DEFERENCE REALITIES: JUDICIAL DEFERENCE AND LITIGATION OUTCOMES IN THE APPELLATE REVIEW ERA

EDWARD H. STIGLITZ[*]

ABSTRACT

The Supreme Court recently overturned Chevron*, potentially re-shaping the relationship between courts and administrative agencies. Observers were quick to fête or mourn the decision. However, the relationship between deference law and agency litigation outcomes is unclear. It is possible that deference law is immaterial to agency litigation outcomes.*

This Article undertakes a sweeping historical-empirical examination of the relationship between deference regimes and agency litigation outcomes. The analysis considers three distinct candidate periods: an early deference regime, which runs from the birth of federal-question jurisdiction in 1875 to the judicial revolution of 1937; a transitional deference regime, which emerged in the aftermath of that revolution and persisted until the Court's seminal Chevron *decision in 1984; and the* Chevron *deference regime, which governed judicial review of agency actions for nearly four decades. A novel dataset of approximately 500,000 published cases from the Supreme Court and Courts of Appeals provides the foundation for this research. That dataset is assessed via a pipeline of validated, fine-tuned, and domain-adapted large language models, built on months of careful annotation by domain experts.*

There are many threats to straightforward inferences about the consequences of deference law for agency win-rates. Using a theoretical model, the Article first lays out what can and cannot be learned about deference law through the study of win-rates. More can be learned than might first appear, and the fundamental strategy of the analysis is to account for what can be accounted for, and to calibrate understanding of what cannot be accounted for. The results do not reveal a consistent association between deference regime and the success of agencies in court. The shift from one deference regime to another—for example, from the transitional regime to the Chevron *regime—does not appear to be*

* Cornell University. I thank participants at workshops or conferences at Cornell University, Chicago Law School, Yale Law School, and ALEA for helpful comments, as well as very helpful criticism from Cary Coglianese and Matthew Stephenson. I am extremely grateful for the careful work of a team of research assistants: Garry Blum, Alex Strohl, Julia Risi, Jamie Li, Akshey Mulpuri, and Whitney Dawson.

accompanied by substantial changes in the trajectory of agency litigation outcomes.

The later parts of this Article argue for a re-orientation toward deference realities, including the bureaucratic constraints of judicial administration. The study proposes orientation toward historical expansions of judicial clerkships and judgeships as sources of variation in judicial-bureaucratic capacity. Such deference realities may both inform predictions of the nascent Loper Bright *regime and suggest a research agenda focused on institutional levers rather than doctrinal formulations.*

TABLE OF CONTENTS

INTRODUCTION

The Court overturned a cornerstone of the American legal landscape in 2024, holding that *Chevron* deference was inconsistent with the

Administrative Procedure Act (APA).[1] Proponents herald this as a righting to the proper judicial role.[2] Critics regard this as a troubling development and worry that it will substantially rework the operation of our government, effectively hobbling administrative agencies.[3] Administrative bodies carry out critical public responsibilities—safeguarding clean water, food quality, labor standards, the integrity of financial institutions[4]—and the critics, if right, sound a justified alarm.

But what is the relationship between deference regimes and litigation outcomes for agencies? On the one hand, from the earliest days of the *Chevron* revolution, skeptical observers questioned whether the doctrine affected the success of agencies before courts. They reasoned either that *Chevron* changed little, merely summarizing existing doctrine, even if doing so more cleanly than usual;[5] or that the doctrine, though an innovation, was sufficiently pliable so that jurists would not be constrained by it.[6] On the other hand, legal observers and litigants themselves lavish attention on the doctrine, making it one of the most cited fixtures in legal debates[7]—its passing variously mourned or fêted. It is hard to rationalize that attention if the doctrine is immaterial to litigation outcomes. Yet there is little direct evidence that *Chevron* or other deference regimes matter.[8]

---

1. Loper Bright Enters. v. Raimondo, 603 U.S. 369 (2024). For a helpful assessment of scholarly predictions regarding the decision, see Cary Coglianese & Daniel E. Walters, *The Great Unsettling: Administrative Governance After* Loper Bright, 77 ADMIN. L. REV. 1 (2025).

2. *E.g.*, Editorial, *Two Big Victories for Liberty at the Supreme Court*, WALL ST. J. (June 28, 2024), https://www.wsj.com/articles/supreme-court-chevron-deference-loper-bright-jan-6-fischer-d5958b01 [https://perma.cc/2ZKK-B3RR].

