, which is the last appeal on matters of state law.[9]

If the state supreme court accepts a case for review, the judges will rehear the case at oral argument and then review the submitted briefs for legal error. Each judge votes whether to affirm or reverse the lower decision. One of the majority judges then researches and writes an opinion explaining the decision (with the help of clerk staff). In our data we cannot directly disentangle a judge's work from that of their clerks. We return to this issue further below.

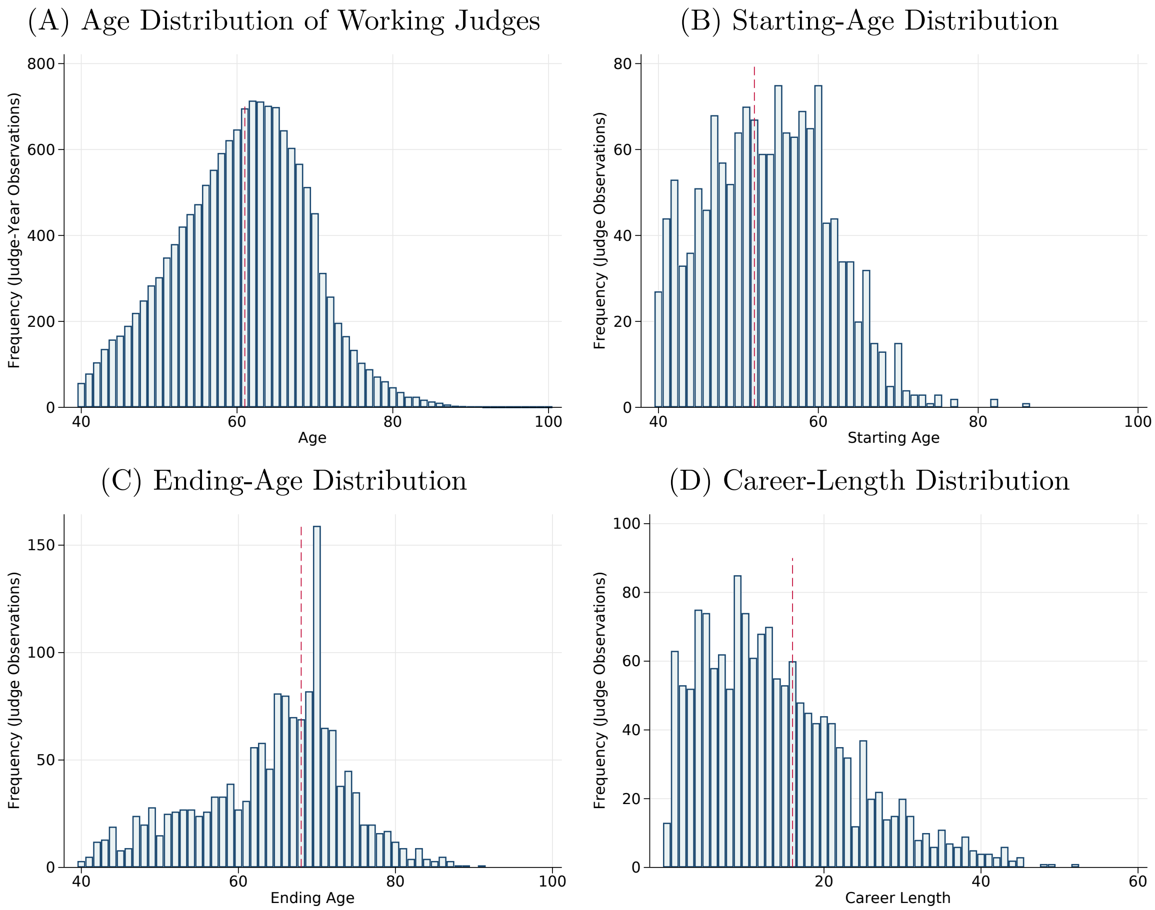
2.3 Judge Age and Retirement Decisions

The starting point for data collection is the existing data on state supreme courts from Ash and MacLeod (2015). A team of research assistants collected these data from a range of sources and built biographies for each judge in the sample. The key sources include state court web sites, judge obituaries, and Marquis Who's Who. Items that were unavailable from these sources were obtained through records requests or interviews of state court administration staff.

[8] Retirement choices have also been used for identification of electoral-retention effects. For jurisdictions with competitive reappointment processes, judges planning to retire do not face the same retention-related incentives as judges who intend to stay in office (Shepherd, 2009a; Gordon and Huber, 2007; Shepherd, 2009b).

[9] In rare cases when federal law (rather than state law) is pivotal, state supreme court decisions can be appealed to the U.S. Supreme Court. In two states (Texas and Oklahoma), there are separate high courts for criminal and civil matters.

Figure 2: Summary Histograms on Judge Age



Notes. Distributions of judge age, starting age, ending age, and career length, as indicated. Vertical dashed line at median.

The dataset includes 1,558 state supreme court judges, for which we were able to obtain birth date information for all but 3 individuals. Beyond birth, we collected date information on judgeship starts, judgeship terminations, and judge deaths. We also collected information on how judgeships ended and previous and subsequent career information.

Figure 2 provides some visual descriptives on the age and retirement decisions of state supreme court judges. These graphs only include states that do not have a mandatory retirement age. First, Panel A shows the age distribution for all state supreme court judges working between 1947 and 1994. We can see a wide range of ages of active working state supreme court judges. The other panels show the distributions of the starting age (Panel B), ending age (Panel C), and career length (Panel D). Judges tend to start in their position late in life (in their 50s) and work late as well (into their 70s). Appendix Figure A.1 shows the wide distribution over birth cohort decades in the dataset (from the 1860s to the 1950s).

The average retirement age of judges has not changed much over the time period 1947-1994 (Appendix Figure A.2, Panel B). This is somewhat different from the long-run changes in the broader economy which include an aging work force (Appendix Figure A.2, Panel A). The retirement age for judges has been consistently 4-5 years higher than the average worker over this time period. Meanwhile, the proportion of judges working in the private sector after leaving the court has increased over this time period (Appendix Figure A.3).

2.4 Mandatory Retirement Rules

The second data collection performed is on the rules on mandatory retirement. Table 1 reports these rules and records on their reforms in the 1947-1994 period. As discussed already in the introduction, we have 14 reforms to analyze in the empirical part. These include maximum ages of 70, 72, or 75.

In some states, the chief justice has the right to keep on a retired judge in what the federal courts call "senior status" or "active retirement status". Appendix Table A.2 presents some information on how these rules work by state. When senior status is available under mandatory retirement, the control rights on the retirement decision switch to the court, which can decide to keep a high-performing judge.

One practical implication of senior status is that we do see some judges in our data working past the official retirement age. This can be seen in Figure 3, showing the probability of retirement at any given age, separately by the mandatory retirement rule. The blue line, with no mandatory retirement, is relatively smooth, peaking in the early 70s. The red line, with mandatory retirement at age 70, shows big increases for ages 69 and 70. We see corresponding jumps for retirement at 72 (green line) and 75 (purple line). We do see, however, that the rules are not perfectly enforced. Some judges stay on past the mandatory retirement age due to senior judge status. Since the maximum age rule is not 100% binding, our estimates from the rule changes should be interpreted as intention-to-treat effects.[10]

[10]Results using variation in senior status rules are reported in Appendix Table B.3. We did not find differential impacts of reforms in states with formal senior-status policies.

Figure 3: Retirement Rates by Age, by Mandatory Retirement Age

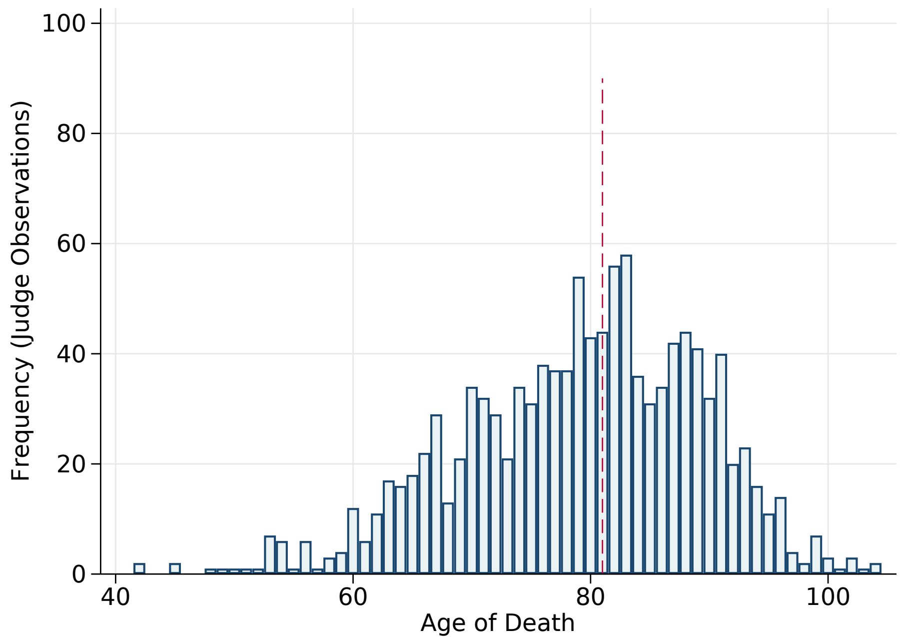
Notes. Probability that a judge retires at a particular age, conditional on working at that age. Plotted separately by mandatory retirement rule.

To round out the evidence on the judge life cycle, Figure 4 provides some statistics on the timing of judge deaths. Panel A shows that the judges have relatively long lifespans, with most living into the eighties. Panel B looks at how judge retirement is related to judge longevity, separately for mandatory retirement (left panel) and voluntary retirement (right panel). The figure shows that with voluntary retirement, judges are much more likely to die within a year of leaving office. This difference supports the idea that mandatory retirement is an impactful policy: judges are more likely to stay in their jobs until death under voluntary retirement. On the other hand, there is still a relatively high chance of death in the first year out of office under mandatory retirement (left panel), which may hint at a causal impact of retirement on mortality (as found in Sullivan and von Wachter, 2009). This is a promising area for future work.[11]

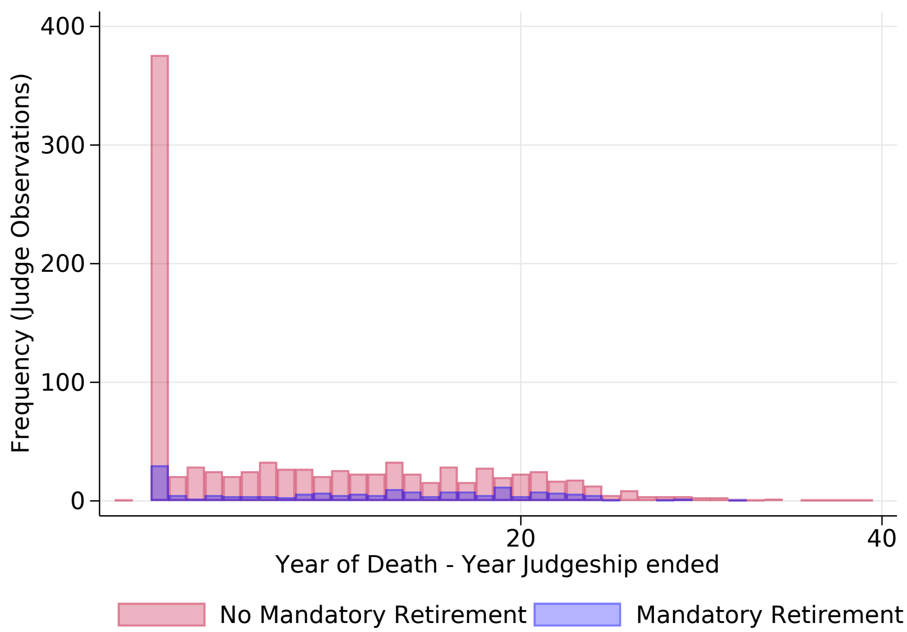
[11]Meng et al. (2017) review the mixed evidence about the cognitive impacts of retirement, noting that there is a "major knowledge gap in regards to the impact of retirement on cognitive decline."

Figure 4: Mandatory Retirement and Deaths on the Job

(A) Judge Age-at-Death Distribution



(B) Distribution of Years Between Termination and Death, With/Without Mandatory Retirement



Notes. Distributions of judge age at death (panel A, vertical dashed line at median) and death year minus year judgeship ended (panel B).

2.5 Measuring Judge Performance

Our performance measures are constructed from published state supreme court opinions for the years 1947 through 1994, obtained (along with some annotated meta-data) from Bloomberg Law. The metadata include records on forward citations up until 2012. The full sample includes 1,024,261 cases. We drop opinions that do not have a named author (per curiam decisions), resulting in a sample of 404,928 majority opinions.[12] In the final sample we have an average 25.7 cases per judge per year.

As mentioned in the introduction, a special feature of appellate judging is that one can obtain useful measures of individual work performance. While there is a collaborative element of judging, almost all significant decisions have a single authoring judge who is individually responsible. This is different from many other occupations, such as science/innovation, which are inextricably collaborative and it is technically challenging to disentangle the contributions of individual team members.

Another nice feature of judging is that the quantity of assigned work is constant. When scientists or inventors get more papers or patents, that is the result of a joint choice on the number of projects and work quality on those projects. So one cannot disentangle the work quality choice from the extensive margin. In the courts, judges have to do the same number of cases as every other judge, so it is possible to extract the work quality factor.

Our preferred measure of judge performance is *work impact*. This is the total number of positive forward citations to a judge in a year. Judges in a common-law system cite previous cases that are useful to their decision, and therefore citations can be seen as an expert evaluation of peer decision quality (Posner, 2008; Choi et al., 2010; Epstein et al., 2013; Ash and MacLeod, 2015).[13] Note that this is not a measure of whether the decision is correct or not, which we do not observe. But on average, more citations means that a case was useful to a future judge and thereby makes a stronger influence on the path of the law.[14]

Ash and MacLeod (2015) measure the effect of exogenous reductions in workload for judges. Judges could consume this extra time in leisure, but we find that judges

[12]Appendix Figure B.3 shows our retirement reform results with the full dataset of cases, collapsed by court-year rather than judge-year. The results are similar.

[13]This expert peer evaluation is quite different from Carrell and West (2010), for example, who use a quality evaluation by non-peers (college students evaluating lecture instructors).

[14]In U.S. states, judges are hired by the governor or legislature, not by the judges themselves. This means that, unlike scientists, citation choices are not used to strategically influence future hiring or promotion decisions.

use the time to produce more highly cited decisions. This result is consistent with the hypothesis that judges themselves value citations as a measure of performance. In addition, relative citation counts across judges within court persist over time, consistent with a judge-specific ability component. If citations were due to other factors, then under random assignment of cases there should be no persistent variation in citations due to a judge fixed effect. Meanwhile, it remains a challenging open question whether or not more highly cited decisions have a positive effect upon measures of social well-being, such as economic growth or inequality.

Appendix Figure A.4 shows the distribution of the number of positive citations at the case level. We can see a steadily decreasing power-law distribution with a median of about 7 cites per case. While there is a large minority of cases with zero cites, overall the distribution is smoothly and slowly decreasing. Citations are relatively widely distributed, rather than showing up only in a few blockbuster cases.[15]

Citations are annotated as positive, negative, or distinguishing by the data provider. For the baseline, we look only at positive citations. As a more inclusive measure of performance that does not rely on subjective annotations about "positive", we use all cites (including negative and distinguishing).[16] As more restrictive measures of performance, we use discussion cites (where the case was discussed at length by the citing court) and out-of-state cites (only citations in other jurisdictions). Because state supreme court precedents are not binding in other states, out-of-state citations serve as an especially strong signal of legal usefulness or influence (Choi et al., 2010). In addition, while older judges might have time to network and influence colleagues in their own court to earn cites, this concern is less pronounced for out-of-state cites.

The citations data were collected in 2012, so all citations through 2012 are included. Since the most recent cases are in 1994, we are able to measure the impact of work with at least 18 years of performance data. The median delay between a case and a forward citation is about ten years. Appendix Figure A.4 shows that the distribution of citation counts per case is close to log normal, with some extra mass at zero. About half (49 percent) of cases have at least one citation, and the average case receives 5.52 cites. Given that our window is at least 18 years after a decision is published, that ensures

[15]We show robustness of our results to dropping these blockbuster cases (formally, the top five percent of cases in each court-year by citations).

[16]The citation measure available from the data provider includes citations to concurring and dissenting opinions. From annotating a sample of cases, we found that (unlike the U.S. Supreme Court) citations to discretionary opinions are extremely rare in state supreme courts (well under 1% of cites).

Table 2: Summary Statistics on Outcomes

	Levels		Logs	
	Mean	S.D.	Mean	S.D.
Positive Cites	297.7	273.5	5.360	0.954
All Cites	355.9	322.5	5.521	0.982
Out-State Cites	44.60	81.27	3.297	0.998
Discussion Cites	66.77	53.54	3.912	0.878
# of Opinions	25.73	15.86	3.131	0.565
Cites per Opinion	13.05	12.52	2.419	0.668
# of Words Written	56352.7	32538.4	10.77	0.622
Addendum Opinions	6.400	9.253	1.492	0.995

Notes. Summary statistics (mean and standard deviation) on judicial opinion outcomes, at the judge-year level.

that we capture most of the variation in case citations. Moreover, we are comparing judges conditional upon year of the opinion (year fixed effects), ensuring that sample selection of citations does not affect the relative ranking of cases. In the empirical distribution one does observe a drop off in citation counts around year 18, though it is also clear that the bulk of citations occur before then.

Our measure of work impact is a combination of quantity (number of opinions) and quality (number of citations per opinion). For analyzing the impacts of reforms, such as mandatory retirement, we feel that the impact measure is the most policy-relevant. To the extent that the number of opinions is invariant, work impact is a measure of work quality. Still, to decompose the importance of quantity and quality, we report as additional outcomes the number of opinions written (quantity), and the number of positive citations per opinion (work quality). In addition, as a measure of *work output* we report effects on the total number of words written in opinions during a year. The appendix includes analysis for a range of other outcomes, including measures of caselaw research and number of discretionary opinions written.

For all these variables, the baseline measure is the log of the average value for the judge in a year. Summary statistics for all of our outcome measures are reported in Table 2. The main outcome is the log of positive citations, with additional measures meant to provide additional dimensions of performance, quantity, and quality.

Appendix Figure A.5 shows the distributions for the main outcome variables.

We report a range of alternative specifications for the outcomes in the appendix. When comparing judges to their colleagues, we also highlight the use of a rank percentile specification. These transformations are discussed further below.

2.6 Case Assignment and Case Characteristics

The citation count for a decision is a joint product of both the type of case and the judge's efforts. For example, cases that review the constitutionality of statutes will generally get more citations than summary habeas denials. When looking at the effects of reforms or aging on quality, we have to check whether that is driven by changes in the composition of the caseload, rather than changes in judge work quality.

A relevant institutional rule is how cases are assigned to judges, especially when comparing judges to their colleagues. There are three systems for case assignment, collected by the State Supreme Court Project (`http://www.ruf.rice.edu/~pbrace/statecourt/`) and updated by Christensen et al. (2012).[17] Appendix Table A.4 lists the state supreme courts by rule. Discretionary assignment by the chief justice (the rule at the U.S. Supreme Court) is the minority rule followed in just 15 states. In 13 states, cases are randomly assigned by lottery to authoring judges. In the remaining 22 states, cases are assigned on a rotating system, with cases arbitrarily assigned to judges based on their order on the docket.[18] Christensen et al. (2012) found that in state supreme courts, case characteristics and judge characteristics are correlated even under random/rotation assignment. This is important for interpreting any effects, which could be due to changes in case types.

At the decision level, we have data on the area of law of a case, as well as the related industries of a case. These are coded for each case by the data provider, and there may be up to three legal areas and three related industrial sectors for any particular case. Appendix Table A.3 reports summary tabulations of these characteristics. In the data, we include a vector of dummy variables for each area and sector, equaling one if the case is annotated as that area or sector. Because there are so many of these

[17]These rules were confirmed by Brace and Hall in the early 1990s and late 2000s. We tried to check the rules for earlier years. We could not get comprehensive information, but for those states where we could find information, it comported with the Christensen et al. (2012) information.

[18]There are complex rules across states that affect the rotation. Senior judges have fewer cases. Judges can occasionally recuse themselves. On appeal after remand, the same panel normally reviews a case. There can be exceptions for specialized cases such as those involving the death penalty.

categories, including separate covariates for every category would almost saturate the dataset. Instead, we construct the first five principal components of this matrix of categorical variables, which explains 65% of the variance.[19] We will uses these factors as controls, but also look at how they respond to the treatments.

3 Mandatory Retirement and Court Performance

3.1 Empirical Approach

The main estimating equation is

$$y_{ist} = \alpha_s + \alpha_t + \rho M_{st} + \alpha_s \cdot t + X_s^0 \cdot \alpha_t + X'_{st}\beta + \epsilon_{ist} \tag{1}$$

where y_{ist} is an annual performance metric (e.g. log citations, described in Subsection 2.5 above) for judge i working in court s during year t. Coefficients are estimated by ordinary least squares regression with standard errors clustered by state. The right-hand side items in the equation are described as follows.

The empirical approach is generalized differences-in-differences. To control for time-invariant court characteristics that may be correlated with the retirement system and with performance, we include court fixed effects α_s. To control for national trends in performance, we include year fixed effects α_t. Standard errors are clustered by state to allow correlation in the residuals over time across judges and within time in the same state.

The treatment indicator M_{st} equals one for years after introducing mandatory retirement. Fourteen states introduced a mandatory retirement age during the time period of our data (see Table 1). The coefficient ρ measures the corresponding causal effect of interest. Due to the length of the panel, in the baseline specification we estimate effects in a ten-year window before and after the reforms.[20] Formally, X_{st} includes an indicator equaling one for the baseline time window of ten years before and ten years after a change to the retention system. In turn, M_{st} is a dummy for the ten years after the change. Thus, as y_{st} is specified in logs, the estimates can be interpreted as the average proportional difference in within-court performance for the

[19] Using more or fewer components does not change anything.

[20] In Appendix Table B.2 we use a six-year window, fourteen-year window, or no window (all years). The main effects are robust across these specifications.

ten years after the policy change relative to the ten years before the policy change. For additional flexibility, we also allow for state-specific treatment windows.

The question we address in this section is the effect of mandatory retirement on *court* performance. In our sample the number of judges varies by court, and so the effect of mandatory retirement is correlated to court size. We address this issue by running the regression at the judge level, so that the coefficient on M_{st}, ρ, is interpreted as how mandatory retirement affects the per-judge performance of the court. By clustering at the court level, the standard errors are adjusted for variation in the size of the court.

A number of factors can motivate the introduction of a mandatory retirement age. Having older judges on the court could be problematic if they are ill and have to recuse themselves from more cases. With older judges, there is a higher risk of a judge having to step down unexpectedly due to illness, or even to die on the court. The courts have an odd number of judges to avoid deadlocks; having a missing judge increases the likelihood of a deadlock. Under voluntary retirement, there could be issues of inefficient strategic timing in order to influence the political affiliation of the successor. For our purposes, the most important potential driver is (the perception) that older judges are not delivering sufficiently high decision quality.[21] On the other hand, there have been some recent moves to repeal mandatory retirement rules, so the pressure to change operates in both directions.[22]

In any case, these reforms are not exogenously assigned. But that is not needed for identification. Consistent estimation of ρ requires parallel trends between treated states and comparison states (Bertrand et al., 2004). Put differently, the comparison states should provide a counterfactual for the trend in the treated states in the absence of the rule change. Still, the standard assumption of parallel trends is strong in this setting if the reforms are implemented in response to pre-existing trends in performance. In our regressions, we allow for pre-existing state trends in performance that may be confounded with the reforms by including state-specific linear trends $\alpha_s \cdot t$. Further, the term $X_s^0 \cdot \alpha_t$ includes initial-period court characteristics – institutional rules, case types, and judge age distribution – interacted with year fixed effects. These covariates allow for different trends along these different dimensions of state court characteristics.

To formally test for parallel trends and assess the dynamics of the effect, we use a

[21] See (Posner, 1995, ch.8) and Goldstein (2011) for discussions of the issue with respect to federal judges.

[22] See `http://ncsc.contentdm.oclc.org/cdm/ref/collection/judicial/id/440` for an update on the state situation.

panel event-study specification. Formally, we estimate

$$y_{ist} = \alpha_s + \alpha_t + \sum_{k=-6,k\neq-1}^{12} \rho_k M_{st}^k + \alpha_s \cdot t + X_s^0 \cdot \alpha_t + X_{st}' \beta + \epsilon_{ist} \quad (2)$$

where all of the items are as above, except the singular treatment indicator M_{st} is replaced with a sequence of event-study year indicators M_{st}^k. We let k index the years before and after treatment, with the year before treatment being left out as the comparison year. Then $\hat{\rho}_k$ give the dynamic effects on performance k years before/after the reform. Parallel trends are consistent with $\hat{\rho}_k = 0$ for $k < -1$. For $k >= 0$, $\hat{\rho}_k$ will elucidate the dynamics of the difference-in-differences effect measured by $\hat{\rho}$ from Equation (1).

3.2 Effect on Judge Age Distribution

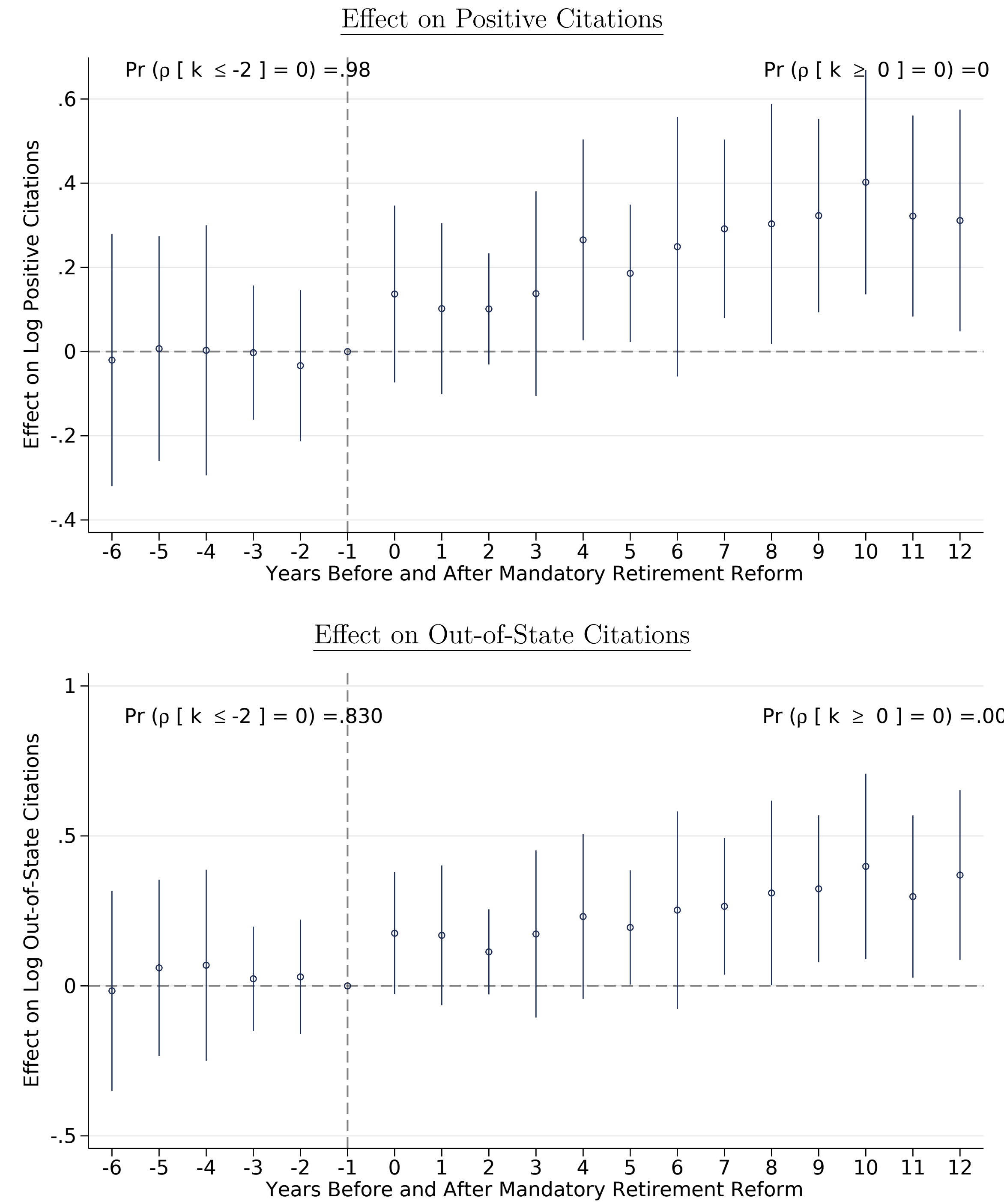
An initial question is whether mandatory retirement reforms have an effect on the age distribution in the courts. To add to the descriptive statistics presented earlier, we report the causal treatment effect using the reforms. This can be seen as a "first stage" to see if the "instrument" (mandatory retirement) is affecting the relevant "treatment variable" (judge age).

Regression estimates for Equation (1) with age statistics as the outcome are reported in Table 3. Relative to the pre-reform trend, judges after the reform are about 3 years younger (Columns 1 and 2). Columns 3 through 7 look at the effect on distributional statistics using court-year-level aggregates. We see that the age effect is present across the age distribution. Both the youngest and oldest ends of the distribution are shifting down. This is consistent with the replacement of the oldest judges with overall younger judges. Presumably, this result is what the reforms were meant to achieve.

3.3 Effect on Court Performance

Figure 5 shows event-study estimates (Equation 2) with log citations per judge per year as the outcome. We use six years before the reform, up until twelve years after, as the event window. In the top panel, the outcome includes all positive citations. In the bottom panel, the citation count is limited to citations from other states. Out-of-state citations may better reflect quality because they proxy for the persuasive influence of an opinion as rulings are not binding in other states (Choi et al., 2010). For both outcomes,

Figure 5: Event-Study Effect of Retirement Reform on Court Performance



Notes. Court performance before and after reforms implementing retirement ages of 70, 72 or 75. Top panel outcome is log positive citations for a judge in a year; bottom panel is only citations from courts in other states. Time series is a coefficient plot from the event study regression (2), with coefficents estimated relative to the year before the reform. Regression includes court and year fixed effects and court-specific event windows. 95% confidence intervals constructed with standard errors clustered by state.

Table 3: First Stage Effect of Mandatory Retirement Reform on Judge Age

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
					Court-Level Distributional Statistics		
	Judge Age		Age (Min)	Age (Q25)	Age (Median)	Age (Q75)	Age (Max)
Ret. Reform	-3.279**	-3.072**	-2.881+	-3.323*	-4.601**	-4.970**	-4.734**
	(0.717)	(0.848)	(1.461)	(1.615)	(1.145)	(1.410)	(1.753)
Year FE, Court FE	X	X	X	X	X	X	X
Windows/Trends		X	X	X	X	X	X
N	14989	14989	2357	2357	2357	2357	2357
R-sq	0.078	0.121	0.597	0.577	0.624	0.644	0.643

Notes. DD effect of mandatory retirement reform on judge age statistics in ten years after reform, relative to ten years before reform. Observation is a judge working in a year. "Ret. Reform" is a treatment indicator for the ten years after the introduction of mandatory retirement. In Columns 3 through 7, dependent variables are computed at the court-year level. In particular, "Age (Min)" the minimum age, "Age (Q25)" the age at the 25th percentile, "Age (Median)" the median age, "Age (Q75)" the age at the 75th percentile and "Age (Max)" the maximum age. Court Treat Windows means court-specific treatment windows (ten years before and after reform). Standard errors clustered by state in parentheses. + p<.0.1, * p<0.05, ** p<0.01.

we can see a clear break and increase after treatment. The effect increases over a number of years, peaking at about a 40 percent increase (relative to counterfactual) ten years after the reform. Meanwhile, the precise zeros before the reform support our identification assumption of parallel trends. According to a joint significance test, the pre-period coefficients are jointly not significant ($p > .8$), while the post-period coefficients are jointly significant ($p < .001$).

The results for the differences-in-differences regressions for Equation (1) are reported in Table 4. Across a range of specifications, there is a positive and significant effect of introducing a mandatory retirement age on a court's influence, as measured by citations. The result is consistently significant when adding controls for initial court rules, case types, and the age distribution, interacted with year. Appendix Figure B.1 shows the event study when adding these interacted covariates. Appendix Table B.11 shows identical results using an inverse hyperbolic sine rather than log transformation.

We undertook an array of checks to assess the robustness of the effect of the retirement reform on court performance. First, Table 5 shows that the effect holds for a number of alternative quality measures besides positive citations. The result is robust to using levels rather than logs (Columns 1 and 2), and note that the estimated effect in levels (60 positive forward citations) is about 24% of the mean in the ten years

Table 4: Effect of Mandatory Retirement Reform on Log Citations

	(1)	(2)	(3)	(4)	(5)
	Effect on Log Positive Cites per Judge-Year				
Retirement Reform	0.228**	0.253**	0.237**	0.322**	0.302**
	(0.0756)	(0.0836)	(0.0818)	(0.0899)	(0.0928)
Court FE, Year FE	X	X	X	X	X
Court Trends/Windows		X	X	X	X
Init Court Rules × Year FE			X	X	X
Init Case Types × Year FE				X	X
Init Age × Year FE					X
N	15010	15010	15010	15010	15010
R-sq	0.460	0.526	0.538	0.555	0.564

Notes. DD effect of mandatory retirement reform on log positive citations to a judge's opinions in ten years after reform, relative to ten years before reform. Observation is a judge working in a year. "Ret. Reform" is a treatment indicator for the ten years after the introduction of mandatory retirement. Court Treat Windows means court-specific treatment windows (ten years before and after reform). "Init X" × year FE means initial values are interacted with year. "Init Court Rules" includes a state's 1947 rules for judge selection/retention system, admin office, intermediate appellate court, number of judges, and term length. "Init Case Types" includes a court's 1947 average values for case characteristics (legal area and related industries). "Init Age" includes the initial mean and standard deviation for judge age on the court. Standard errors clustered by state in parentheses. + p<.0.1, * p<0.05, ** p<0.01.

Table 5: Effect of Mandatory Retirement Reform, Other Quality Measures

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Cites in Levels		Within 10 years		All Cites		Discuss Cites		Out-of-State Cites	
Ret. Reform	57.51*	60.50*	0.333**	0.338**	0.224**	0.249**	0.183**	0.175*	0.154	0.208*
	(23.27)	(25.41)	(0.0921)	(0.102)	(0.0715)	(0.0806)	(0.0569)	(0.0654)	(0.0920)	(0.0846)
Year / Court FE	X	X	X	X	X	X	X	X	X	X
Trends/Windows		X		X		X		X		X
N	15010	15010	15010	15010	15010	15010	15010	15010	15010	15010
R-sq	0.351	0.415	0.537	0.613	0.470	0.530	0.463	0.523	0.471	0.520

Notes. Observation is a judge working in a year. "Retirement Reform" is an indicator for the ten years after the introduction of mandatory retirement. "Cites in Levels" means the outcome is not logged. "Within 10 years" is the log positive cites within ten years of an opinion. "All Cites" is the log number of all citations (positive, negative, and distinguishing) to a judge in a year. "Discuss Cites" is only the positive cites where the latter judge discussed the cited opinion. "Out-of-State Cites" is the count of number of positive citations from courts in other states. "Positive Cites" is the number of positive cites (in levels). Court Treat Windows means court-specific treatment windows (ten years before and after reform). Standard errors clustered by state in parentheses. + p<.0.1, * p<0.05, ** p<0.01.

before the reforms (245 positive forward citations), similar to the effect in log points from above. The effect is robust when holding the number of years of forward citing cases constant to ten years of an opinion (Columns 3 and 4). We report effects on a more inclusive measure of quality (all cites, not just positive) in Columns 5 and 6. We use a more restrictive measure (discussion cites, where the previous case was specifically discussed and applied) in Columns 7 and 8. Finally, we report DD estimates for out-of-state citations (Columns 9 and 10). We see positive effects for all of these alternative measures. The respective event study estimates are reported in Appendix Figure B.2.[23]

The appendix reports robustness checks along a number of additional margins. Appendix Figure B.4 shows robustness to only using treated states in the sample. Appendix Figure B.5 adds more pre-periods and shows there is no sign of a pre-trend even ten years before the reform. Appendix Figure B.6 shows robustness to dropping each treated state individually.[24] Appendix Figure B.3 shows identical results when collapsing to the court-year level, rather than using judge-year-level data. Appendix Table B.8 shows that a qualitatively similar effect on judge performance is observed when each of the specified maximum judge ages (70, 72, or 75) are analyzed separately.

Our setting features a staggered treatment, where different treated states adopt the reforms at different times. As noted in Goodman-Bacon (2018), in this setting the standard two-way fixed-effects estimator can be biased due to the presence of previously treated states in the comparison group. To address this issue, we follow the method in Callaway and Sant'Anna (2019) and implement the following alternative specification. For each treated state, we estimate the event study (2) dropping all other treated states, and including in the comparison group only the set of 18 states who had not adopted mandatory retirement before the end of our time period. We then take the average of these estimates and plot as a normal event study. Appendix Figure B.8 reports these adjusted event study estimates, which again show a poitive effect that is

[23]Appendix Table B.1 reports effects on even more measures. Columns 1 and 2 show that there is still an effect for forward citations more than ten years after an opinion is published; presumably, judges have less professional influence on these long-term citations. In Columns 3 and 4, we there that there is no effect on the proportion of cases by a judge that has at least one positive citation; this is perhaps unsurprising since an average of 87% of authored cases have at least one cite. Third, Columns 5 and 6 show that the effect is robust and similar when dropping the top 5% of cases by number of cites in each court-year; this means that our effects are not driven by the production of more "blockbuster" cases.

[24]Results are also robust if dropping all states that have an electoral reform within five years of the retirement reform.

similar to the results of the baseline model.

Next, in Appendix Table B.7 we add additional time-varying controls to X_{ist} , which are probably "bad controls" or "colliders" in the sense that they could be affected by the retirement reforms. We include controls for the case characteristics, time-varying court rules (election system, number of judges, and government expenditures on the judiciary), and fixed effects for the number of years a judge has been on the court. Finally, we include the lagged dependent variable, which can perform better in panel data models with persistent shocks (Gentzkow et al., 2011; Caughey and Warshaw, 2018).

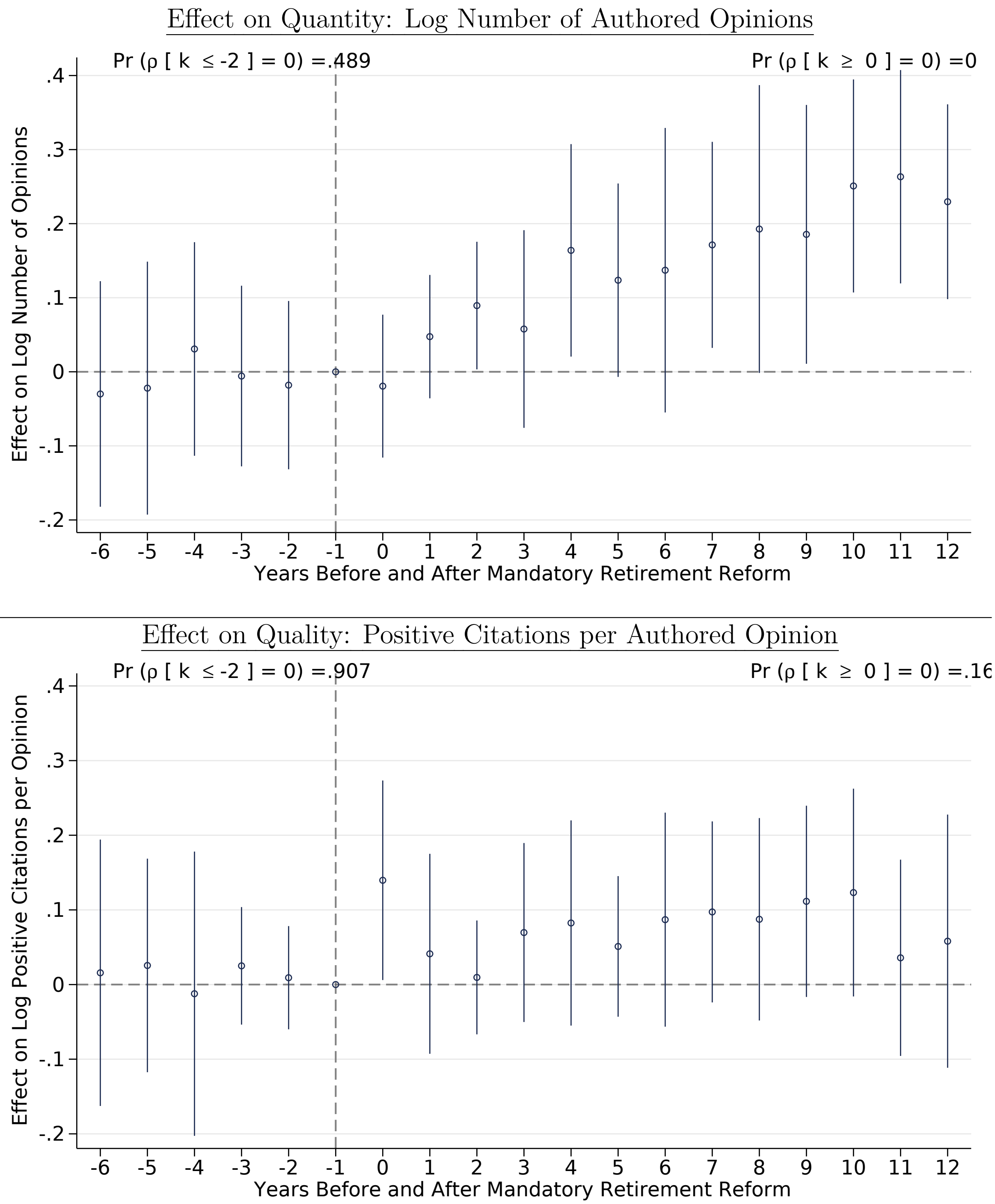
3.4 Is it Quality or Quantity?

As discussed, the main outcome, total citations to a judge or court per year, is a combined measure of quantity and quality. Posner (1995) suggests that older judges often maintain high quality levels by reducing quantity in terms of the number of opinions they write. Ash and MacLeod (2015) provide evidence that these are separate choices in a judge's work and that judges tend to care more about quality at the margin. A clearly important question, then, is whether the effect on total citations is driven by total number of opinions, or citations per opinion.

We provide some evidence in this direction by running our analysis with these separated quantity and quality measures. Figure 6 shows the event study estimates. We can see a clear positive effect of the reform on both outcomes.

Table 6 provides the differences-in-differences estimates. We see that there is an increase in the number of opinions (Columns 1-3), reflecting an increase in quantity. In addition, the point estimates on the effect of cites per opinion are positive (Columns 4-6), even when limiting to out-of-state cites (Columns 7-9). Though, the effect on citations per case is less precisely estimated than the effect on quantity. Thus, given that the effect on mandatory retirement on citations per case likely is positive, while the effect on quantity is a precisely estimated positive effect, we can conclude that mandatory retirement has a positive effect on state supreme court performance in the United States.

Figure 6: Reform Effect on Quantity and Quality



Notes. Court performance effects before and after reforms implementing retirement ages of 70, 72 or 75. Top panel outcome is log number of opinions by a judge in a year; bottom panel is average log positive citations per opinon. Time series is a coefficient plot from the event study regression (2), with coefficents estimated relative to the year before the reform. Regression includes court and year fixed effects and court-specific event windows. 95% confidence intervals constructed with standard errors clustered by state.

Table 6: Reform Effect on Quantity and Quality

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	# of Opinions			Cites per Case			Out-State Cites / Case		
Retirement Reform	0.169**	0.142**	0.184**	0.0368	0.0751	0.0856+	-0.00514	0.0348	0.0500
	(0.0415)	(0.0467)	(0.0580)	(0.0456)	(0.0497)	(0.0461)	(0.0468)	(0.0404)	(0.0463)
Year FE, Court FE	X	X	X	X	X	X	X	X	X
Court Trends/Windows		X	X		X	X		X	X
Init Covars × Year FE			X			X			X
N	15010	15010	15010	15010	15010	15010	15010	15010	15010
R-sq	0.326	0.510	0.555	0.649	0.710	0.736	0.629	0.668	0.691

Observation is a judge working in a year. "Retirement Reform" is an indicator for the ten years after the introduction of mandatory retirement. "# of Opinions" is the number of majority opinions written by a judge in a year. "Cites per Case" is number of citations per published opinion. "Out-of-State Cites / Case" is number of out-of-state citations per published opinion. Court Treat Windows means court-specific treatment windows (ten years before and after reform). Init Covars × Year FE includes initial rule, case type, and age statistics of a court interacted with year. Standard errors clustered by state in parentheses. + p<.0.1, * p<0.05, ** p<0.01.

Table 7: Effect of Retirement Reform on Case Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Effect on Log Positive Cites					Share	Case
	Random Assigment		Discretionary Assigment		Case Controls	Crim.Cases	Importance
Retirement Reform	0.239+	0.300*	0.263**	0.216**	0.148*	0.0301*	0.0350
	(0.125)	(0.134)	(0.0845)	(0.0608)	(0.0703)	(0.0133)	(0.0300)
Year FE, Court FE	X	X	X	X	X	X	X
Court Trends / Windows		X		X	X	X	X
N	10861	10861	4149	4149	15010	15010	15010
R-sq	0.466	0.532	0.419	0.476	0.652	0.649	0.472

Observation is a judge working in a year. "Retirement Reform" is an indicator for the ten years after the introduction of mandatory retirement. Court Treat Windows means court-specific treatment windows (ten years before and after reform). "Share Criinal Cases" is the share on criminal law. "Case Importance" is the predicted citations to a case based on case characteristics (legal area and related industries). Standard errors clustered by state in parentheses. + p<.0.1, * p<0.05, ** p<0.01.

3.5 Relevance of Case Characteristics

As mentioned above, judicial citations are a combined result of judge input and the importance of a case. We are therefore concerned with whether these effects are driven in part by the types of cases. Table 7 reports some regression results along these lines. First, in Columns 1 through 4 we show that the effect of the reform on quality is seen both in states with random case assignment and in states with discretionary case assignment. Appendix Figure B.9 shows the event study effects separately by case assignment rule, which are similar. Column 5 shows the main results is robust to including controls for case type.

To further unpack this dimension of the effect, we look at how the types of cases that the judges review changed after the reform. In Table 7 Column 6, we see that the share of criminal cases increased after the reform, reflecting a change in the composition of the caseload that judges review. However, in Column 7 we do not see a significant change in the importance of opinions, as indicated from the case characteristics . Therefore it appears that the portfolio of cases that the court accepts for appeal is not a major factor in the increase in court citations.

3.6 Additional Results

We look at additional outcomes in Appendix Table B.12, and report effects on work output (log number of words written), case-law research, and the rate of being overruled by the U.S. Supreme Court. There are no effects on these additional dimensions. There is an increase in the number of addendum opinions (log of the count of dissents and concurrences), but that is proportional to the total number of opinions produced by the court.

Appendix Table B.10 provides some estimates of the effects on related institutional factors. Judge experience (years on the court) decreases (mechanically) after the reforms, but not quite as much as by judge age. We see no immediate effect (comparing the ten years before and after a reform) on how a judge's tenure ended (whether by retirement, death in office, etc).

4 Aging and Judge Performance

The evidence thus far shows that the mandatory retirement reform results in increased output and citations per judge. Conceptually, the experiment is to compare average court performance before and after the change in the mandatory retirement rule. The next question is to understand why this is occurring. As discussed in Section 2.1, there is plenty of previous evidence that age is negatively correlated with measures of of an individual's cognitive and physical performance. The difficulty is that age is not a treatment in the sense used by the potential outcomes framework. Holland (1986) observes that individual characteristics such as age, sex, or race cannot be experimentally varied, and hence one cannot properly discuss a causal effect of age.

One way to attack this problem is to compare judges of different ages in the same court-year. The benefit of this approach is that judges are being evaluated using exactly the same criteria – the number of cases and future citations for a specific setting and period. We can take advantage of the panel structure and control for heterogeneity in a judge's base-line performance. We can then ask if the relative ranking of the judge's performance changes with age.

4.1 Empirical Approach

This approach is implemented with the following specification. Assume a quadratic age model of performance variable y_{ist} for judge i working in court s at year t:

$$y_{ist} = \alpha_{st} + \alpha_i^0 + \gamma_1 A_{ist} + \gamma_2 A_{ist}^2 + X'_{ist}\beta + \epsilon_{ist} \quad (3)$$

where A_{ist} is the age (in years) for judge i in court s at t, and α_i^0 is a judge-specific intercept for performance in their first year on the court. This provides a judge-specific baseline value such that our estimates for the effect of age, $\hat{\gamma}_1$, are relative to the individual's baseline. The main source of bias comes from the time-varying changes in the court work environment which are systematically correlated with age. Thus, we include a full set of court-year fixed effects. Therefore any estimated coefficients are also relative to the court average in each year. The regressions effectively compare judges sitting on the same court, working at the same time, but who are of different ages. Again, standard errors are clustered by state.

X_{ist} includes a number of additional items which we add in follow-up specifications.

First, we have cohort fixed effects – indicators for each decade that the judge started on the court. This covariate is meant to rule out mechanical variation due to cohort differences across the time period. In the same vein, we have court-specific linear trends in judge starting cohort: formally, judge starting-year interacted with court fixed effect. This allows for judges in different states to have a different confounding trend in starting year and performance.

In the baseline specification, we specify y_{ist} in logs (as above) and the age variable is in levels. This means that the coefficients can be interpreted as the proportional change in performance due to a one-year increase in judge age. In follow-up results, we report results with y_{ist} in levels and in rank percentiles by court-year. That is, the judge with the highest measure in a year is given a 1, the lowest a zero, and all other judges uniformly distributed on that interval according to rank. In this specification, the level differences do not matter, and the measure is more robust to outliers (Chetty et al., 2014). The interpretation of coefficients is similar to median (quantile) regression.

In addition to the quadratic model regressions, we plot out the dynamic changes in performance over the lifespan by estimating

$$y_{ist} = \alpha_{st} + \alpha_i^0 + \sum_{g \in G} \gamma_g A_{ist}^g + X_{ist}' \beta + \epsilon_{ist} \tag{4}$$

where now instead of a quadratic specification for age, we have a vector of indicator variables A_{ist}^g, $g \in G$, equaling one when judge i is in age group g. The age groups, chosen for convenience, are 0-44, 45-49, 50-54, 55-59, 60-64, 65-60, 61-64, 65-69, 70-74, and 75+. Appendix Figure C.1 shows the distribution in our data for these age groups. The qualitative results of the dynamic age analysis are not sensitive to how these groups are defined.

For the dynamic analysis we drop the first and last years of each judge career. We drop these observations because in the first and last years, judges have a partial caseload by construction. In the quadratic age model, this sample adjustment does not make a big difference in estimates. But it can shift estimates around a lot for younger and older judges when estimating the age group effects.

It is important to note why we do not report a specification with judge fixed effects. Age (in years) is perfectly linear within judge. Therefore, γ_1 and γ_2 are not identified when including both judge fixed effects and year (or court-year) fixed effects. There is no straight-forward estimation approach that would allow judge fixed effects and also account for the large global variation in citations over time across all courts. Further, with judge fixed effects the age and experience effect cannot be distinguished.

Since judge fixed effects are not an option, a potential issue with the previous specifications is that they are applied to an unbalanced sample of judges. Judges start and end at different ages, so the estimated effects could be driven by selection of different types of judges into different starting and ending ages. To address this issue, we produce a set of regressions using balanced samples of judges. First, we take overlapping ten-year age groups across the lifespan (45-54, 50-59, 55-64, 60-69, 65-74) and limit to judges that worked continuously in that age group. The distribution over these groups is shown in Appendix Figure C.2. Next, we produce estimates from Equation 3 restricting to those balanced samples. This is similar to using judge fixed effects as we are keeping the set of judges constant. On the other hand, the judges come from many different courts and years, and there could be confounding variation in citations across courts over time. Therefore for the balanced-sample analysis, the preferred outcome is the rank percentile in citations. The rank percentile provides a performance measure that is comparable across courts and over time.

4.2 Main Results on Judge Age and Judge Performance

Now we report the main results for variation of judge performance by age. The regression estimates from Equation (3) are reported in Table 8.[25] There is a highly significant negative relationship between age and judge performance as measured by positive citations to a judge's opinions in a year. The estimate doesn't change that much when adding the judge's first-year as a baseline to the regression (Column 2), or when adding cohort fixed effects and state-specific cohort trends (Column 3). As seen in Column 5, the effect also holds using a rank percentile specification in work performance.

Columns 4 and 6 assume a quadratic model for age. Both of these columns show that the negative linear estimates from the other columns conceal a concave relationship, where the linear term is actually positive. The quadratic term is negative and significant, indicating that the negative relationship between age and performance accelerates at later ages. Taking the quadratic model at face value, the coefficient estimates indicate that judge performance is maximized around age 47 (Column 6) to 48 (Column 4). Given that the median age of working judges is 61, these estimates

[25] See Appendix Table C.1 for identical results using inverse hyperbolic sine of cites, rather than log of cites.

Table 8: Judge Age and Judge Performance

	(1)	(2)	(3)	(4)	(5)	(6)
		Log Positive Cites			Rank Percentile Cites	
Judge Age (Years)	-0.00822**	-0.00812**	-0.00712**	0.0322*	-0.00468**	0.0166+
	(0.00134)	(0.00107)	(0.00121)	(0.0126)	(0.000826)	(0.00894)
Age Squared				-0.000331**		-0.000179*
				(0.000111)		(0.0000788)
Court-Year FE	X	X	X	X	X	X
First-Year Baseline		X	X	X	X	X
Cohort FE / Trends			X	X	X	X
N	14981	14981	14981	14981	14972	14972
R-sq	0.670	0.695	0.700	0.701	0.109	0.112

Observation is a judge working in a year. "Log Positive Cites" is log of positive cites to a judge in a year. "Rank Percentile Cites" means judges are uniformly distributed between zero and one based on number of positive citations within court-year (0 is lowest, 1 is highest). Court-Year FE is interacted court-year fixed effects. First-Year Baseline means a judge's value for the outcome in their first year on the court is included as a control. Cohort FE means fixed effect for decade that the judge started on the court. Cohort Trends means judge starting-year interacted with court fixed effect. Standard errors clustered by state in parentheses. + p<.0.1, * p<0.05, ** p<0.01.

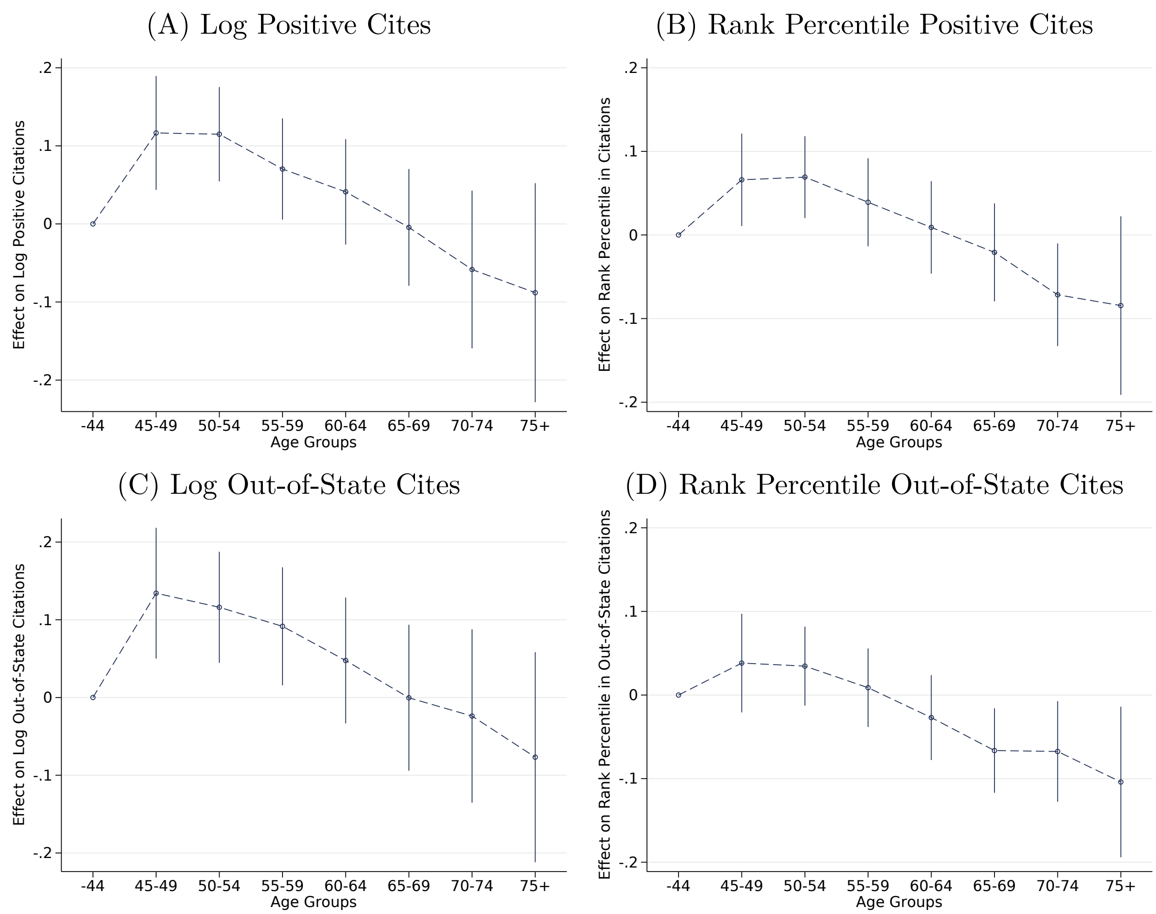
suggest a negative age-performance relationship for almost all judges.

To show further the dynamics of the effect, Figure 7 reports the age-group estimates from Equation (4). Each item in the graphs shows the difference in citations between the respective five-year age group and the baseline group of judges who are 44 or younger. These regressions are consistent with the quadratic model, showing an initial increase in performance and then a steady decrease starting in the late 40s or early 50s. This trend holds for both positive citations and out-of-state citations, and for the log outcome as well as the within-court-year percentile rank.

As seen in Figure 8, the age-performance effect holds for alternative quality measures: cites within ten years of an opinion, all cites (not just positive), discussion cites, and out-of-state cites. Appendix Tables C.2 and C.4 report the associated regression results. For each of these variables, the relationship is concave in the quadratic specification.[26]

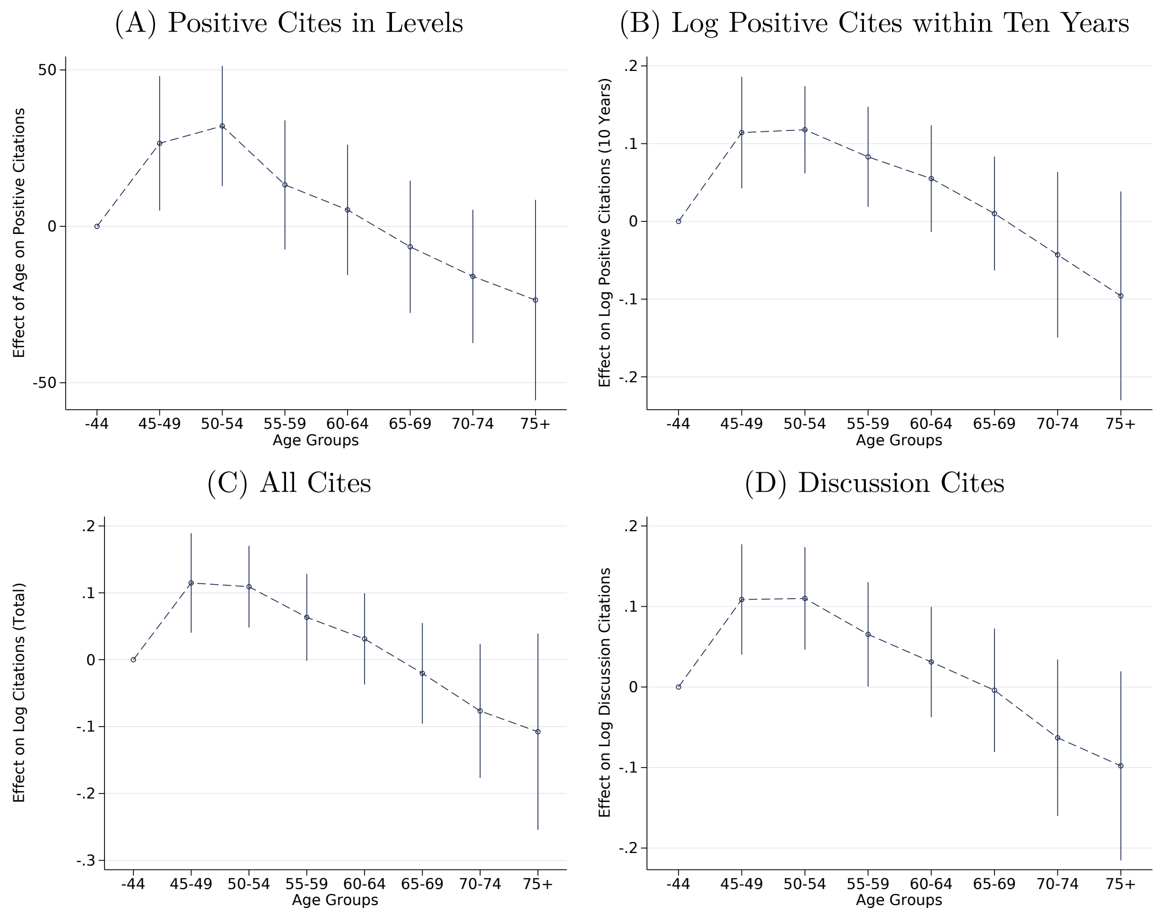
[26] Appendix Table C.3 reports the life-cycle estimates for three additional measures. Columns 1 and 2 show that there is a similar aging effect for forward citations more than ten years after an opinion is published; presumably, judges have less professional influence on these long-term citations. Columns 3 and 4 show that there is a similar effect on the proportion of cases by a judge that has at least

Figure 7: Dynamic Analysis of Judge Age and Judge Performance



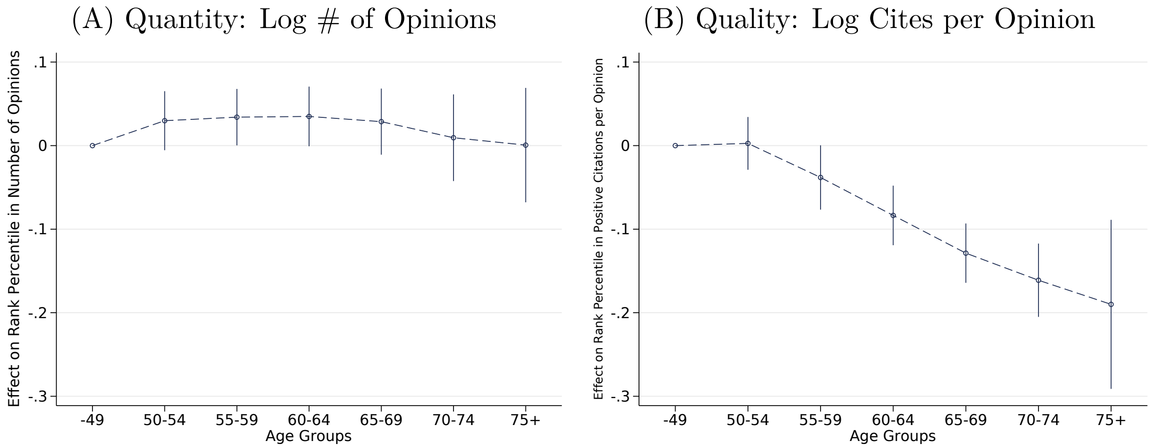
Dynamic coefficient plots for estimates of five-year age group differences, relative to the age $<$ 45 group. Observation is a judge working in a year. All graphs contain court-year interacted fixed effects, first year baselines, and cohort fixed effects. Outcomes are in logs or rank percentiles, as indicated. 95% confidence intervals constructed using standard errors clustered by state.

Figure 8: Dynamic Analysis: Additional Performance Measures



Dynamic coefficient plots for estimates of five-year age group differences, relative to the age < 45 group. Observation is a judge working in a year. All graphs contain court-year interacted fixed effects, first year baselines, and cohort fixed effects. Outcomes are as indicated. 95% confidence intervals constructed using standard errors clustered by state.

Figure 9: Dynamic Analysis: Quantity vs. Quality



Dynamic coefficient plots for estimates of five-year age group differences, relative to the age < 45 group. Observation is a judge working in a year. All graphs contain court-year interacted fixed effects, first year baselines, and cohort fixed effects. 95% confidence intervals constructed using standard errors clustered by state.

Figure 9 breaks out the age effect by quantity (Panel A) and quality (Panel B). As can be seen in the figure, there is not much of an age trend for the caseload, but a large and significant negative effect on citations per opinion. These estimates suggest that the aging effect on performance is driven by changes in work quality, rather than by changes in work quantity. At a minimum, we can rule out older judges compensating lower quality with higher quantity.

Next we show that aging effects are not driven by changes in the types of cases that older judges work on. As shown in Appendix Table C.5 and Appendix Figure C.3, case categories and case importance are not significantly related to judge age. This null is seen under both random/rotating assignment of judges to cases and under discretionary assignment. Overall, these statistics suggest that observed differences in work quality across the lifespan are not mainly driven by changes in the types of cases that older judges rule on.

The effect of age on performance is robust to a variety of alternative specifications. Appendix Figure C.4 shows that the results look very similar without any fixed effects (Panel A), or instead adding experience fixed effects (Panel B). We include results

one positive citation, suggesting the older judges work on more cases that have no impact. Third, Columns 5 and 6 show that the effect is robust and similar when dropping the top 5% of cases by number of cites in each court-year; this means that the variation over the lifespan is not driven by the production of more or fewer "blockbuster" cases.

separately by mandatory and voluntary retirement (Appendix Table C.6); they are similar. The results are robust to alternative weighting (Appendix Figure C.6) and alternative clustering of standard errors (Appendix Figure C.7).

We also separate our results before/after 1970. Appendix Figure C.5 shows that the age trend is shallower in the latter period. Judges from more recent cohorts show decreasing performance starting at later ages, and it remains high until after the age of 70. In the earlier period, performance dips significantly starting at age 65. This difference could be due to significant health differences between cohorts of judges in the early and late period. Another possibility is that support technologies, such as higher-skill clerks or legal research databases, have made judges more productive later in life.[27]

We analyze a number of alternative outcomes (see Appendix Table C.7). Older judges tend to write the same amount in terms of total words written, have cases that are similar in length, and do about the same amount of caselaw research. These null effects are inconsistent with an alternative "verbosity" explanation of the citations effect. That is, citations do not reflecting variation in quality due to aging, but rather that younger judges are more verbose in research and explaining the previous caselaw, and therefore get more cites due to explaining rules more. These null estimates are also inconsistent with a "clerks" explanation for age-performance variation – that older judges rely more or less on clerks than younger judges. If that were the case, we would expect systematic variation in the tasks that clerks are most responsible for – writing and researching.

Appendix Table C.7 also shows that older judges affirm (rather than reverse) lower-court cases at about the same rate as younger judges. This is important for our interpretation of the results because it suggests that older judges are not systematically more accepting about the legal status quo. If our result on cites were driven by older judges being less innovative, we would expect to see that in fewer reversals of lower courts. We see no difference in reversal rate by age. [28]

For completeness, Appendix Table C.8 reports a two-stage least-squares (2SLS) specification where we instrument for judge age using the mandatory retirement treat-

[27]As additional heterogeneity analysis, we compare judges by gender, and by the judge appointment system. We find that aging effects are the same across genders (unreported).

[28]We also see in Appendix Table C.7 that older judges are overruled by the U.S. Supreme Court at about the same rate, tend to dissent less often, and are superseded more often by statute. The latter is a rare outcome.

ment indicators. This specification relies on unrealistic exogeneity assumptions, as the reforms affect performance through other channels besides age. That said, we get a strong first stage and estimate a significantly negative 2SLS effect of age on performance. The 2SLS coefficient is about ten times as large as the OLS coefficient, bringing the estimates more in line with the effect of mandatory retirement on performance measured above.

4.3 Analysis of Balanced Judge Samples

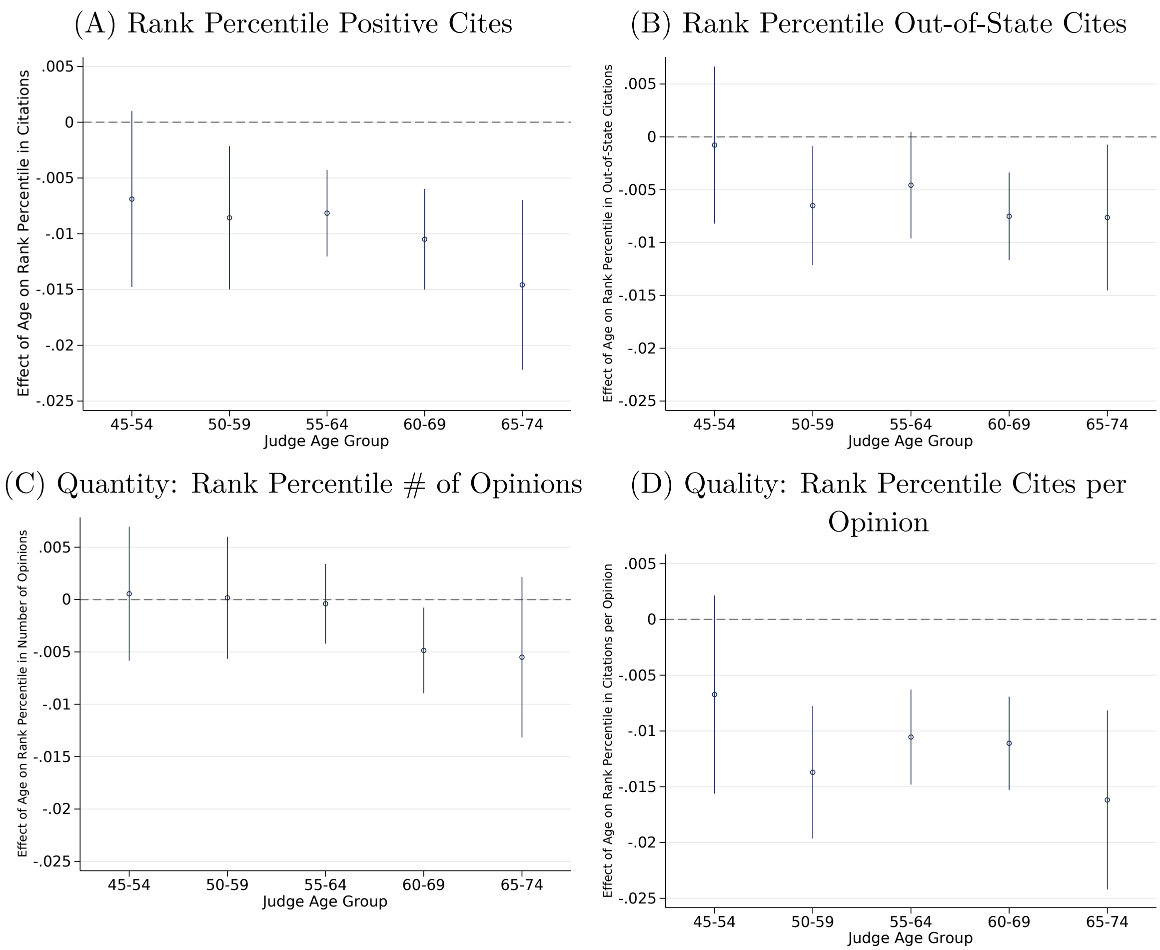
As discussed in Section 4.1 above, an issue with the previous results is that they are made on unbalanced samples of judges. Judges are entering and leaving the sample at different ages. It could very well be that the negative effect of aging on performance is driven by this attrition bias. To address this issue, we construct balanced samples of judges by age group across the lifespan.

The regression results using balanced samples are visualized by the coefficient plots in Figure 10. Each of these coefficient plots reports the estimate for $\hat{\gamma}$ (with 95% confidence interval) from Equation 3 (the linear model, without the quadratic age term) using a sample of judges that work continuously for the ten years of the age group indicated in the horizontal axis. The plots go from left to right starting from 45-54, to 50-59, and so on, up until 65-74. Because the judges are working continuously, the estimates give within-judge changes over time and are not biased by selective entry or exit. Because the outcomes are within-court-year rank percentiles, no fixed effects are needed and the performance measures are comparable across courts and years.

We can see in Panel A that there is a negative effect of age on performance across the samples. The effect in the earliest cohort is not significant. As we move to later samples, the effect becomes more negative and more statistically significant. This is consistent with an increasingly negative effect of aging on performance later in the lifespan. These baseline results are robust to alternative weighting and alternative clustering of standard errors (Appendix Figure C.8). In Panel B, we see a similar trend for out-of-state cites.

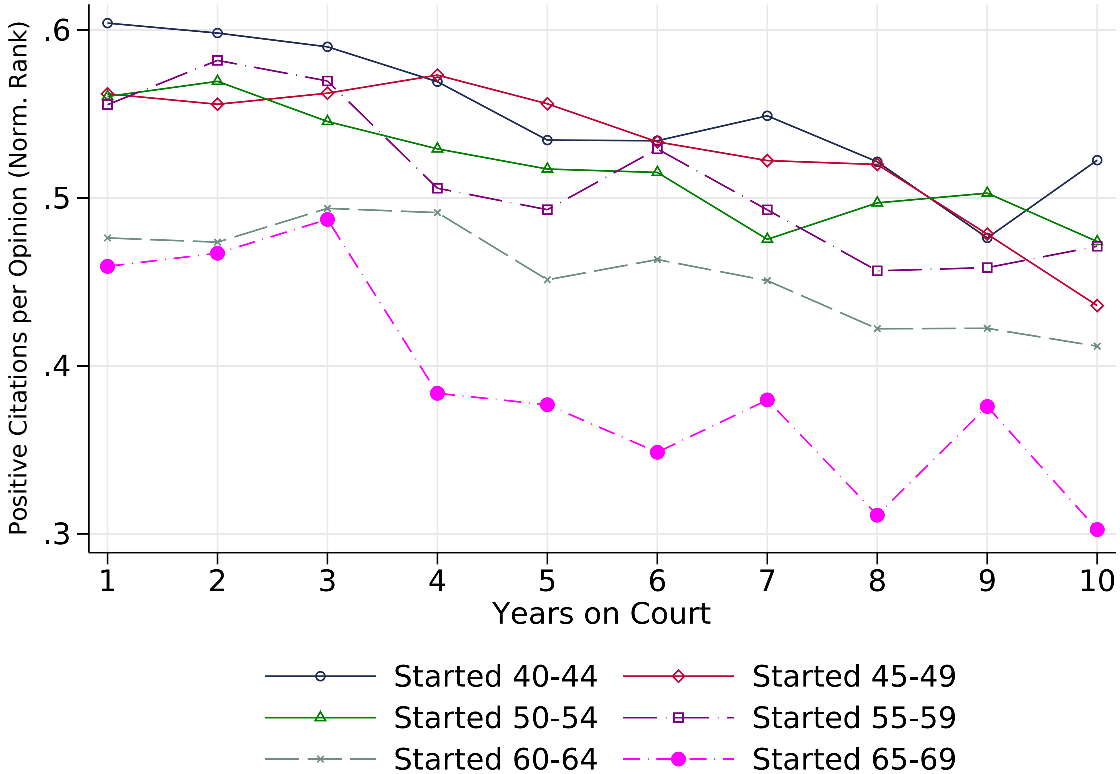
In Panel C and Panel D, we look at quantity (number of opinions) and quality (citations per opinion) respectively, in the balanced samples of judges. We see mostly insignificant effects on the workload (number of opinions). Meanwhile, there is a large negative effect for quality (citations per opinion) starting quite early: in the 50-59 sample. These results provide additional support for the view that the age-performance

Figure 10: Effect of Age on Performance in Balanced Judge Samples



Performance-Age estimates for separate balanced samples of judges based on age group. Observation is a judge working in a year. Outcomes are in rank percentiles, regressed on age (in years) for the specified group. 95% confidence intervals constructed using standard errors clustered by state.

Figure 11: Effect of Age on Performance in First Years of Judgeship



Notes. Time series for average rank percentile (within court year) in positive citations for the first ten years of a judge career, separately by starting age (indicated in legend).

effect is driven by work quality (rather than quantity).

4.4 Experience and Attrition

There are two final issues that are relevant to the interpretation of an aging effect on judge performance. First, older judges mechanically tend to have more judicial experience than younger judges. We would like to disentangle these effects by holding experience constant. Second, we only observe the performance of judges who are on the court – we do not observe performance in other jobs. We analyze variation in performance according to entry and exit timing, to understand better what direction selective attrition is influencing our estimates.

First, to get at the experience issue, we look at how judge performance evolves over

Table 9: Effects of Starting Age and Ending Age

	(1)	(2)	(3)	(4)	(5)
	Effect on Log Positive Cites				
Judge Start Age	-0.00700**	-0.000885			-0.00147
	(0.00155)	(0.00232)			(0.00222)
Judge End Age			-0.000612	0.0151**	0.0152**
			(0.00132)	(0.00177)	(0.00181)
Court-Year FE	X	X	X	X	X
Age FE		X		X	X
N	13655	13643	14618	13643	13643
adj. R-sq	0.672	0.678	0.668	0.682	0.682

Effect of judge start age and end age on judge performance. Observation is a judge working in a year. Standard errors clustered by state in parentheses. + p<.0.1, * p<0.05, ** p<0.01.

the first ten years on the court, but comparing judges by their starting age. We can see this variation in Figure 11. The judges in the different time series have the same amount of experience (years on the state supreme court), but they start at different ages. As before, we see a generally negative trend in quality over time. Moreover, we can see clearly that younger starting judges begin at a higher quality level than older starting judges. This difference is maintained over time. Therefore, increased age reduces work performance even holding experience constant.

To get at selective attrition, we hold age constant with age fixed effects, and then regress performance on starting age or ending age. Table 9 provides this additional analysis of how age and performance relate to career transition choices. First, we see that holding age fixed, the start age has no effect (Column 2). This result again supports the view that there is no important experience effect relative to the aging effect. Second, we look at the effect of judge's ending age, holding current age constant (Column 4). We see a significant positive effect. At a given age, judges who will end up working longer have relatively high performance in their job. This means that there is positive selection in terms of judges staying on the court. Lower-performing judges tend to leave the court at younger ages.

The reported life-cycle evidence must be interpreted with this selective exit result in mind. It suggests that all of our estimates for late-life performance decreases are, if anything, an underestimate. If our regressions included the judges who tend to leave at younger ages, the estimated performance declines would be even steeper.

4.5 The Role of Clerks

As mentioned above, state supreme court judges receive the help of clerks in research and writing their opinions. Therefore, the observed effects of age on performance could be driven in part by changes in how judges hire or manage their staff. Ideally, we could check whether there are systematic differences in the types of clerks that older or younger judges hire. But there are no datasets available on clerks in the state court system.

In the federal courts, however, there is a dataset on clerkships collected by Bonica et al. (2017). These authors have kindly shared their data on clerk law schools. As described in further detail in Appendix A.5, we link the clerkships to judge age using the Federal Judicial Center dataset. We measure clerk quality using Leiter's (2000) "quality of student body" ranking, which is constructed from LSAT scores and college GPA.

We then regress clerk quality on judge age using court-year fixed effects. Appendix Table A.5 shows that there is no relationship, positive or negative, between a judge's age and the quality of their clerks. With the caveat that this analysis is for federal courts, rather than state courts, it still provides suggestive evidence that clerks are probably not the main driver of observed variation in work performance over the judge life cycle.

5 Team Effects of Aging

Now we have analyzed the court-level effect of mandatory retirement (Section 3) and the judge-level life-cycle effect of aging (Section 4). To connect these analyses, let us use the life-cycle estimates to form a prediction of the effects on court performance due to the retirement reform's effect on age. Specifically, we can multiply the reform's effect on age (decrease by 3 years, Table 3) by the life-cycle effect of age on citations (about 1 percent decrease per year, Table 8). Together, these estimates suggest that the reform would increase cites by about 3 percent due to the aging effect. But what we see in the reform DD regression is a much larger effect: about 25 percent. The bias due to selective attrition (Section 4.4) could account for some of this difference, but given the especially large difference it seems that the reform is affecting performance through channels other than shifting the court's age composition.

The disconnect between the reform and the life-cycle estimates can also be seen in the relative importance of quantity and quality. The retirement reform effect is driven mostly by an increase in quantity (number of published opinions), with relatively small increases in measured work quality (citations per opinion). Meanwhile, the aging effect is driven mostly by quality (citations per opinion), with not much effect on quantity (number of authored opinions). Again, the effect on individual judge age does not seem to be the principal mechanism by which retirement reforms affect court performance.

To explore further these issues, we first revisit the retirement-reform regressions using judge fixed effects instead of court fixed effects. Formally, we estimate (1) including judge indicators α_i rather than state indicators α_s. These regressions will summarize the within-judge effect of the reforms, focusing on those judges that stay on the court and are not removed due to old age. If the retirement reform works primarily through composition, the results with judge fixed effects would be zero. If the estimates are non-zero, we can reject that composition is the primary driver of how mandatory retirement improves performance.

Table 10 shows the difference-in-difference estimates for the specification with judge fixed effects. There is a positive and significant effect, about half of the magnitude of the full effect measures from Table 4. As before, the estimates are robust to the addition of a number of exogenous state characteristics interacted with year fixed effects. Appendix Figure D.1 shows the corresponding event study estimates, which tell the same story. Appendix D reports additional robustness checks along the same lines as Section 3. Again suggesting that the within-judge effects are not as large as the court-wide effects, the results are generally not as robust as the court-level effects in terms of magnitude and statistical significance.

As before, we look at quantity and quality separately using judge fixed effects. Table 11 reports the results. We see, similar to the baseline results, that quantity goes up robustly, but not as much as the court-level effect in terms of magnitudes. The quality effects are positive, yet not as robust across specifications. These results suggest that after the older judges are removed, the remaining younger judges take on more work without a reduction in work quality.

More generally, these estimates suggest that the effect of the mandatory retirement

Table 10: Effect of Reform on Log Cites, Judge Fixed Effects

	(1)	(2)	(3)	(4)	(5)
	Effect on Log Cites per Judge-Year				
Retirement Reform	0.175*	0.173*	0.170*	0.221*	0.216*
	(0.0768)	(0.0726)	(0.0728)	(0.0872)	(0.0868)
Judge FE, Year FE	X	X	X	X	X
Court Trends/Windows		X	X	X	X
Init Court Rules × Year FE			X	X	X
Init Case Types × Year FE				X	X
Init Age × Year FE					X
N	14905	14905	14905	14905	14905
R-sq	0.675	0.683	0.691	0.700	0.707

Observation is a judge working in a year. "Retirement Reform" is an indicator for the ten years after the introduction of mandatory retirement. Court Treat Windows means court-specific treatment windows (ten years before and after reform). "Init X" × year FE means initial values are interacted with year. "Init Court Rules" includes a state's 1947 rules for judge selection/retention system, admin office, intermediate appellate court, number of judges, and term length. "Init Case Types" includes a court's 1947 average values for case characteristics (legal area and related industries). "Init Age" includes the initial mean and standard deviation for judge age on the court. Standard errors clustered by state in parentheses. + p<.0.1, * p<0.05, ** p<0.01.

Table 11: Effect of Reform on Quantity/Quality, Judge Fixed Effects

	(1)	(2)	(3)	(4)	(5)	(6)
	Quantity (# of Opinions)			Quality (Cites per Opinion)		
Retirement Reform	0.0863*	0.0869*	0.0931+	0.0594	0.0564	0.0948*
	(0.0369)	(0.0383)	(0.0502)	(0.0399)	(0.0380)	(0.0382)
Judge FE, Year FE	X	X	X	X	X	X
Court Trends/Windows		X	X		X	X
Init Court Rules × Year FE			X			X
Init Case Types × Year FE			X			X
Init Age × Year FE			X			X
N	14905	14905	14905	14905	14905	14905
R-sq	0.648	0.674	0.706	0.815	0.821	0.836

Observation is a judge working in a year. "Retirement Reform" is an indicator for the ten years after the introduction of mandatory retirement. Court Treat Windows means court-specific treatment windows (ten years before and after reform). "Init X" × year FE means initial values are interacted with year. "Init Court Rules" includes a state's 1947 rules for judge selection/retention system, admin office, intermediate appellate court, number of judges, and term length. "Init Case Types" includes a court's 1947 average values for case characteristics (legal area and related industries). "Init Age" includes the initial mean and standard deviation for judge age on the court. Standard errors clustered by state in parentheses. + p<.0.1, * p<0.05, ** p<0.01.

reform is driven not just by older judges being replaced. In addition to that, there is a positive effect on the younger judges who are working before and after the reform. Thus, there is not just an individual but a team effect of aging workers upon co-workers. The presence of older judges appears to reduce the productivity of younger judges.

There could be two reasons for this team effect. First, it could be that the presence of older judges imposes unobserved work burdens on younger judges. That is, to maintain a reasonable work product level, the younger judges have to help the older judges with their work. The reforms remove the older judges and remove this burden, so then the younger judges have more time to provide higher performance.

A second, perhaps subtler, explanation of this effect is an "equal workload" norm that, combined with the presence of relatively unproductive older judges, imposes a workload cap on younger judges. As we saw in the life cycle regressions, all judges are responsible for the same number of cases regardless of age. Thus, as a team, the court would not want to increase the caseload above what can be handled by its lowest-performing judges. In particular, before mandatory retirement a court has to maintain a relatively low caseload so that the oldest judges can keep up with the equal load.

Consistent with this latter equal-workload explanation, the effect of the reforms is driven mostly by an increase in quantity rather than quality. With the oldest judges removed, the younger judges coordinate on a higher workload. If the first explanation (younger judges helping older judges) were more important, we would expect a similar or larger effect on quality, consistent with the evidence in Ash and MacLeod (2015) that judges prefer to increase quality rather than quantity at the margin when time pressure is relieved.

We find additional supporting evidence for the equal-workload mechanism in the patterns of authorship rates before/after the reform. That is, we look at the court-level proportion of cases that are authored (an individual judge signs his/her name) rather than unsigned (per curiam) opinions of the court. Under an equal workload norm with variation in productivity due to age, the best-performing younger judges have excess work capacity and can expend that on unsigned opinions. After the reform and the removal of older judges, therefore, we might expect fewer unsigned opinions. In our data that is exactly what we find: The reform increases the proportion of cases that are authored (Appendix Table D.2 Columns 1 and 2).[29]

[29]Changes in authorship rate are not driving are main reform effects. The results are similar when controlling for authorship rates (Appendix Table D.2 Columns 3 through 8).

6 Conclusion

The goal of this paper has been to measure the effects of aging on judicial behavior and court performance. Given that judges have low-powered incentives that do not explicitly link pay to performance, these factors likely have a significant impact on judge behavior. We find that mandatory retirement rules increase the performance of courts as a whole. An important contributor to this effect is that physical aging is associated with a reduction in work quality over the lifespan. We also have evidence for a team effect of aging, where older judges reduce the performance of younger colleagues.

These effects could be explained by the cognitive impacts of physiological aging and the resulting interactions between judges. But there could be other mechanisms besides cognitive factors. Part of the variation over the lifetime could be changes in incentives, for example due to pensions vesting. A third possibility is that preferences, rather than ability, change over the lifetime. Older judges might have different priorities than younger judges, for example preferring to seek justice in particular cases rather than seeking to influence the decisions in future cases. Younger judges might have more career concerns and want to have a reputational impact. Finally, younger judges might be noisier in their decision-making, which could be error-prone but also more innovative, thereby leading to more cites. Future work could do more to disentangle these mechanisms, perhaps in the lower courts where career concerns and decision errors are easier to identify.

We found a different age-performance profile than the previous work on federal circuit judges (Posner, 1995) and Australia High Court judges (Smyth and Bhattacharya, 2003). There could be many reasons for this difference, starting with our much larger sample of judges. The U.S. federal courts (and the Australia High Court) are consistently professionalized and selective, while state supreme courts have much more variance in the types of individuals who become judges.

In the state supreme courts, our results remain policy-relevant. As recently as 2016, Pennsylvanians voted (by a 2 percent ballot initiative margin) to increase the mandatory retirement age for state supreme court judges from 70 to 75. It is likely that other state-level changes in the retirement age will be proposed in the coming years. New York State has had mandatory retirement at 70 since 1869, the same rule that persists in many states. It cannot be, given changes in health and other technologies, that the same maximum age could be optimal both now and in 1869.

In the case of federal judges, our results are especially relevant given that imposing mandatory retirement is unconstitutional (Posner, 1995).[30] While the use of senior status and other incentives can put pressure on older judges (Choi et al., 2013), it is still the case that many federal judges stay on the bench past their prime (e.g. Goldstein, 2011). Our evidence provides more support for proposed reforms (including a constitutional amendment) to mandate retirement for the oldest federal judges.

What does our evidence say about an optimal retirement age? In the case of judges, our estimates – along with some more structural assumptions – could be used to compute an optimal retirement age (Diermeier et al., 2005; Chetty, 2009). If replacing judges were cost-less, then we could impose mandatory retirement at the age where performance peaks and judges begin to decrease in decision quality. According to Figure 7, for example, this would be about 55 years old. Replacing judges is not cost-less, however. Such a structural approach would have to estimate such costs in order to deliver useful policy parameters.

In labor economics, this research highlights the usefulness of direct measures of employee performance. A major challenge is that it is difficult to evaluate employees over long periods of time. In particular, in the last century there have been transformative changes in the nature of work. Computers are much more important, and jobs are more complex and include "soft" factors such as the ability to manage employees (e.g. Autor et al., 2008). A second challenge is that even if one could measure these factors, using them in hiring and promotion decisions could have unexpected agency effects.

These results may be useful to policymakers seeking to design better retirement policies for judges and other high-skill jobs. When productivity decreases with age, mandatory retirement can increase productivity on average. In particular, the results are useful in an era where an aging workforce is resulting in large structural changes to the economy (Acemoglu and Restrepo, 2017). Our results show that age matters, and that it can effect team performance. Though we have used the mandatory retirement rule change to estimate these effects, it does not imply that mandatory retirement is necessarily the optimal rule. That depends on many factors, including the counterfactual occupations for both the potential retirees and new judges. It may be that an increased use of senior status to reduce the negative output effect of older judges on younger judges is appropriate.

Finally, the Age Discrimination Act of 1967 requires as a matter of law that work-

[30] Federal judges "shall hold their offices during good behavior" (U.S. Const. Art. III Sec. 1).

ers be evaluated by their performance, and not by their age. In this study we have connected the age of a judge to a performance measure that uses 18 or more years of observations after they have worked. But in order to have employment reflect performance, one would need a contemporaneous measure of performance. This is a problem that is faced for many occupations filled by highly skilled professionals. It is certainly a challenge for future work to build better contemporaneous evaluation measures, and to design systems that comply with the law, while ensuring the productive employment of older workers.

References

Abraham, K. G. and Farber, H. S. (1987). Job duration, seniority, and earnings. *American Economic Review*, 77(3):278–297.

Acemoglu, D. and Restrepo, P. (2017). Secular stagnation? the effect of aging on economic growth in the age of automation. *NBER*.

Agarwal, S., Driscoll, J. C., Gabaix, X., and Laibson, D. (2009). The age of reason: Financial decisions over the life cycle and implications for regulation. *Brookings Papers on Economic Activity*, 2009:51–101.

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

Ashenfelter, O. and Card, D. (2002). Did the elimination of mandatory retirement affect faculty retirement? *American Economic Review*, 92(4):957–980.

Autor, D. H., Katz, L. F., and Kearney, M. S. (2008). Trends in us wage inequality: Revising the revisionists. *Review Of Economics and Statistics*, 90(2):300–323.

Ballesteros, S., Nilsson, L.-G., and Lemaire, P. (2009). Ageing, cognition, and neuroscience: An introduction. *European Journal of Cognitive Psychology*, 21(2-3):161–175.

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

Bloom, N., Eifert, B., Mahajan, A., McKenzie, D., and Roberts, J. (2012). Does management matter? evidence from India. *Quarterly Journal of Economics*, 128(1):1–51.

Bloom, N., Lemos, R., Sadun, R., and Van Reenen, J. (2015). Does management matter in schools? *Economic Journal*, 125(125):647–674.

Bloom, N. and Reenen, J. V. (2007). Measuring and explaining management practices across firms and countries. *The Quarterly Journal of Economics*, 122(4):1351–1408.

Blundell, R. and Macurdy, T. (1999). Chapter 27 labor supply: A review of alternative approaches. In Ashenfelter, O. and Card, D., editors, *Handbook of Labor Economics*, volume 3, Part A, pages 1559–1695. Elsevier.

Bonica, A., Chilton, A. S., Goldin, J., Rozema, K., and Sen, M. (2017). Measuring judicial ideology using law clerk hiring. *American Law and Economics Review*, 19(1):129–161.

Borsch-Supan, A. and Weiss, M. (2016). Productivity and age: Evidence from work teams at the assembly line. *Journal of the Economics of Ageing*, 7:30–42.

Callaway, B. and Sant'Anna, P. H. (2019). Difference-in-differences with multiple time periods.

Carrell, S. E. and West, J. E. (2010). Does professor quality matter? evidence from random assignment of students to professors. *Journal of Political Economy*, 118(3):409–432.

Caughey, D. and Warshaw, C. (2018). Policy preferences and policy change: Dynamic responsiveness in the american states, 1936–2014. *American Political Science Review*, 112(2):249–266.

Chetty, R. (2009). Sufficient statistics for welfare analysis: A bridge between structural and reduced-form methods. *Annual Review of Economics*, 1(1):451–488.

Chetty, R., Hendren, N., Kline, P., and Saez, E. (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., Gulati, G. M., and Posner, E. A. (2008). Judicial evaluations and information forcing: Ranking state high courts and their judges. *Duke LJ*, 58:1313.

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

Choi, S. J., Gulati, M., and Posner, E. A. (2013). The law and policy of judicial retirement: An empirical study. *The Journal of Legal Studies*, 42(1):pp.111–150.

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

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

Coviello, D., Ichino, A., and Persico, N. (2014). Time allocation and task juggling. *American Economic Review*, 104(2):609–23.

Coviello, D., Ichino, A., and Persico, N. (2015). The inefficiency of worker time use. *Journal of the European Economic Association*, 15(5):906–947.

Desjardins, R. and Warnke, A. J. (2012). Ageing and skills: A review and analysis of skill gain and skill loss over the lifespan and over time. Technical report, ZBW-German National Library of Economics.

Diermeier, D., Keane, M., and Merlo, A. (2005). A political economy model of congressional careers. *The American Economic Review*, 95(1):pp. 347–373.

Dimitrova-Grajzl, V., Grajzl, P., Zajc, K., and Sustersic, J. (2012). Judicial incentives and performance at lower courts: Evidence from slovenian judge-level data. *Review of Law & Economics*, 8(1):215–252.

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

Frederiksen, A. and Flaherty Manchester, C. (2019). Responding to regulation: The effects of changes in mandatory retirement laws on firm-provided incentives.

Galenson, D. W. (2011). *Old masters and young geniuses: The two life cycles of artistic creativity*. Princeton University Press.

Gentzkow, M., Shapiro, J. M., and Sinkinson, M. (2011). The effect of newspaper entry and exit on electoral politics. *American Economic Review*, 101(7):2980–3018.

Goldstein, J. (2011). Life tenure for federal judges raises issues of senility, dementia. *ProPublica*.

Goodman-Bacon, A. (2018). Difference-in-differences with variation in treatment timing. Working Paper 25018, National Bureau of Economic Research.

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

Grossmann, I., Na, J., Varnum, M. E., Park, D. C., Kitayama, S., and Nisbett, R. E. (2010). Reasoning about social conflicts improves into old age. *Proceedings of the National Academy of Sciences*, 107(16):7246–7250.

Gustman, A. L. and Steinmeier, T. L. (1986). A structural retirement model. *Econometrica*, 54(3):pp. 555–584.

Gustman, A. L. and Steinmeier, T. L. (1991). The Effects of Pensions and Retirement Policies on Retirement in Higher Education. *American Economic Review*, 81(2):111–15.

Gustman, A. L. and Steinmeier, T. L. (2005). The social security early entitlement age in a structural model of retirement and wealth. *Journal of Public Economics*, 89(2):441 – 463.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Keane, M. P. and Merlo, A. (2010). Money, political ambition, and the career decisions of politicians. *American Economic Journal: Microeconomics*, 2(3):pp. 186–215.

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

Landes, W. M. and Posner, R. A. (2009). Rational judical behavior: A statistical study. *The Journal of Legal Analysis*, 1:775–831.

Lazear, E. P. (1979). Why is there mandatory retirement? *Journal of Political Economy*, 87:1261–64.

Leiter, B. (2000). Measuring the academic distinction of law faculties. *The Journal of Legal Studies*, 29(S1):451–493.

Lemieux, T., MacLeod, W. B., and Parent, D. (2009). Performance pay and wage inequality. *Quarterly Journal of Economics*, 124(1):1–49.

Levin, S. G. and Stephan, P. E. (1991). Research productivity over the life cycle: Evidence for academic scientists. *The American Economic Review*, 81(1):pp. 114–132.

Lindenberger, U. (2014). Human cognition again: Corriger la fortune. *Science*, 346(6209):572–578.

Lumsdaine, R. L. and Mitchell, O. S. (1999). Chapter 49 new developments in the economic analysis of retirement. In Ashenfelter, O. and Card, D., editors, *Handbook of Labor Economics*, volume 3, Part C, pages 3261 – 3307. Elsevier.

Medoff, J. L. and Abraham, K. G. (1980). Experience, performance and earnings. *Quarterly Journal of Economics*, 95:703–36.

Meng, A., Nexø, M. A., and Borg, V. (2017). The impact of retirement on age related cognitive decline–a systematic review. *BMC geriatrics*, 17(1):160.

Munnell, A. H. (2015). The average retirement age–an update. *Center for retirement research*, 1920:1960–1980.

Nixon, D. C. and Haskin, J. D. (2000). Judicial retirement strategies the judge's role in influencing party control of the appellate courts. *American Politics Research*, 28(4):458–489.

Oster, S. M. and Hamermesh, D. S. (1998). Aging and productivity among economists. *Review of Economics and Statistics*, 80(1):154–156.

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

Posner, R. A. (1995). *Aging and old age*. University of Chicago Press.

Prendergast, C. (2001). Selection and oversight in the public sector, with the los angeles police department as an example. Technical report, National Bureau of Economic Research Working Paper: 8664.

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., and Baayen, H. (2014). The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Topics in Cognitive Science*, 6:5–42.

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.

Shi, L. (2009). The limit of oversight in policing: Evidence from the 2001 Cincinnati riot. *Journal of Public Economics*, 93(1-2):99–113. Shi, Lan.

Small, B. J., Dixon, R. A., and McArdle, J. J. (2011). Tracking cognition–health changes from 55 to 95 years of age. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 66(suppl 1):i153–i161.

Smyth, R. and Bhattacharya, M. (2003). How fast do old judges slow down?: A life cycle study of aging and productivity in the federal court of australia. *International Review of Law and Economics*, 23(2):141–164.

Stock, J. H. and Wise, D. A. (1990). Pensions, the option value of work, and retirement. *Econometrica*, 58(5):pp. 1151–1180.

Sullivan, D. and von Wachter, T. (2009). Job displacement and mortality: An analysis using administrative data. *Quarterly Journal Of Economics*, 124(3):1265–1306.

Tanaka, H. and Higuchi, M. (1998). Age, exercise performance, and physiological functional capacities. *Advances in exercise and sports physiology*, 4(2):51–56.

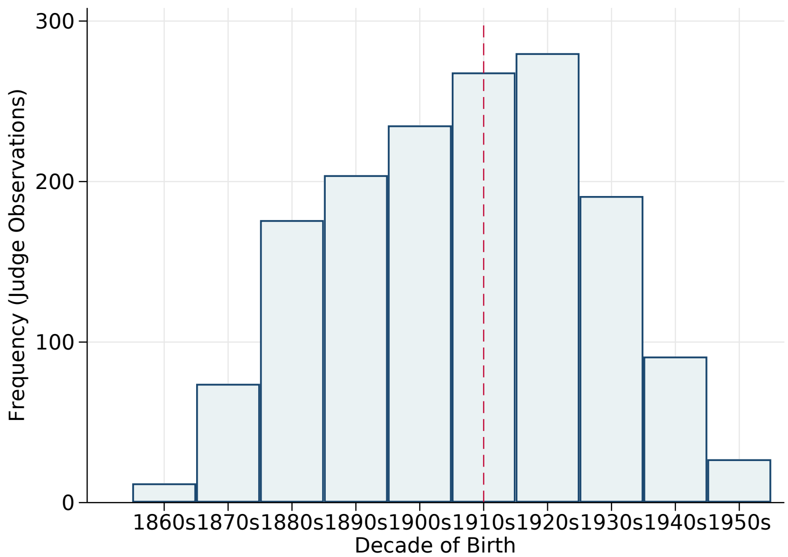
Teitelbaum, J. C. (2006). Age and tenure of the justices and productivity of the u.s. supreme court: Are term limits necessary? *Florida State University Law Review*, 34:161–182.

Appendix

A Background and Data

A.1 Aging and Retirement Decisions

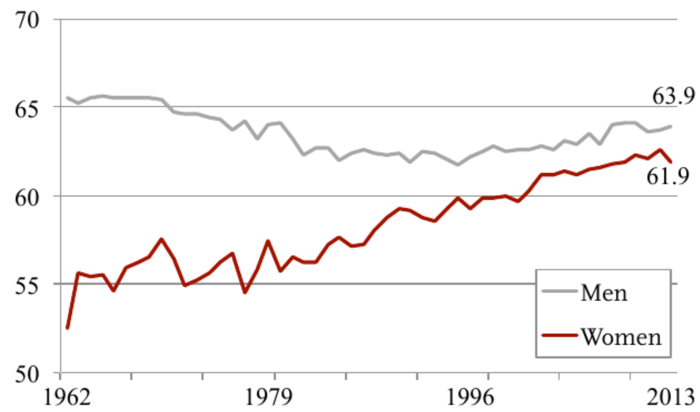
Figure A.1: Distribution of Judge Birth Decades



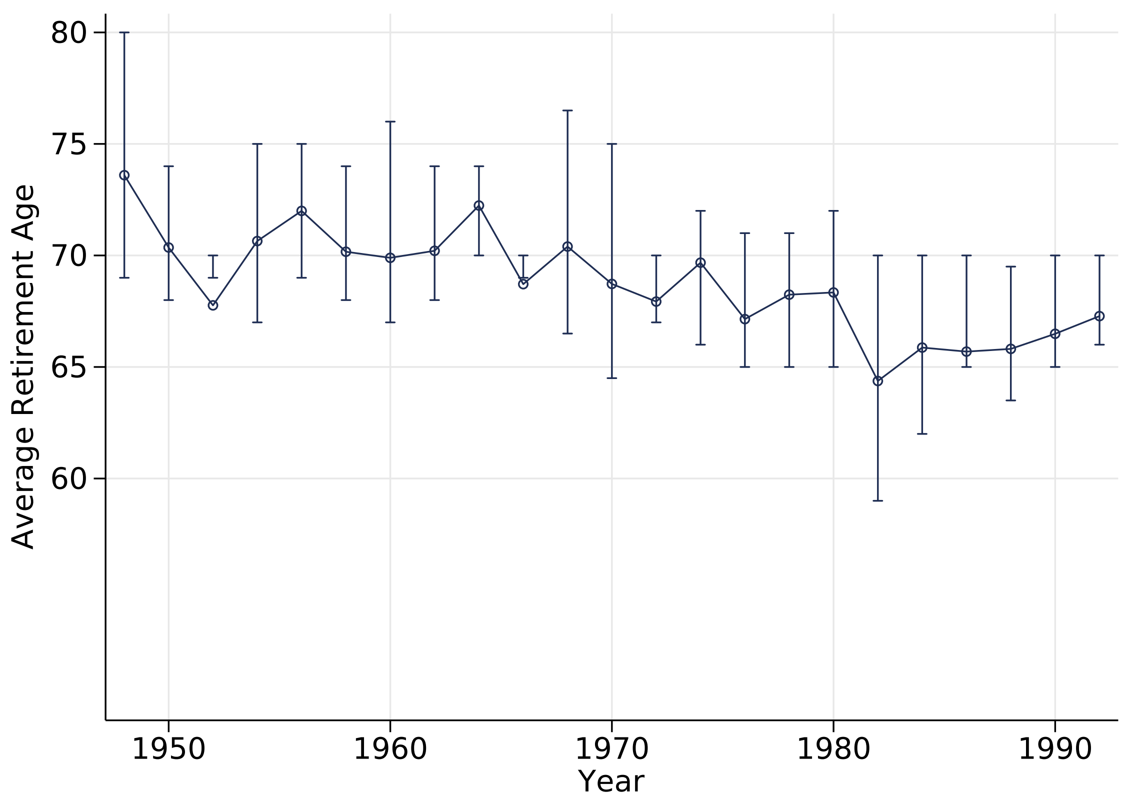
Number of judges in each birth decade cohort. Vertical dashed line at median.

Figure A.2: Average Retirement Age for U.S. Workers and Judges,

(A) Average Retirement Age for U.S. Workers, by Gender, 1962-2013

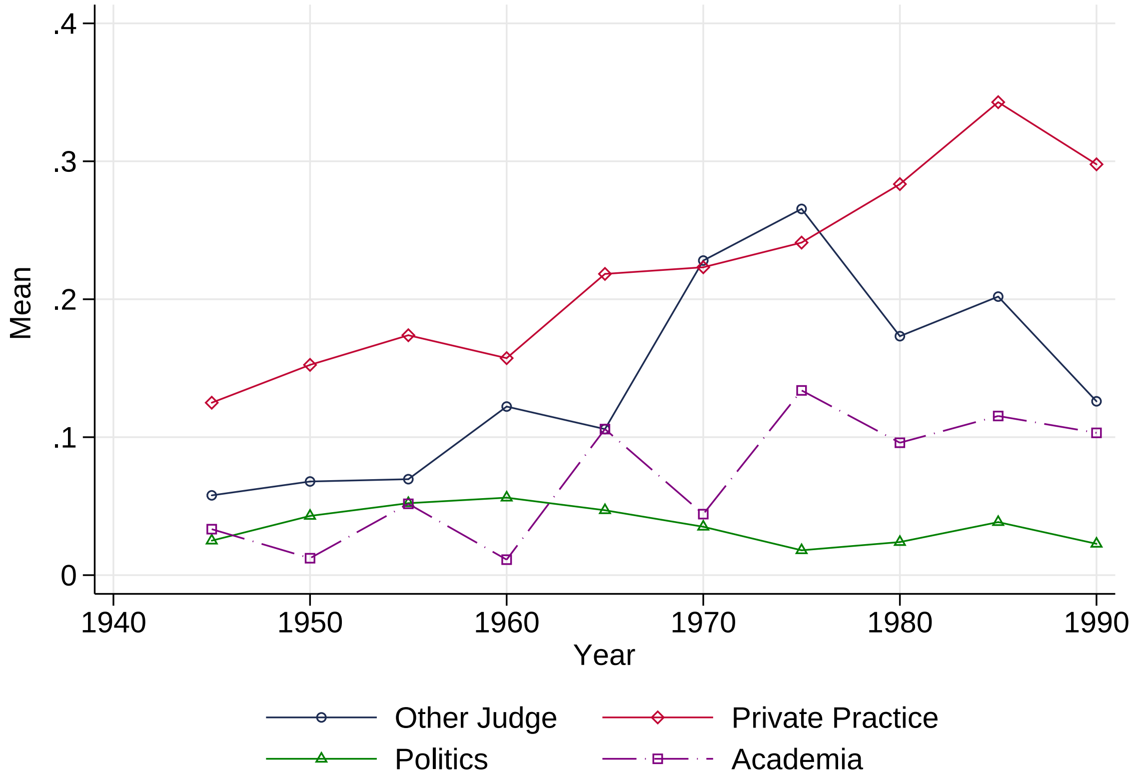


(B) Average Retirement Age for State Supreme Court Judges, 1948-1994



Panel (A): Average retirement age by gender for U.S. workers, computed from CPS by Munnell (2015). Panel (B): Average retirement age of state supreme court judges, by year. Error spikes give 25th and 75th percentiles.

Figure A.3: Post-Judgeship Careers



Proportion of judges with documented careers after their state supreme court judgeship, including other judgeship, private practice, politics, and academia. Plotted by five-year bins.

What do judges do after retirement? Figure A.3 shows the trends in these career choices. At the beginning of the sample, few judges took on more work after their judgeship. That has become more common in recent years. If they do take another career, it is usually in private practice as an attorney.

A.2 Mandatory Retirement

Table A.1: Tabulations on Treatment and Control Judges

Reform	Number of Courts with Treatment Variation	Number of Obs (Judge-Year)			Number of Judges		
		Controls	Treated	Total	Controls	Treated	Total
Retirement Age 70	9	11265	1390	12655	1240	172	1368
Retirement Age 72	3	12373	884	13257	1348	88	1421
Retirement Age 75	2	10584	259	12373	1191	34	1219

Notes. Summary tabulations on Retirement Reform Judges. The column gives the number of courts that experience a change in the retirement rule. The second set of columns gives the number of judge-year observations in the control and treatment groups (and total) when a change in retirement rule occurs in that year. The third list of columns gives the number of judges in these respective groups.

Appendix Table A.1 provides tabulations on the relevant treatment variation in the data for mandatory retirement reforms. 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.

Table A.2: Rules on Judge Senior Status, by State

State	Retirement Age	Seniority	Note
Alabama	70	Yes	Supernumerary judge
Alaska	70	No	Only to work on temporary assignments
Arizona	70	No	Only to work on temporary assignments
Arkansas	none	—	No retirement benefits if seek re-election past age 70
California	none	—	
Colorado	72	No	
Conneticut	70	Yes	State referee
Delaware	none	—	
Florida	75	No	Allows temporary assignments
Georgia	none	—	
Hawaii	70	No	
Idaho	none	—	
Illinois	none	—	Retirement Act for age 75, declared unconstitutional in 2009
Indiana	75	No	
Iowa	72	Yes	
Kansas	75	No	
Kentucky	none	—	
Louisiana	70	No	
Maine	none	—	
Maryland	70	No	
Massachusetts	70	No	
Michigan	70	No	
Minnesota	70	No	
Mississippi	none	—	
Missouri	70	Yes	Senior judge
Montana	none	—	
Nebraska	none	—	
Nevada	none	—	
New Hampshire	70	No	
New Jersey	70	No	
New Mexico	none	—	
New York	70	Yes	May serve after 70 until 76
North Carolina	72	No	Only to work on temporary assignments
North Dakota	none	—	
Ohio	70	No	
Oklahoma	none	—	
Oregon	75	No	Legislature may ask retired judges to work on temporary assignments
Pennsylvania	75	Yes	Senior judge
Rhode Island	none	—	
South Carolina	72	—	
South Dakota	70	No	
Tennessee	none	—	
Texas	75	No	Conditions may vary based on Art. 5 of Texas Constitution
Utah	75	No	
Vermont	90	No	
Virginia	73	No	
Washington	75	No	
West Virginia	none	—	
Wisconsin	70	Yes	Can serve a judge on a temporary basis
Wyoming	70	Yes	

A.3 Case Assignment

Table A.3: Summary Statistics on Area of Law and Related Industries

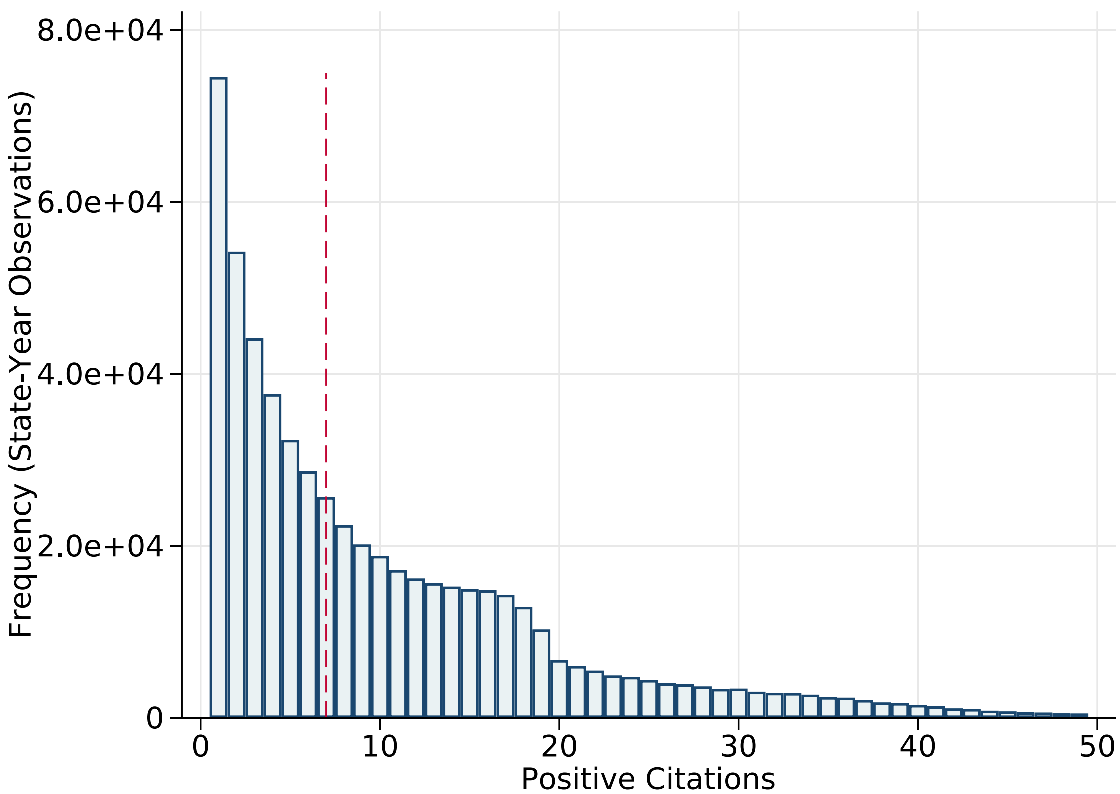
Area of Law	Freq.	Percent	Related Industrial Sector	Freq.	Percent
Criminal Law	191810	21.85	Real Estate	28527	13.64
Civil Procedure	74757	8.52	Law Enforcement	10758	5.14
Evidence	66377	7.56	Automobiles	10206	4.88
Torts	57915	6.6	Insurance	9158	4.38
Damages & Remedies	45073	5.14	Tax	8509	4.07
Contracts	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 A.4: 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).

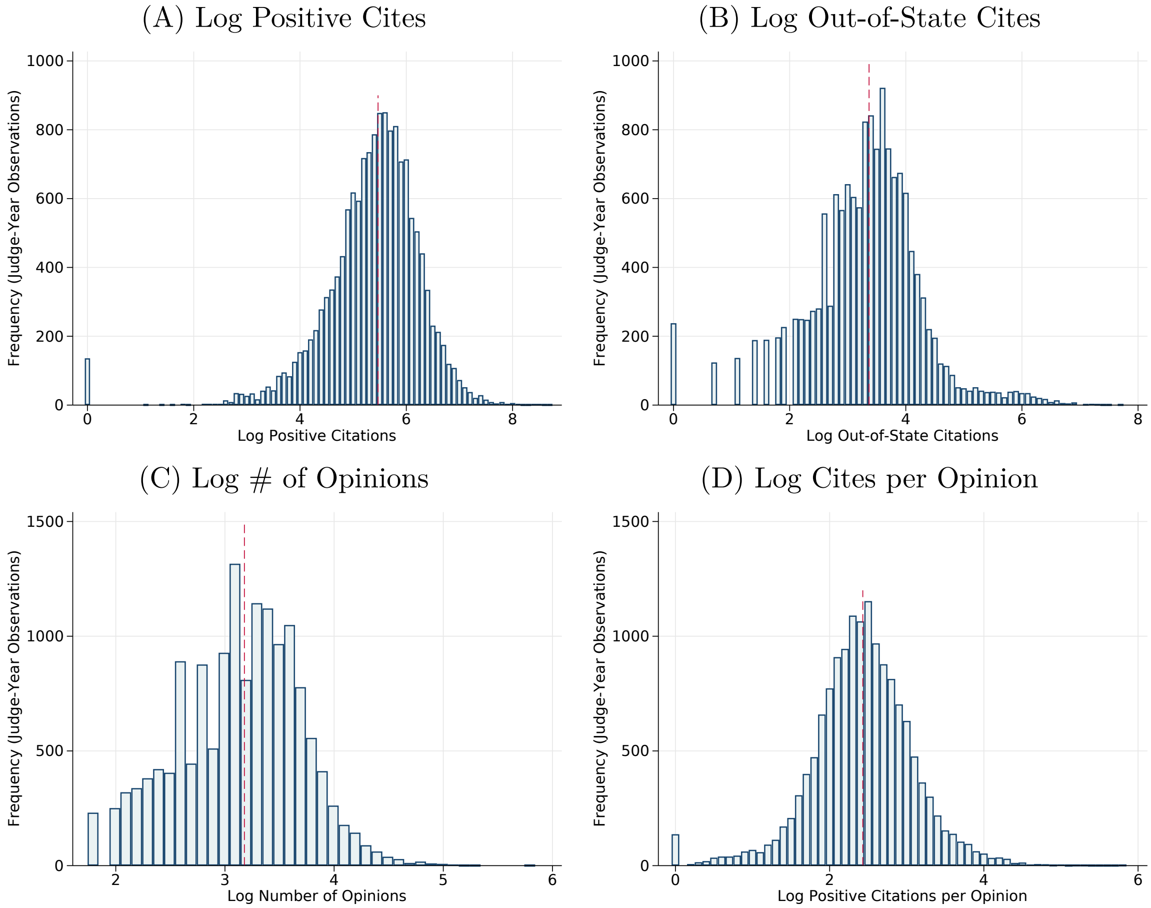
Figure A.4: Distribution of Citations by Case



Notes. Hstiogram of the number of positive cites per case in the dataset. Vertical dashed line at the median.

A.4 Judge Performance

Figure A.5: Distributions of Outcome Variables - Judge-Year



Notes. Histograms of judge-year performance measures. Vertical dashed line at median.

A.5 Clerk Quality and Judge Age

A potential confounder for the relationship between judge age and judge performance is selection of different types of clerks to judges by different ages. To check this, we require a dataset on clerk quality, which we could then match to judge age and look for systematic variation in clerk quality. Here, we do this with a dataset of federal court clerks for which we have the ranking of the law schools they attended.

Data on clerks is not available for state supreme courts. But it is available for a large sample of judges in the federal courts for the years 1995 through 2004, collected by Bonica et al. (2017). We used the information on clerk law school attendance from this dataset.

We have data on law school rankings from Leiter (2000), who produced a variety of rankings using data from 1998. For our purposes, it is helpful that there is a ranking on "quality of student body", constructed from LSAT scores and college GPA. This ranking is available for the top 50 schools, with Yale at #1, Georgetown at #10, Fordham at #25, and Chicago-Kent at #50.

For each clerkship, we have information on the judge and the clerk. For the judge, we have information on the court and the age in years. For circuit judges, we have the specific circuit (1st, 2nd, D.C., etc). For the district courts, the data do not include the specific court. Therefore the best specification, in terms of netting out all court-year-level factors, restricts analysis to the circuit courts.

Table A.5: Judge Age and Clerk Quality (Evidence from U.S. Federal Courts

	(1)	(2)	(3)	(4)	(5)	(6)
	Clerk Law School Ranking (Higher is Worse Ranking)					
Judge Age (Years)	-0.00162	0.0593	0.0306	0.0199	0.0767*	0.0396
	(0.0304)	(0.0539)	(0.0643)	(0.0214)	(0.0388)	(0.0453)
Judge Sample	Only Circuit Courts			All Federal Courts		
Court-Year FE	X	X	X			
Court-Type X Year FE				X	X	X
School Ranks Included	1-25	1-50	1-50	1-25	1-50	1-50
Max Age 70			X			X
N	3215	3683	3483	7623	9632	9174
Adj. R-sq	0.040	0.052	0.047	0.045	0.071	0.068

Notes. Observation is a clerk-year. "Clerk Law School Ranking" is the student body quality ranking, 1-50, from Leiter (2000). Standard errors clustered by judge in parentheses. + p<.0.1, * p<0.05, ** p<0.01.

Appendix Table A.5 reports the estimates for a regression of clerk law school ranking on judge age. The regression for circuit courts (Columns 1 through 3) include court-year fixed effects. The regression for all federal courts (Columns 4 through 6) includes court-type × year fixed effects. Columns 1 and 4 include clerks from the top 25 schools. Columns 2, 3, 5, and 6 include the full ranking of 50 schools. Columns 3 and 6 exclude judges above the age of 70. Standard errors are clustered by judge.

Overall, there is no consistent relationship between judge age and clerk quality. In the preferred specifications in the circuit courts, there is no relationship. When including all courts (where we cannot include court-year fixed effects), there is a statistically significant positive effect, suggesting that older judges get worse-ranked clerks.

B Additional Analysis of Mandatory Retirement Reforms

Table B.1: Effect of Reform on Citations: Additional Measures

	(1)	(2)	(3)	(4)	(5)	(6)
	Cites After 10 years		Has Cite		Drop 5% Outliers	
Ret. Reform	0.140*	0.169*	0.0218	0.0307	0.233**	0.263**
	(0.0648)	(0.0688)	(0.0243)	(0.0268)	(0.0774)	(0.0836)
Year / Court FE	X	X	X	X	X	X
Trends/Windows		X		X		X
N	15010	15010	15010	15010	14763	14763
R-sq	0.410	0.473	0.773	0.804	0.454	0.520

Notes. Observation is a judge working in a year. "Retirement Reform" is an indicator for the ten years after the introduction of mandatory retirement. "Cites after 10 Years" means the log of the positive cites to a judge's cases in a year, from cases more than ten years later. "Has Cite" means the proportion of cases with at least one positive citation. "Drop 5% Outliers" is the baseline outcome (log of positive cites to a judge's cases in a year) but for each court-year, the top 5% of cases by cite count are dropped. Court Treat Windows means court-specific treatment windows (ten years before and after reform). Standard errors clustered by state in parentheses. + p<.0.1, * p<0.05, ** p<0.01.

Table B.2: Effect of Reform on Citations: Different Windows

	(1)	(2)	(3)	(4)	(5)	(6)
	Effect on Log Positive Cites					
Treatment Window	6		14		All	
Retirement Reform	0.171**	0.159**	0.218**	0.204**	0.0493	0.303**
	(0.0552)	(0.0532)	(0.0776)	(0.0693)	(0.114)	(0.0929)
Year FE, Court FE	X	X	X	X	X	X
Court Trends/Windows		X		X		X
N	2357	2357	2357	2357	2357	2357
R-sq	0.811	0.883	0.812	0.885	0.810	0.880

Notes. Observation is a judge working in a year. "Retirement Reform" is an indicator for the ten years after the introduction of mandatory retirement. Court Treat Windows means court-specific treatment windows (ten years before and after reform). Standard errors clustered by state in parentheses. + p<.0.1, * p<0.05, ** p<0.01.

Table B.3: Effect of Reform on Log Cites: Senior Status Rules

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Effect on Log Citations per Judge-Year							
Retirement Reform	0.209*	0.282*	0.245**	0.246*	0.251**	0.243*	0.202*	0.270*
	(0.0879)	(0.119)	(0.0809)	(0.0904)	(0.0799)	(0.0896)	(0.0904)	(0.103)
× Grandfather Rule	0.0657	-0.127						
	(0.111)	(0.146)						
× Finish Term			-0.115	-0.0939				
			(0.0762)	(0.0983)				
× Finish Term Half					-0.355**	-0.0573		
					(0.0749)	(0.109)		
× Finish Year							0.134	-0.135
							(0.120)	(0.115)
Year FE, Court FE	X	X	X	X	X	X	X	X
Court Trends/Windows		X		X		X		X
N	9868	9868	9868	9868	9868	9868	9868	9868
R-sq	0.370	0.438	0.370	0.438	0.371	0.438	0.370	0.438

Observation is a judge working in a year. "Retirement Reform" is an indicator for the ten years after the introduction of mandatory retirement. Coefficients are interacted with respective senior status rules (respectively: the rule not applying to sitting judges, being allowed to finish the term, being allowed to finish terms that are over halfway finished, and being able to finish out the year). Court Treat Windows means court-specific treatment windows (ten years before and after reform). Standard errors clustered by state in parentheses. + p<.0.1, * p<0.05, ** p<0.01.

Table B.4: Effect of Reform on Log Cites, Only Reform States

	(1)	(2)	(3)	(4)	(5)
	Effect on Log Citations per Judge-Year				
Retirement Reform	0.295**	0.262**	0.273*	0.349*	0.552**
	(0.0818)	(0.0869)	(0.113)	(0.158)	(0.176)
Court FE, Year FE	X	X	X	X	X
Court Trends/Windows		X	X	X	X
Init Court Rules × Year FE			X	X	X
Init Case Types × Year FE				X	X
Init Age × Year FE					X
N	5562	5562	5562	5562	5562
R-sq	0.288	0.346	0.382	0.435	0.457

Observation is a judge working in a year. "Retirement Reform" is an indicator for the ten years after the introduction of mandatory retirement. Court Treat Windows means court-specific treatment windows (ten years before and after reform). "Init X" × year FE means initial values are interacted with year. "Init Court Rules" includes a state's 1947 rules for judge selection/retention system, admin office, intermediate appellate court, number of judges, and term length. "Init Case Types" includes a court's 1947 average values for case characteristics (legal area and related industries). "Init Age" includes the initial mean and standard deviation for judge age on the court. Standard errors clustered by state in parentheses. + p<.0.1, * p<0.05, ** p<0.01.

Table B.5: Effect of Reform on Citations: Alternative Clustering

	(1)	(2)	(5)	(6)
Clustering Group	State and Year		None (Robust)	
Retirement Reform	0.228**	0.253**	0.228**	0.253**
	(0.0458)	(0.0436)	(0.0291)	(0.0300)
Court FE, Year FE	X	X	X	X
Court Trends/Windows		X		X
N	15010	15010	15010	15010
R-sq	0.460	0.526	0.460	0.526

Notes. Observation is a judge working in a year. "Retirement Reform" is an indicator for the ten years after the introduction of mandatory retirement. Court Treat Windows means court-specific treatment windows (ten years before and after reform). Standard errors clustered by state in parentheses. + p<.0.1, * p<0.05, ** p<0.01.

Table B.6: Effect of Reform on Log Cites, Alternative Weighting

	(1)	(2)	(3)	(4)
	Effect on Log Positve Cites per Judge-Year			
Retirement Reform	0.211**	0.253**	0.133+	0.181*
	(0.0750)	(0.0858)	(0.0736)	(0.0802)
Weighting	# of Opinions		Judges Equal	
Court FE, Year FE	X	X	X	X
Court Trends/Windows		X		X
N	15010	15010	15010	15010
R-sq	0.496	0.569	0.421	0.499

Observation is a judge working in a year. "Retirement Reform" is an indicator for the ten years after the introduction of mandatory retirement. Court Treat Windows means court-specific treatment windows (ten years before and after reform). Standard errors clustered by state in parentheses. + p<.0.1, * p<0.05, ** p<0.01.

Table B.7: Effect of Reform on Log Cites, with Time-Varying Controls

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Effect on Log Positive Cites per Judge-Year							
Retirement Reform	0.106+	0.148*	0.251**	0.271**	0.202**	0.225**	0.143**	0.172*
	(0.0572)	(0.0703)	(0.0886)	(0.0804)	(0.0725)	(0.0804)	(0.0504)	(0.0644)
Court FE, Year FE	X	X	X	X	X	X	X	X
Court Trends/Windows		X		X		X		X
Case Controls	X	X						
Rule Controls			X	X				
Judge Experience FE					X	X		
Lagged Dep. Var.							X	X
N	15010	15010	13863	13863	15006	15006	13304	13304
R-sq	0.611	0.652	0.457	0.523	0.478	0.545	0.585	0.609

Observation is a judge working in a year. "Retirement Reform" is an indicator for the ten years after the introduction of mandatory retirement. Court Treat Windows means court-specific treatment windows (ten years before and after reform).Case controls means the first five principal components of the matrix of controls for legal topic and related industries. Rule controls means rules for selection and retention of juges and other institutional items. Judge Experience FE means fixed effects for yeras on the court. Lagged Dep. Var. means the court-year lag of the dependent variable (log positive cites in the previous year). Standard errors clustered by state in parentheses. + p<.0.1, * p<0.05, ** p<0.01.

Table B.8: Effect of Reform on Citations: Separately by Maximum Age Imposed

	(1)	(2)	(3)	(4)	(5)	(6)
	Effect on Log Positive Cites					
Maximum Age	70		72		75	
Retirement Reform	0.212+	0.291*	0.396**	0.310**	0.155+	0.165**
	(0.117)	(0.133)	(0.113)	(0.0773)	(0.0899)	(0.0384)
Court FE, Year FE	X	X	X	X	X	X
Court Trends/Windows		X		X		X
N	15010	15010	15010	15010	15010	15010
R-sq	0.459	0.524	0.459	0.518	0.458	0.517

Notes. Observation is a judge working in a year. "Retirement Reform" is an indicator for the ten years after the introduction of mandatory retirement. Court Treat Windows means court-specific treatment windows (ten years before and after reform). Standard errors clustered by state in parentheses. + p<.0.1, * p<0.05, ** p<0.01.

Table B.9: Effect of Mandatory Retirement Reform, Merit vs. Other Systems

	(1)	(2)	(3)	(4)
	Effect on Log Positive Cites			
Appointment System	Merit		Other System	
Retirement Reform	-0.124	-0.0314	0.193*	0.243*
	(0.130)	(0.112)	(0.0849)	(0.107)
Year FE, Court FE	X	X	X	X
Court Trends/Windows		X	X	X
N	2442	2442	12568	12568
R-sq	0.375	0.428	0.474	0.535

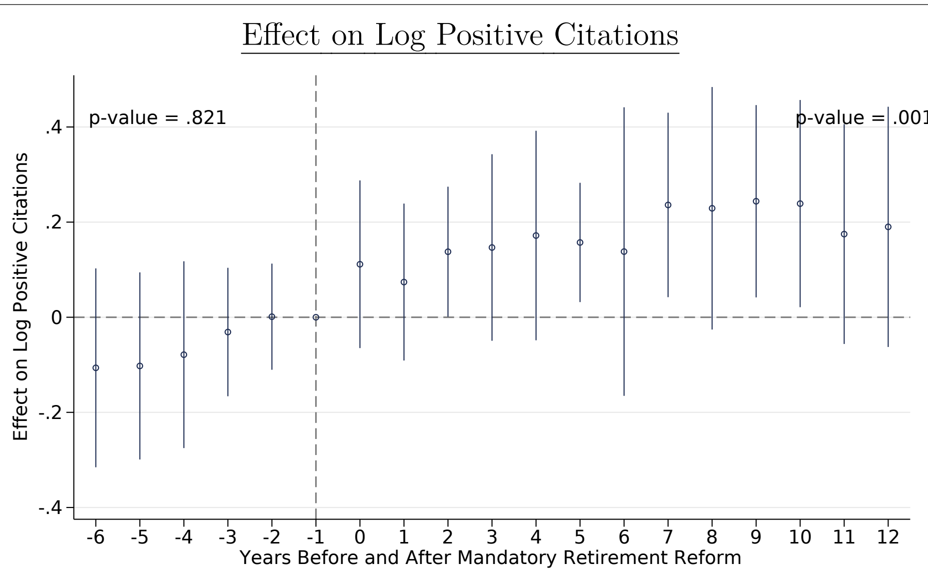
Notes. Observation is a judge working in a year. "Retirement Reform" is an indicator for the ten years after the introduction of mandatory retirement. "Experience" is the years of experience of each judge. Each of the outcome about the end of judgeship are dummy equal to 1 indicating how the judgeship ended. Court Treat Windows means court-specific treatment windows (ten years before and after reform). Standard errors clustered by state in parentheses. + p<.0.1, * p<0.05, ** p<0.01.

Table B.10: Effect of Mandatory Retirement Reform, Other Institutional Features

	(1)	(2)	(3)	(4)	(5)	(6)
			Effect on How Judgeship Ended			
	Experience		Retirement	Death	Other Job	Lost Election
Retirement Reform	-1.301*	-1.464*	-0.0123	0.00633	0.00687	-0.0134
	(0.593)	(0.695)	(0.0463)	(0.0281)	(0.0505)	(0.0163)
Year FE, Court FE	X	X	X	X	X	X
Court Trends/Windows		X	X	X	X	X
N	15010	15010	15010	15010	15010	15010
R-sq	0.162	0.189	0.155	0.114	0.125	0.074

Notes. Observation is a judge working in a year. "Retirement Reform" is an indicator for the ten years after the introduction of mandatory retirement. "Experience" is the years of experience of each judge. Each of the outcome about the end of judgeship are dummy equal to 1 indicating how the judgeship ended. Court Treat Windows means court-specific treatment windows (ten years before and after reform). Standard errors clustered by state in parentheses. + p<.0.1, * p<0.05, ** p<0.01.

Figure B.1: Event-Study Effect of Reform with Initial Covariates



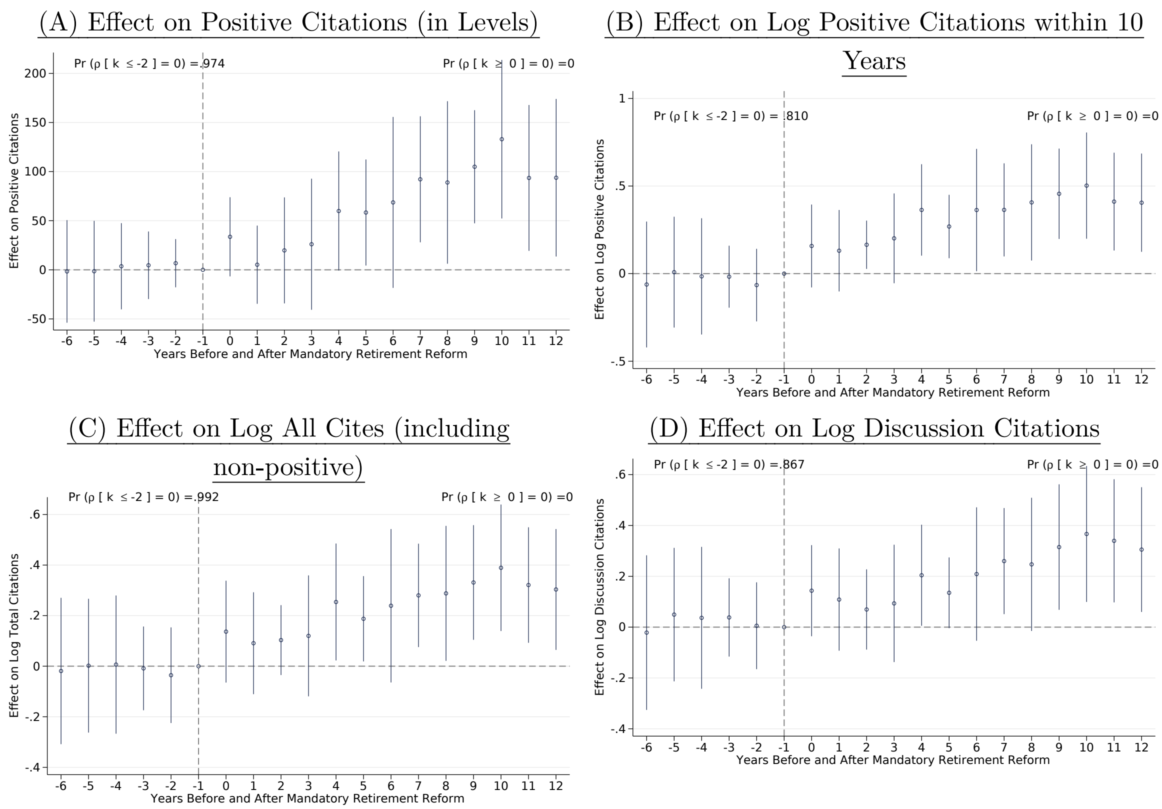
Court performance before and after reforms implementing retirement ages of 70, 72 or 75. The outcome is the log positive citations of a judge in a year. Time series is a coefficient plot from the event study regression (2), with coefficents estimated relative to the year before the reform. Regression includes court and year fixed effects and court-specific event windows, plus initial rules, cases, and age variables, interacted with year. 95% confidence intervals constructed with standard errors clustered by state.

Table B.11: Effect of Reform on Inverse Hyperbolic Sine of Citations

	(1)	(2)	(3)	(4)	(5)
	Inverse Hyperbolic Sine of Positive Cites per Judge-Year				
Retirement Reform	0.231**	0.257**	0.240**	0.324**	0.304**
	(0.0787)	(0.0868)	(0.0860)	(0.0941)	(0.0960)
Court FE, Year FE	X	X	X	X	X
Court Trends/Windows		X	X	X	X
Init Court Rules × Year FE			X	X	X
Init Case Types × Year FE				X	X
Init Age × Year FE					X
N	15010	15010	15010	15010	15010
R-sq	0.451	0.519	0.530	0.546	0.556

Notes. DD effect of mandatory retirement reform on inverse hyperbolic sine (asinh) positive citations to a judge's opinions in ten years after reform, relative to ten years before reform. Observation is a judge working in a year. "Ret. Reform" is a treatment indicator for the ten years after the introduction of mandatory retirement. Court Treat Windows means court-specific treatment windows (ten years before and after reform). "Init X" × year FE means initial values are interacted with year. "Init Court Rules" includes a state's 1947 rules for judge selection/retention system, admin office, intermediate appellate court, number of judges, and term length. "Init Case Types" includes a court's 1947 average values for case characteristics (legal area and related industries). "Init Age" includes the initial mean and standard deviation for judge age on the court. Standard errors clustered by state in parentheses. + p<.0.1, * p<0.05, ** p<0.01.

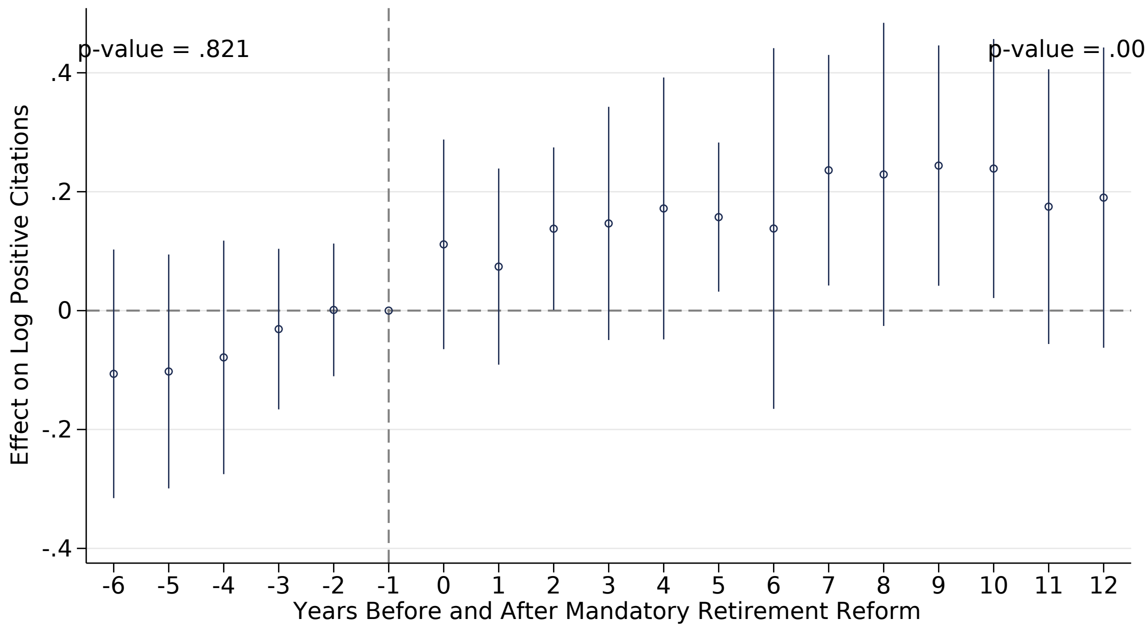
Figure B.2: Event-Study Effect on Performance: Alternative Cite Measures



Judge performance before and after reforms implementing retirement ages of 70, 72 or 75. Panel A: outcome is number of positive citations of a judge in a year (in levels, rather than logs). Panel B: outcome is the log positive citations of a judge in a year that were made within ten years of a case. Panel C: outcome is the log total citations of a judge in a year (including non-positive negative cites). Panel D: outcome is the log discussion citations of a judge in a year. Time series is a coefficient plot from the event study regression (2), with coefficents estimated relative to the year before the reform. Regression includes court and year fixed effects and court-specific event windows. 95% confidence intervals constructed with standard errors clustered by state.

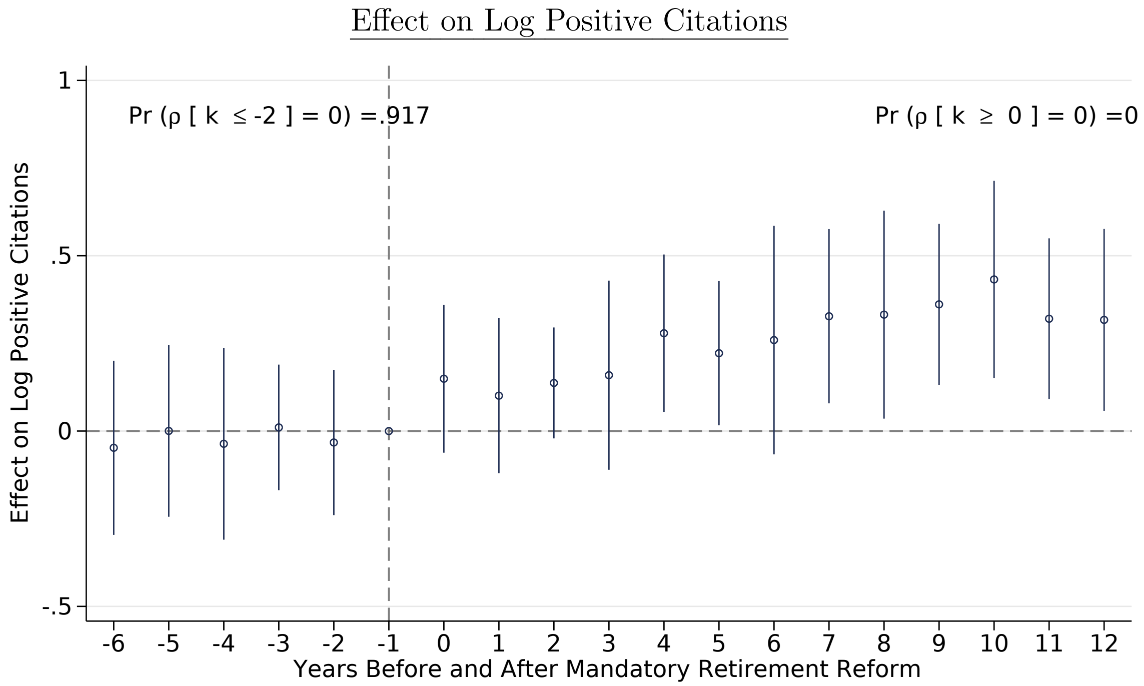
Figure B.3: Event-Study Effect of Reform: Collapsed Court-Year Data

Effect on Log Positive Citations



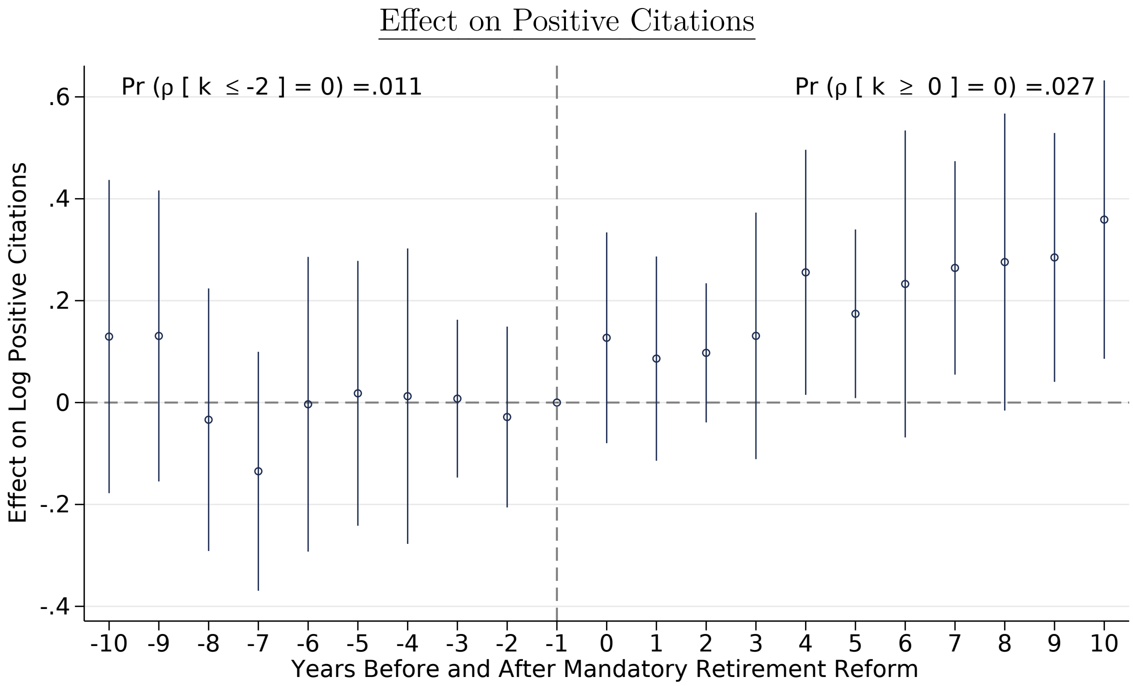
Court performance before and after reforms implementing retirement ages of 70, 72 or 75. The outcome is the log positive citations of a court in a year. Time series is a coefficient plot from the event study regression (2), with coefficents estimated relative to the year before the reform. Regression includes court and year fixed effects and court-specific event windows. 95% confidence intervals constructed with standard errors clustered by state.

Figure B.4: Event-Study Effect of Reform on Court Performance: Only Reform States



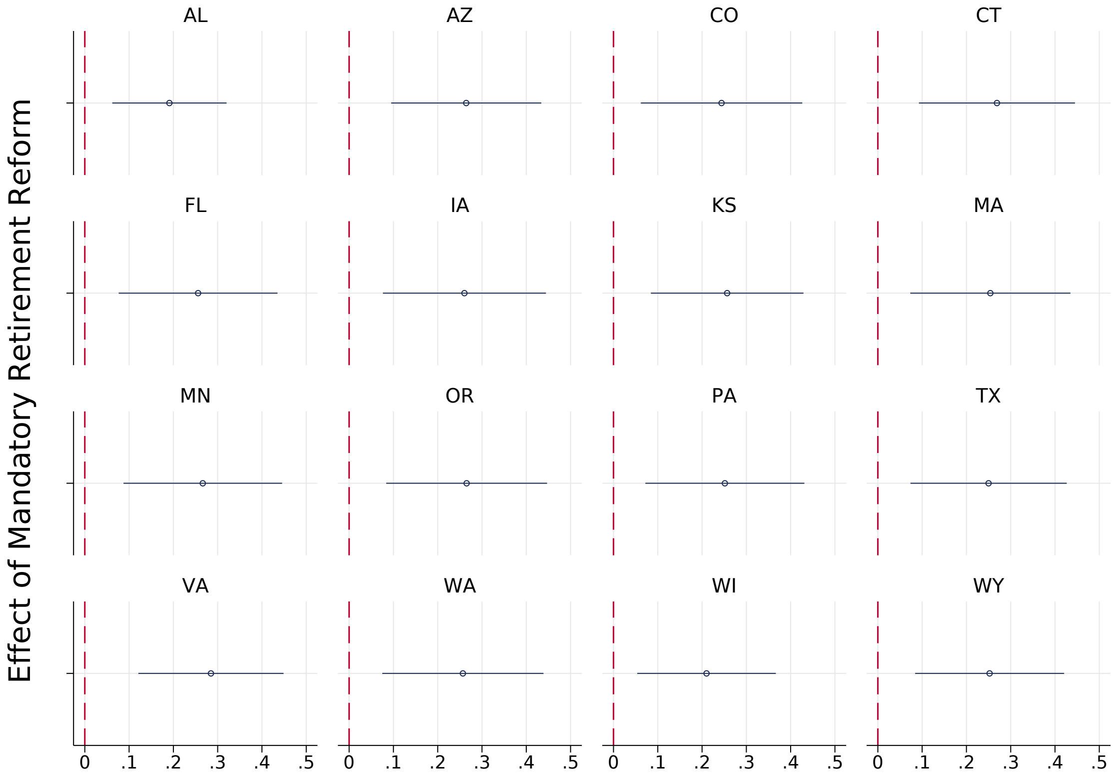
Court performance before and after reforms implementing retirement ages of 70, 72 or 75. Sample limited to reform states. The outcome is the log positive citations of a judge in a year. Time series is a coefficient plot from the event study regression (2), with coefficents estimated relative to the year before the reform. Regression includes court and year fixed effects and court-specific event windows. 95% confidence intervals constructed with standard errors clustered by state.

Figure B.5: Event-Study Effect of Reform on Court Performance: 10 Pre-Periods



Judge performance before and after reforms implementing retirement ages of 70, 72 or 75. The outcome is the log positive citations of a judge in a year. Ten periods before and after included. Time series is a coefficient plot from the event study regression (2), with coefficents estimated relative to the year before the reform. Regression includes court and year fixed effects and court-specific event windows. 95% confidence intervals constructed with standard errors clustered by state.

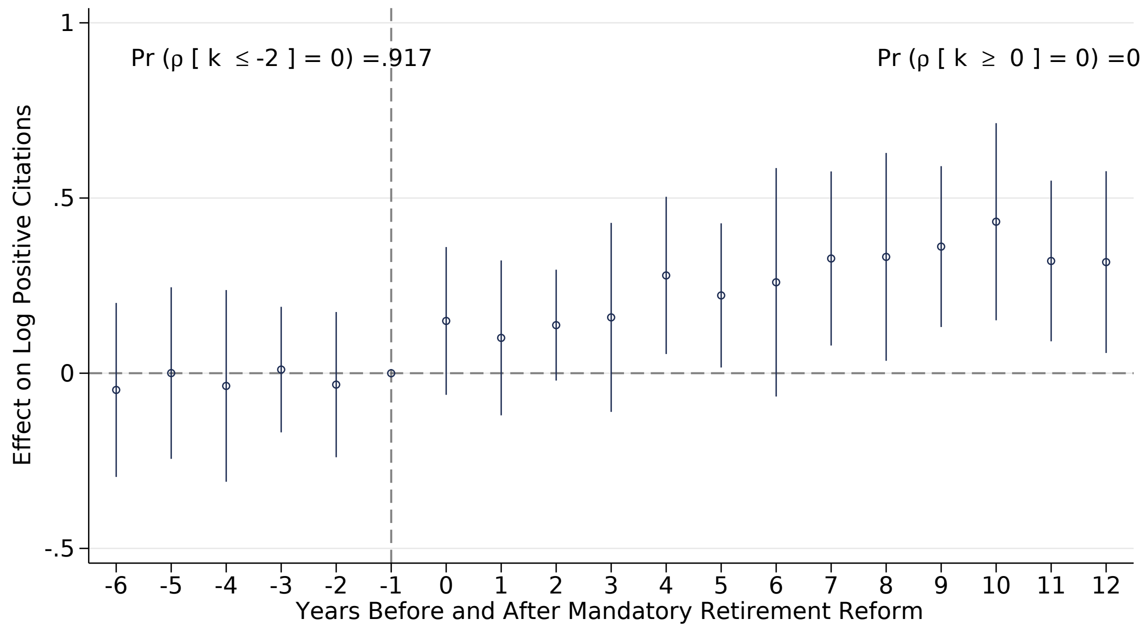
Figure B.6: DD Effect of Reform, Dropping each Treated State Individually



Coefficient for the effect of mandatory retirement at ages of 70, 72 or 75 on judge performance. The outcome is the log positive citations of a judge in a year. Each subfigure plots the coefficient from regression 1 excluding one treated state at a time. Includes court and year fixed effects, court-specific windows and trends.

Figure B.7: Event-Study Effect on Court Performance: Only Reform States

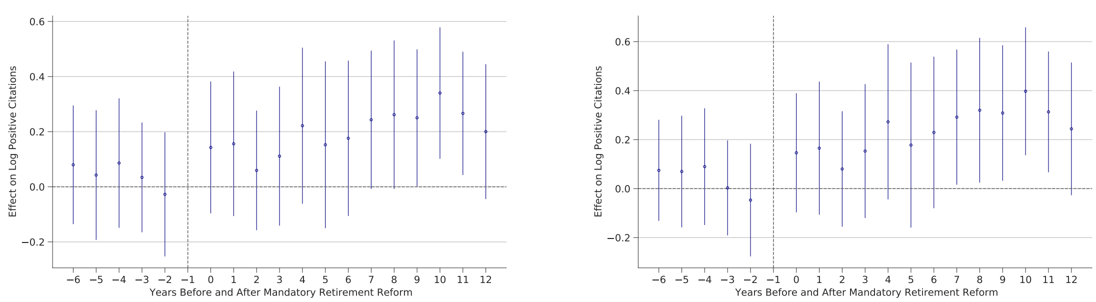
Effect on Log Positive Citations



Court performance before and after reforms implementing retirement ages of 70, 72 or 75. Sample limited to reform states. The outcome is the log positive citations of a judge in a year. Time series is a coefficient plot from the event study regression (2), with coefficents estimated relative to the year before the reform. Regression includes court and year fixed effects and court-specific event windows. 95% confidence intervals constructed with standard errors clustered by state.

Figure B.8: Event-Study Effect on Performance: Adjustment for Staggered Treatment

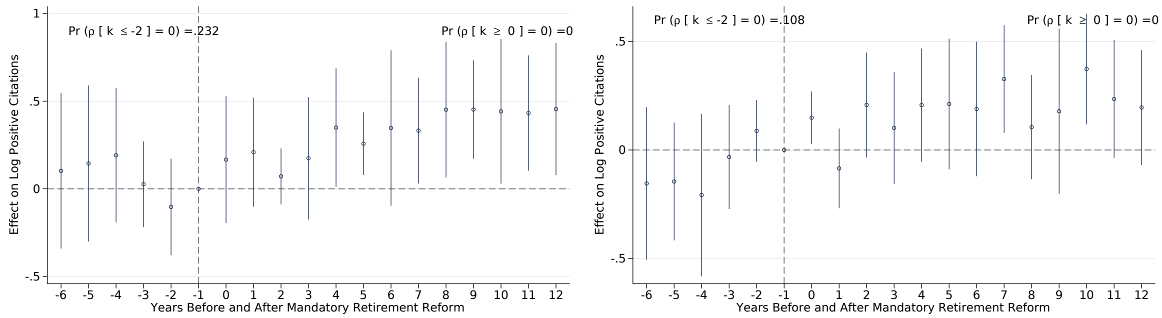
(A) Un-Weighted Averaged Estimates (B) Weighted Average Estimates



Log positive cites to court before and after reforms implementing retirement ages of 70, 72 or 75. Time series is a coefficient plot from the event study regression (2), with coefficents estimated relative to the year before the reform. Regression includes court and year fixed effects and court-specific event windows. Coefficients and standard errors (clustered by state) adjusted for staggered treatment timing, following the method in Callaway and Sant'Anna (2019), as described in Section 3.3. Panel B uses a weighted average of the estimates based on the number of judge-year observations in the event-study window for each treated state.

Figure B.9: Event-Study Effect, by Random and Discretionary Case Assignment

(A) Random Assignment (B) Discretionary Assignment



Notes. Court performance before and after reforms implementing retirement ages of 70, 72 or 75. The outcome is the log positive citations to the court in a year. Panel A includes courts with random or rotating assignment of cases. Panel B includes courts with discretionary assignment of cases. Time series is a coefficient plot from the event study regression (2), with coefficents estimated relative to the year before the reform. Regression includes court and year fixed effects and court-specific event windows. 95% confidence intervals constructed with standard errors clustered by state.

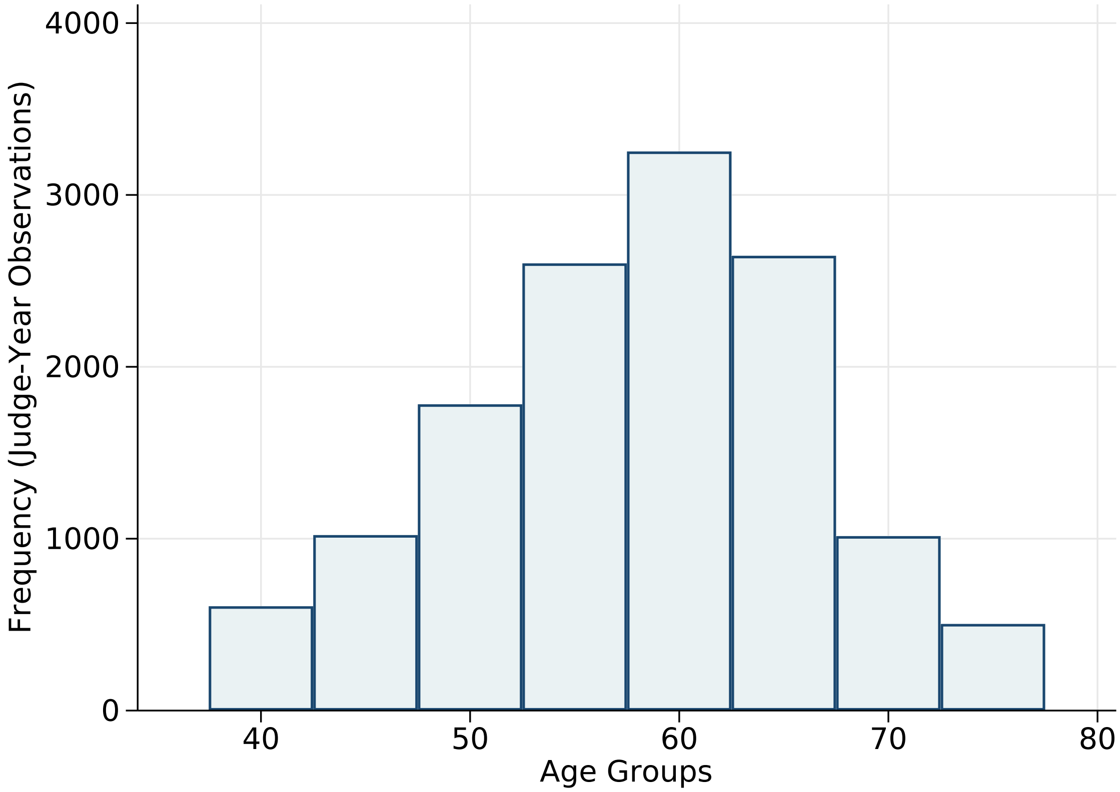
Table B.12: Effect of Mandatory Retirement Reform, Other Behavioral Dimensions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Work Output		Caselaw Research		Overruled Rate		Addendum Ops	
Retirement Reform	0.0310	0.0455	-2.557	-0.211	-0.0195**	-0.00439	0.283**	0.231**
	(0.0643)	(0.0592)	(1.718)	(1.433)	(0.00589)	(0.00593)	(0.0699)	(0.0769)
Year FE, Court FE	X	X	X	X	X	X	X	X
Court Trends/Windows		X		X		X		X
N	15010	15010	15010	15010	15010	15010	15010	15010
R-sq	0.266	0.379	0.456	0.604	0.100	0.143	0.401	0.444

Observation is a judge working in a year. “Retirement Reform” is an indicator for the ten years after the introduction of mandatory retirement. “Work Output” is log number of words written in a year. “Caselaw Research” is number of previous cases cites. “Overruled rate” is being overruled by a higher court. “Addendum Ops” is number of dissenting and concurring opinions (in logs). Court Treat Windows means court-specific treatment windows (ten years before and after reform). Standard errors clustered by state in parentheses. + p<.0.1, * p<0.05, ** p<0.01.

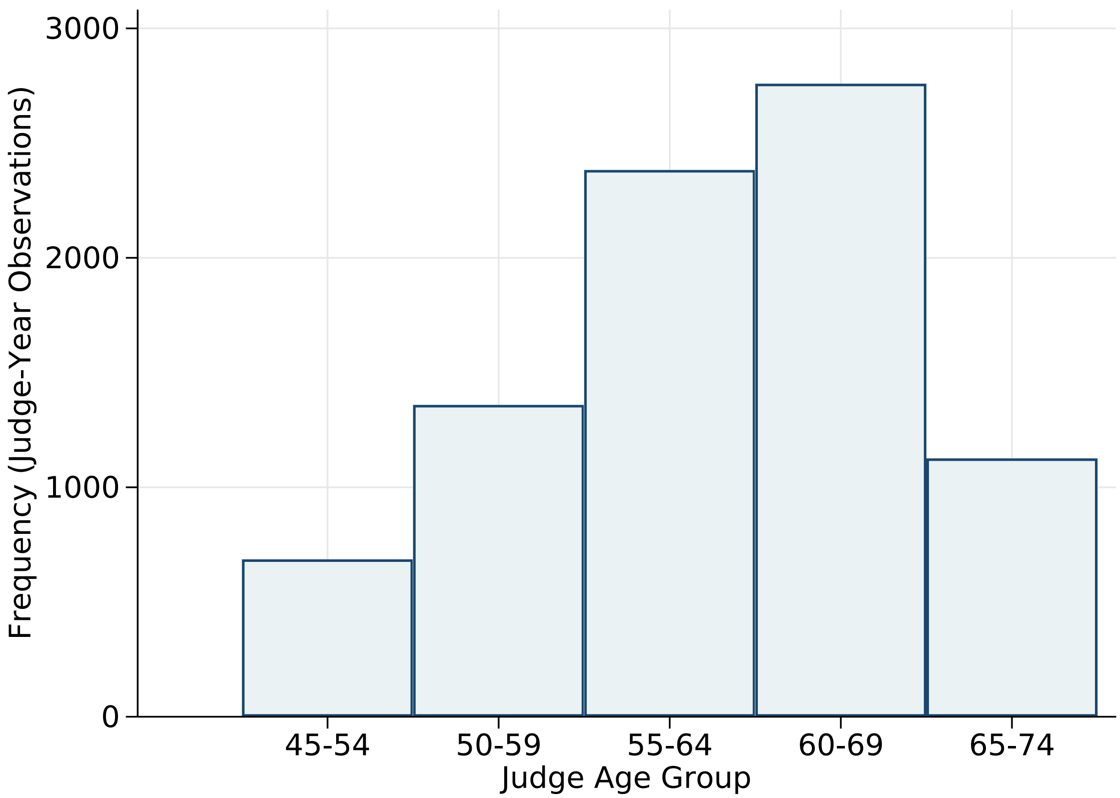
C Additional Material on Aging and Performance

Figure C.1: Distribution over Age Groups for Life Cycle Coefficient Plots



Notes. Number of judge-year observations in each five-year age group for the life cycle coefficient plots.

Figure C.2: Distribution over Age Groups for Balanced Sample Analysis



Notes. Histogram of age groups for balanced sample analysis. Number of judge-years in each overlapping ten-year balanced sample.

Table C.1: Judge Age and Inverse Hyperbolic Sine of Citations

	(1)	(2)	(3)	(4)
	Inverse Hyperbolic Sine of Positive Cites			
Judge Age (Years)	-0.00831**	-0.00821**	-0.00714**	0.0327*
	(0.00138)	(0.00111)	(0.00125)	(0.0128)
Age Squared				-0.000335**
				(0.000114)
Court-Year FE	X	X	X	X
First-Year Baseline		X	X	X
Cohort FE / Trends			X	X
N	14981	14981	14981	14981
R-sq	0.670	0.695	0.700	0.701

Observation is a judge working in a year. Outcome is inverse hyperbolic since (asinh) of positive cites to a judge in a year. Court-Year FE is interacted court-year fixed effects. First-Year Baseline means a judge's value for the outcome in their first year on the court is included as a control. Cohort FE means fixed effect for decade that the judge started on the court. Cohort Trends means judge starting-year interacted with court fixed effect. Standard errors clustered by state in parentheses. + p<.0.1, * p<0.05, ** p<0.01.

Table C.2: Judge Age and Additional Measures of Performance, Logs

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Cites in 10 Years		All Cites		Discuss Cites		Out-of-State Cites	
Judge Age	-0.00695**	0.0348**	-0.00747**	0.0321*	-0.00844**	0.0338**	-0.00882**	0.0292*
	(0.00126)	(0.0121)	(0.00122)	(0.0130)	(0.00112)	(0.0110)	(0.00136)	(0.0131)
Age Squared		-0.000351**		-0.000333**		-0.000355**		-0.000320**
		(0.000107)		(0.000115)		(0.0000970)		(0.000115)
Court-Year FE	X	X	X	X	X	X	X	X
1st-Year Base	X	X	X	X	X	X	X	X
Cohort FE/Trend	X	X	X	X	X	X	X	X
N	14981	14981	14981	14981	14981	14981	14981	14981
R-sq	0.771	0.772	0.697	0.698	0.690	0.691	0.669	0.670

Observation is a judge working in a year. Outcomes are in logs. "Cites in 10 years" is log of positive cites to a judge in a year, within ten years of a case. "All Cites" includes negative and distinguishing (not just positive) cites. "Discuss cites" means the case was positively discussed and applied. "Out-of-state cites" means citations from courts in other states. Court-Year FE is interacted court-year fixed effects. 1st-Year Base means a judge's value for the outcome in their first year on the court is included as a control. Cohort FE means fixed effect for decade that the judge started on the court. Cohort Trends means judge starting-year interacted with court fixed effect. Standard errors clustered by state in parentheses. + $p<.0.1$, * $p<0.05$, ** $p<0.01$.

Table C.3: Effect of Reform on Citations: Additional Measures

	(1)	(2)	(3)	(4)	(5)	(6)
	Cites After 10 years		Has Cite		Drop 5% Outliers	
Judge Age	-0.00786**	0.0284*	-0.000410**	0.00114	-0.00764**	0.0377**
	(0.00117)	(0.0126)	(0.000131)	(0.00122)	(0.00137)	(0.0127)
Age Squared		-0.000305**		-0.0000130		-0.000381**
		(0.000111)		(0.0000107)		(0.000112)
Court-Year FE	X	X	X	X	X	X
1st-Year Base	X	X	X	X	X	X
Cohort FE/Trend	X	X	X	X	X	X
N	14981	14981	14981	14981	14726	14726
R-sq	0.654	0.655	0.927	0.927	0.707	0.708

Notes. Observation is a judge working in a year. “Retirement Reform” is an indicator for the ten years after the introduction of mandatory retirement. “Cites after 10 Years” means the log of the positive cites to a judge’s cases in a year, from cases more than ten years later. “Has Cite” means the proportion of cases with at least one positive citation. “Drop 5% Outliers” is the baseline outcome (log of positive cites to a judge’s cases in a year) but for each court-year, the top 5% of cases by cite count are dropped. Court-Year FE is interacted court-year fixed effects. 1st-Year Base means a judge’s value for the outcome in their first year on the court is included as a control. Cohort FE means fixed effect for decade that the judge started on the court. Cohort Trends means judge starting-year interacted with court fixed effect. Standard errors clustered by state in parentheses. + p<.0.1, * p<0.05, ** p<0.01.

Table C.4: Judge Age and Additional Performance Measures, Rank Percentiles

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Cites in 10 Years		All Cites		Discuss Cites		Out-of-State Cites	
Judge Age	-0.00445**	0.0190+	-0.00460**	0.0172+	-0.00498**	0.0182*	-0.00366**	0.00134
	(0.000878)	(0.00981)	(0.000815)	(0.00880)	(0.000812)	(0.00691)	(0.000685)	(0.00749)
Age Squared		-0.000198*		-0.000184*		-0.000196**		-0.0000422
		(0.0000856)		(0.0000771)		(0.0000600)		(0.0000661)
Court-Year FE	X	X	X	X	X	X	X	X
First-Year	X	X	X	X	X	X	X	X
Cohort FE / T's	X	X	X	X	X	X	X	X
N	13655	13655	8181	13655	13655	13655	13655	13655
R-sq	0.768	0.769	0.195	0.697	0.698	0.690	0.691	0.674

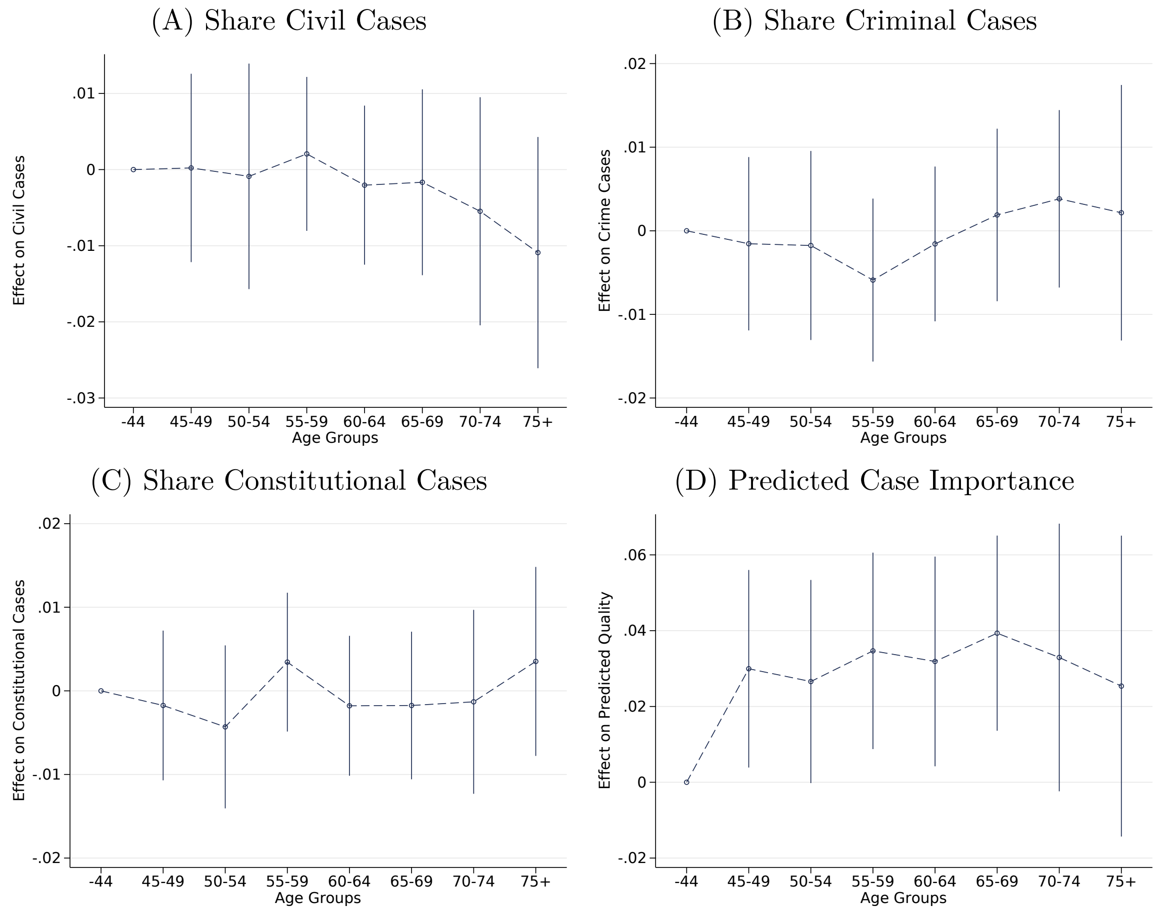
Observation is a judge working in a year. Outcomes are in rank percentiles. "Cites in 10 years" is log of positive cites to a judge in a year, within ten years of a case. "All Cites" includes negative and distinguishing (not just positive) cites. "Discuss cites" means the case was positively discussed and applied. "Out-of-state cites" means citations from courts in other states. Court-Year FE is interacted court-year fixed effects. 1st-Year Base means a judge's value for the outcome in their first year on the court is included as a control. Cohort FE means fixed effect for decade that the judge started on the court. Cohort Trends means judge starting-year interacted with court fixed effect. Standard errors clustered by state in parentheses. + p<.0.1, * p<0.05, ** p<0.01.

Table C.5: Case Type and Importance by Judge Age and Case Allocation Rule

	(1)	(2)	(3)	(4)	(5)
	Crim Cases	Civil Cases	Admin Cases	Con Law Cases	Pred. Cites
Age × Random	0.00427	-0.00435	-0.0164	-0.0196	-0.00127
	(0.00845)	(0.00700)	(0.0115)	(0.0129)	(0.00188)
Age × Not Rand	0.0265	-0.0198	-0.00161	-0.0131	-0.0000133
	(0.0209)	(0.0230)	(0.0176)	(0.0194)	(0.00229)
Court-Year FE	X	X	X	X	X
N	13643	13643	13607	13632	13599
adj. R-sq	0.140	0.209	-0.062	-0.042	0.397

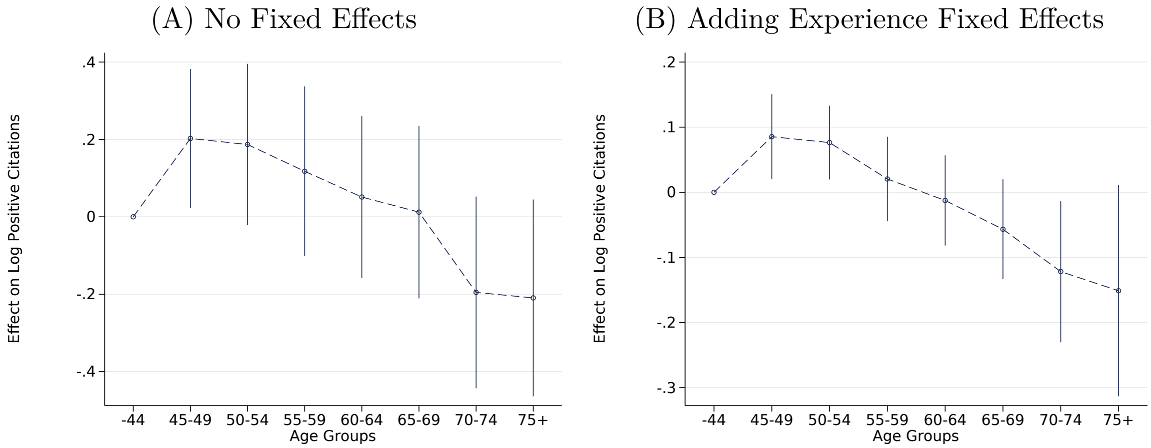
Observation is a judge working in a year. "Random" is an indicator for random-assignment states, while "Not Rand" means discretionary assignment. "Crim Cases" means proportion of cases on criminal law in a year (and respectively for civil cases, administrative cases, and constitutional law cases). "Pred. Cites" means the predicted case quality computed from an OLS regression with case characteristics (legal area and related industries). Standard errors clustered by state in parentheses. + p<.0.1, * p<0.05, ** p<0.01.

Figure C.3: Dynamic Analysis: Case Types



Dynamic coefficient plots for estimates of five-year age group differences, relative to the age < 45 group. Observation is a judge working in a year. All graphs contain court-year interacted fixed effects, first year baselines, and cohort fixed effects. 95% confidence intervals constructed using standard errors clustered by state.

Figure C.4: Judge Age and Judge Performance, Additional FE Specifications



Dynamic coefficient plots for estimates of five-year age group differences, relative to the age < 45 group. Outcom is log positive cites. Observation is a judge working in a year. Panel A has no fixed effects; Panel B contains court-year interacted fixed effects, first year baselines, cohort fixed effects, and fixed effects for number of years on court. 95% confidence intervals constructed using standard errors clustered by state.

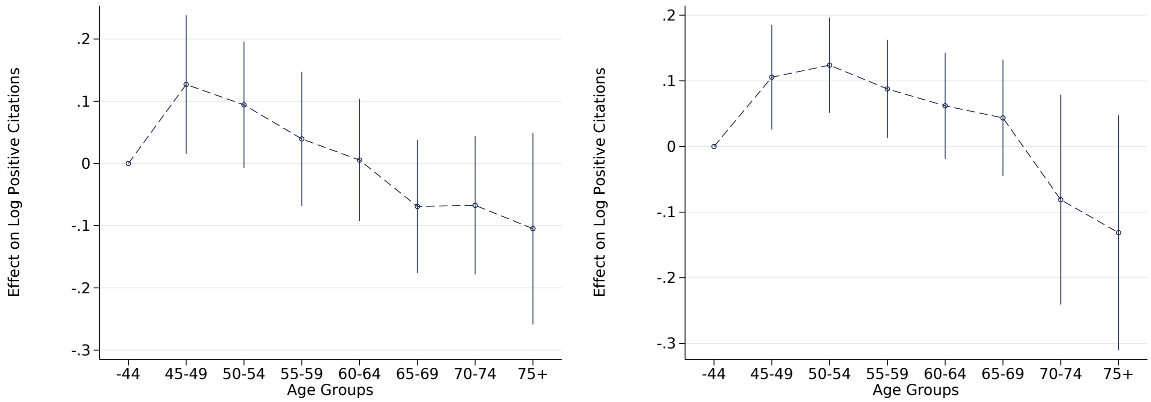
Table C.6: Judge Age and Judge Performance, by Retirement Rule

	(1)	(2)	(3)	(4)
	Voluntary Retirement		Mandatory Retirement	
	Log Cites	Rank	Log Cites	Rank
Judge Age (Years)	-0.00962**	-0.00576**	-0.00615**	-0.00390**
	(0.00320)	(0.00147)	(0.00160)	(0.00122)
Court-Year FE	X	X	X	X
First-Year Baseline	X	X	X	X
Cohort FE / Trends	X	X	X	X
N	4688	4688	8967	8967
R-sq	0.692	0.059	0.613	0.043

Observation is a judge working in a year. "Log Positive Cites" is log of positive cites to a judge in a year. "Rank Percentile Cites" means judges are uniformly distributed between zero and one based on number of positive citations within court-year (0 is lowest, 1 is highest). Court-Year FE is interacted court-year fixed effects. First-Year Baseline means a judge's value for the outcome in their first year on the court is included as a control. Cohort FE means fixed effect for decade that the judge started on the court. Cohort Trends means judge starting-year interacted with court fixed effect. Standard errors clustered by state in parentheses. + p<.0.1, * p<0.05, ** p<0.01.

Figure C.5: Judge Age and Judge Performance, Before/After 1970

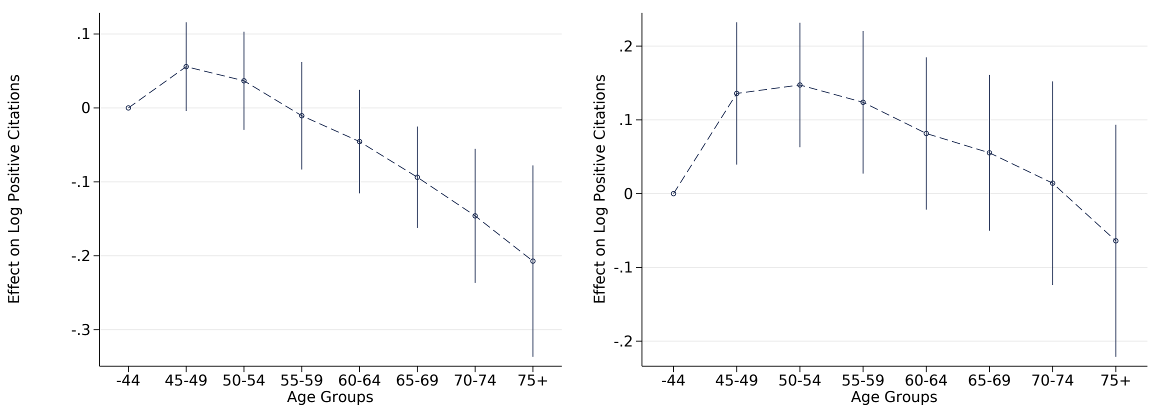
(A) Before 1970 — (B) After 1970



Dynamic coefficient plots for estimates of five-year age group differences, relative to the age < 45 group. Observation is a judge working in a year. All graphs contain court-year interacted fixed effects, first year baselines, and cohort fixed effects. Outcomes are log positive cites to a judge in a year. 95% confidence intervals constructed using standard errors clustered by state.

Figure C.6: Judge Age and Judge Performance, Alternative Weighting

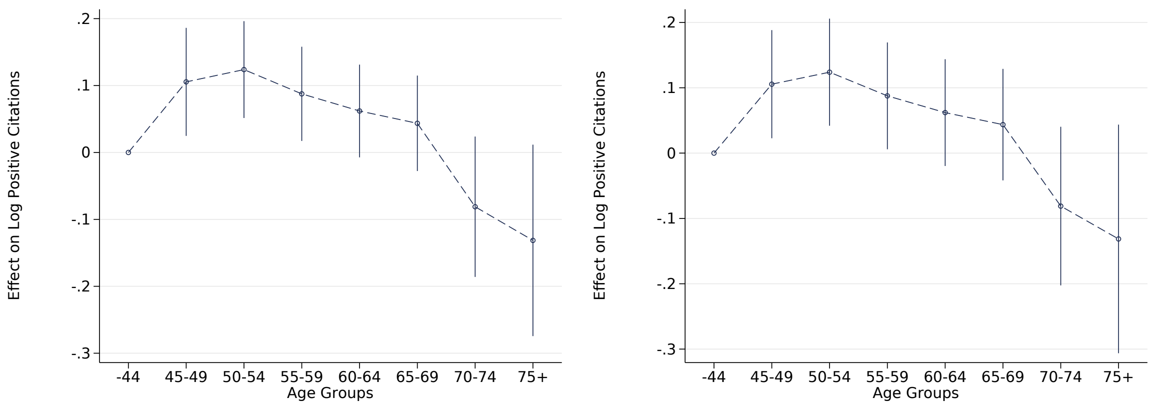
(A) Weighting by Number of Opinions — (B) Weighting by Inverse Career Length



Observation is a judge working in a year. All graphs contain court-year fixed effects. "With Controls" means a baseline for a judge's first year on the court, fixed effects for decade that the judge started on the court, and judge starting-year interacted with court fixed effect.

Figure C.7: Judge Age and Judge Performance, Alternative Clustering

(A) Clustering at State-Year Level | (B) Clustering at Judge Level



Observation is a judge working in a year. All graphs contain court-year fixed effects. “With Controls” means a baseline for a judge’s first year on the court, fixed effects for decade that the judge started on the court, and judge starting-year interacted with court fixed effect.

Table C.7: Additional Outcomes for Life Cycle Analysis

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Words Written	Affirm rate	Case Length	Research	Dissent	Overruled	Superseded
Judge Age	-0.00134	-0.00594	0.00154	-1.608	-0.00458*	0.00133	-0.00306**
	(0.00132)	(0.0217)	(0.00109)	(1.043)	(0.00182)	(0.000975)	(0.000883)
Court-Year FE	X	X	X	X	X	X	X
Cohort FE/Trends	X	X	X	X	X	X	X
N	13655	13655	13655	13655	13655	13655	13655
R-sq	0.563	0.747	0.719	0.607	0.504	0.328	0.464

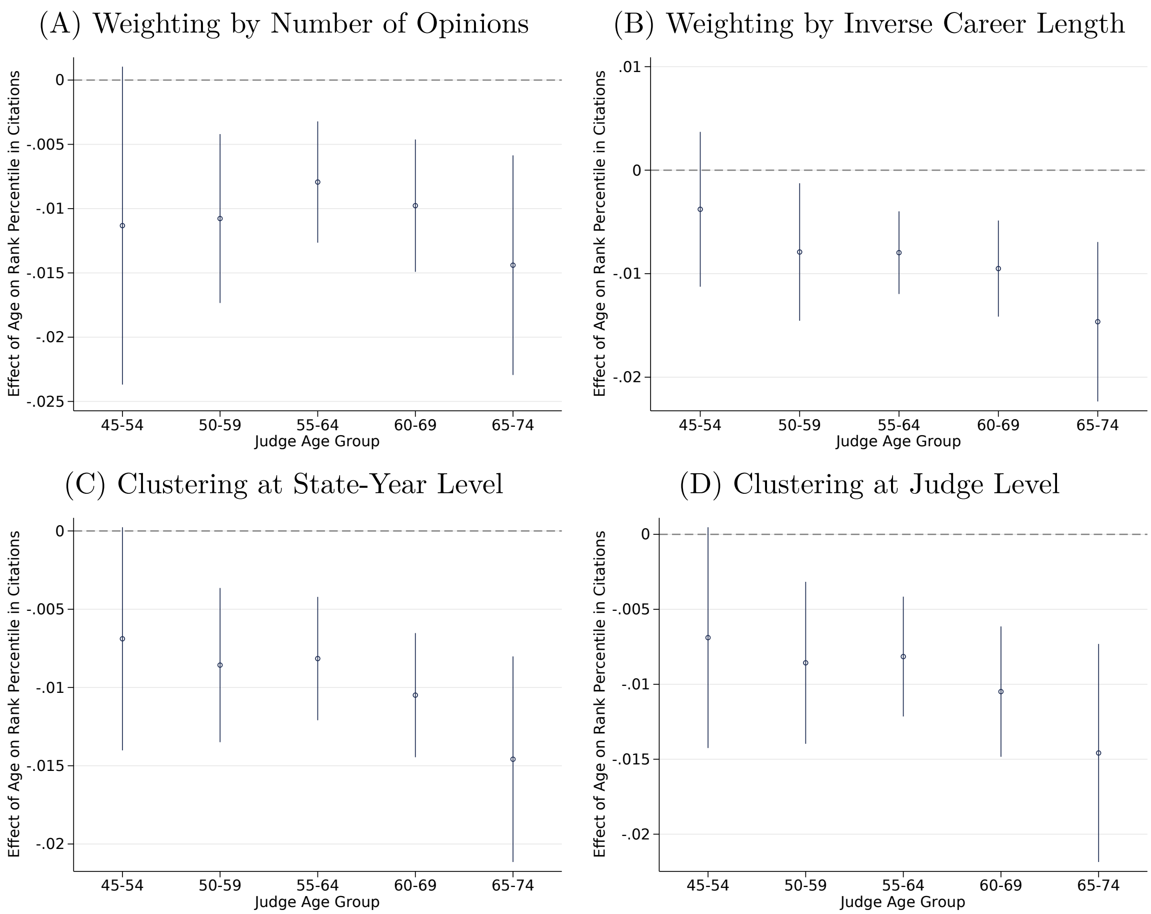
Notes. Observation is a judge working in a year. “Words Written” is log of number of words written in opinions by a judge in a year. “Affirm Rate” is the rate that a judge affirms (rather than reverses) a lower court. “Case length” is words per opinion written. “Research” is previous cases cited in references. “Dissent” is rate of dissenting against other judges. “Overruled” is rate of being overruled by a higher court. “Superseded” is rate of being reversed by legislation. Court-Year FE is interacted court-year fixed effects. First-Year Baseline means a judge’s value for the outcome in their first year on the court is included as a control. Cohort FE means fixed effect for decade that the judge started on the court. Cohort Trends means judge starting-year interacted with court fixed effect. Standard errors clustered by state in parentheses. + p<.0.1, * p<0.05, ** p<0.01.

Table C.8: Instrumenting Age with Reforms

	(1)	(2)	(3)	(4)	(5)
	Effect on Log Positive Cites per Judge-Year				
Judge Age	-0.0916**	-0.139**	-0.0766**	-0.0893**	-0.0888**
	(0.0278)	(0.0395)	(0.0232)	(0.0249)	(0.0244)
Court FE, Year FE	X	X	X	X	X
Court Trends/Windows		X	X	X	X
Init Court Rules × Year FE			X	X	X
Init Case Types × Year FE				X	X
Init Age × Year FE					X
Cragg-Donald F-stat	44.526	39.694	43.803	46.821	45.943
Kleibergen-Paap F-stat	10.372	7.416	17.520	16.991	21.625
N	15010	15010	15010	15010	15010
R-sq	0.460	0.526	0.538	0.555	0.565

Notes. 2SLS estimates for effect of age on performance, instrumenting with the retirement reforms. Observation is a judge working in a year. "Retirement Reform" is an indicator for the ten years after the introduction of mandatory retirement. Court Treat Windows means court-specific treatment windows (ten years before and after reform). "Init X" × year FE means initial values are interacted with year. "Init Court Rules" includes a state's 1947 rules for judge selection/retention system, admin office, intermediate appellate court, number of judges, and term length. "Init Case Types" includes a court's 1947 average values for case characteristics (legal area and related industries). "Init Age" includes the initial mean and standard deviation for judge age on the court. Standard errors clustered by state in parentheses. + p<.0.1, * p<0.05, ** p<0.01.

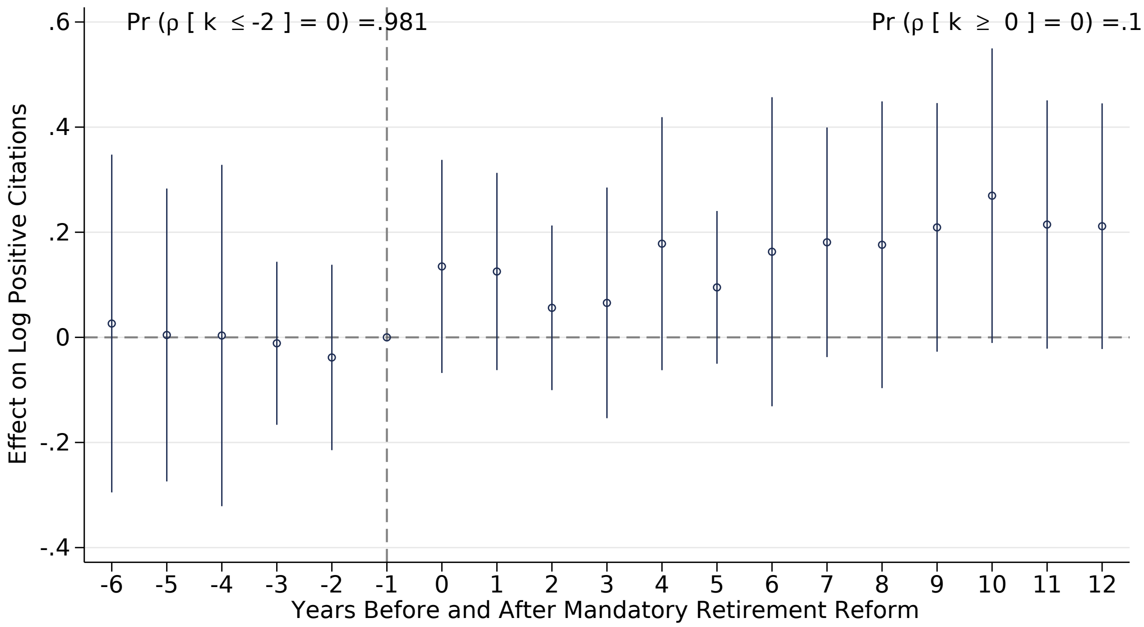
Figure C.8: Balanced Judge Samples: Alternative Weighting and Clustering



Performance-Age estimates for separate balanced samples of judges based on age group. Observation is a judge working in a year. Outcomes are in rank percentiles, regressed on age (in years) for the specified group. 95% confidence intervals constructed using standard errors clustered by state.

D Additional Material on Team Effects of Aging

Figure D.1: Event-Study Effect of Reform on Performance, with Judge Fixed Effects



Judge performance before and after reforms implementing retirement ages of 70, 72 or 75. Outcome is log positive citations for a judge in a year. Time series is a coefficient plot from the event study regression (2), with coefficents estimated relative to the year before the reform. Regression includes court and year fixed effects and court-specific event windows, plus judge fixed effects. 95% confidence intervals constructed with standard errors clustered by state.

Table D.1: Effect of Reform, Other Quality Measures, Judge Fixed Effects

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Cites in Levels		Within 10 years		All Cites		Discuss Cites		Out-of-State Cites	
Ret. Reform	37.25*	18.28	0.238**	0.177	0.167*	0.131	0.109+	0.123	0.155*	0.167
	(17.63)	(17.56)	(0.0869)	(0.122)	(0.0721)	(0.0965)	(0.0543)	(0.0849)	(0.0772)	(0.105)
Year / Judge FE	X	X	X	X	X	X	X	X	X	X
Trends/Windows		X		X		X		X		X
N	14905	14905	14905	14905	14905	14905	14905	14905	14905	14905
R-sq	0.585	0.667	0.739	0.809	0.678	0.753	0.667	0.749	0.641	0.718

Notes. Observation is a judge working in a year. "Retirement Reform" is an indicator for the ten years after the introduction of mandatory retirement. "Cites in Levels" means the outcome is not logged. "Within 10 years" is the log positive cites within ten years of an opinion. "All Cites" is the log number of all citations (positive, negative, and distinguishing) to a judge in a year. "Discuss Cites" is only the positive cites where the latter judge discussed the cited opinion. "Out-of-State Cites" is the count of number of positive citations from courts in other states. "Positive Cites" is the number of positive cites (in levels). Court Treat Windows means court-specific treatment windows (ten years before and after reform). Standard errors clustered by state in parentheses. + p<.0.1, * p<0.05, ** p<0.01.

Table D.2: Mandatory Retirement Reform: Relevance of Authorship Rates

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Authorship Rate		Log Pos. Cites		Log # Opinions		Log Pos. Cites / Case	
Retirement Reform	0.0638	0.0917**	0.206*	0.213*	0.140*	0.0933*	0.0451	0.0855+
	(0.0383)	(0.0304)	(0.0838)	(0.0908)	(0.0549)	(0.0462)	(0.0464)	(0.0480)
Authorship Rate	-	-	0.406**	0.442*	0.526*	0.546**	-0.150	-0.117
			(0.151)	(0.180)	(0.197)	(0.159)	(0.115)	(0.0823)
Year FE, Court FE	X	X	X	X	X	X	X	X
Court Trends/Windows		X		X		X		X
N	2357	2357	15010	15010	15010	15010	15010	15010
R-sq	0.642	0.826	0.465	0.529	0.352	0.522	0.651	0.710

In Columns 1 and 2, an observation is court-year, while in Columns 3 through 8, an observation is a judge-year. "Retirement Reform" is an indicator for the ten years after the introduction of mandatory retirement. "Authorship Rate" is the proportion of total cases that have a signed author in the court-year. Court Treat Windows means court-specific treatment windows (ten years before and after reform). Standard errors clustered by state in parentheses. + p<.0.1, * p<0.05, ** p<0.01.

Table D.3: Comparing Judges Selected Before/After the Reform

	(1) Starting Age	(2) Log Cites / Case	(3) Rank Cites / Case
Selected Post Reform	-0.963	0.136**	0.186**
	(1.464)	(0.0240)	(0.0309)
Court × Year FE	X	X	X
N	2186	2186	2186
R-sq	0.272	0.723	0.029

Observation is a judge working in a year. Sample includes judges selected within ten years before and after the reform. "Selected Post Reform" is an indicator for judges selected within ten years after the introduction of mandatory retirement. Standard errors clustered by state in parentheses. + p<.0.1, * p<0.05, ** p<0.01.

One question not addressed directly in the main text is whether the mandatory retirement effect is driven by selection or incentives. On the selection side, the performance gain comes from the replacement of older judges with new, younger, better-performing judges. On the incentive side, mandatory retirement might increase quality for sitting judges by changing their incentives or their work environment. Incentive effects could include the need to impress the chief justice so that one can stay on as a senior justice, or else to maintain skills for obtaining non-judicial employment outside the court. Work-environment changes could be due to spillovers from the replacement of older judges with younger judges, so judges that stay on have to do less to support the older judges. Our previous estimates have not disentangled the across-judge from the within-judge effects of retirement reforms.

To get at selection, we compare judges working on the same court at the same time, but selected before/after the reforms. These regressions use court-year fixed effects so hold time-varying factors constant. The results, reported in Appendix Table D.3, show that the reforms did not significantly affect the starting age of judges. However, the judges selected after the reforms tended to be high-quality, as measured by citations per opinion.

E The Age Discrimination in Employment Act of 1967 Sec. 621

The Congress hereby finds and declares that

1. in the face of rising productivity and affluence, older workers find themselves disadvantaged in their efforts to retain employment, and especially to regain employment when displaced from jobs;

   (a) the setting of arbitrary age limits regardless of potential for job performance has become a common practice, and certain otherwise desirable practices may work to the disadvantage of older persons;

   (b) the incidence of unemployment, especially long-term unemployment with resultant deterioration of skill, morale, and employer acceptability is, relative to the younger ages, high among older workers; their numbers are great and growing; and their employment problems grave;

   (c) the existence in industries affecting commerce, of arbitrary discrimination in employment because of age, burdens commerce and the free flow of goods in commerce.

   (d) It is therefore the purpose of this chapter to promote employment of older persons based on their ability rather than age; to prohibit arbitrary age discrimination in employment; to help employers and workers find ways of meeting problems arising from the impact of age on employment.