3. *E.g.*, Kate Shaw, Opinion, *The Imperial Supreme Court*, N.Y. TIMES (June 29, 2024), https://www.nytimes.com/2024/06/29/opinion/supreme-court-chevron-loper.html [https://perma.cc/E5YM-5Y9Z] (arguing that the decision will "fundamentally transform" the delivery of critical services to Americans);
Steve Vladeck, Opinion, *The Most Aggressive Restructuring of Government in Almost 90 Years*, CNN (July 2, 2024), https://www.cnn.com/2024/07/02/opinions/supreme-court-radically-restructures-government-vladeck [https://perma.cc/5TLZ-FLF6] (arguing that *Loper Bright* and other actions by the Court represent "the most aggressive restructuring of the federal government in our lifetimes").

4. *E.g.*, EDWARD H. STIGLITZ, THE REASONING STATE 1–20 (2022).

5. For instance, a leading casebook in administrative law titles the heading for its section on the *Chevron* decision as "*Chevron*: Synthesis or Revolution?" STEPHEN G. BREYER, RICHARD B. STEWART, CASS R. SUNSTEIN, ADRIAN VERMUELE & MICHAEL E. HERZ, ADMINISTRATIVE LAW AND REGULATORY POLICY 243 (9th ed. 2022). In line with this heading, Thomas Merrill summarized his review of the justices' papers: "There is no evidence that Justice Stevens understood his handiwork in *Chevron* as announcing fundamental changes in the law of judicial review. . . . Nor is there any evidence that Justice Stevens's colleagues on the Court perceived *Chevron* as some kind of watershed decision, either when it was decided or for some time afterward." Thomas W. Merrill, *The Story of* Chevron*: The Making of an Accidental Landmark*, 66 ADMIN. L. REV. 253, 275–76 (2014).

6. *E.g.*, Thomas W. Merrill, *Judicial Deference to Executive Precedent*, 101 YALE L.J. 969, 971–80 (1992) (referring to *Chevron* as a "revolution on paper").

7. According to Google Scholar, at the time of writing, scholars have written over 18,000 articles that cite *Chevron* (search term: "467 U.S. 837").

8. *See infra* Section II.A for a review of the relevant existing literature.

The main parts of this Article conduct a novel and sweeping historical-empirical investigation of the relationship between deference regimes and litigation outcomes. It covers cases from the Supreme Court and the Courts of Appeals since the birth of federal-question jurisdiction in 1875—in total, the analysis considered approximately 500,000 published cases—with a focus starting in the appellate era.[9] Until recently, a credible analysis of this volume of cases would not have been possible. However, with the latest generation of transformer-based language models, algorithmic performance is human-competitive on many tasks.[10] Building on a months-long human annotation project by domain experts, this analysis develops a pipeline of fine-tuned, validated models to assess the litigation outcomes of agencies before lower and appellate federal courts over the last one hundred and fifty years.

The history of deference does not start with *Chevron*—nor, as the Court recently assured us, does it end with it. The starting point for this analysis is Congress's passage of the Judiciary Act of 1875, which gave federal-question jurisdiction to federal courts.[11] This act moved courts off a mandamus footing for review of agency actions and toward the classes of review that would be familiar to a modern court, coalescing in the appellate review framework in the 1920s.[12] The story of deference from that point is contested, but scholars and jurists tend to divide the span of time into three periods: a first period lasting until approximately the judicial revolution of 1937;[13] a second period lasting from the judicial revolution to the Court's *Chevron* decision in 1984;[14] and a third period from *Chevron* to the Court's *Loper Bright* decision. A fourth period is our young, current *Loper Bright* regime. As explained later, scholars and jurists debate the nature of deference afforded by courts to agencies in these periods—for example, was

9. Thomas W. Merrill, *Article III, Agency Adjudication, and the Origins of the Appellate Review Model of Administrative Law*, 111 COLUM. L. REV. 939, 939 (2011) (noting that the appellate review model, which is the basis of modern administrative law, had firmly established roots by the 1920s).

10. For other examples of a sweeping historical inquiry made possible by these algorithms, see Edward H. Stiglitz & Rosamond Thalken, *Historical Trends in Macro-Jurisprudence: A Language Model Assessment, 1870–2023*, 84 MD. L. REV. 46 (2024); Rosamond Thalken, Edward H. Stiglitz, David Mimno & Matthew Wilkens, *Modeling Legal Reasoning: LM Annotation at the Edge of Human Agreement*, *in* PROCEEDINGS OF THE 2023 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING 9252 (2023).

11. John F. Duffy, *Administrative Common Law in Judicial Review*, 77 TEX. L. REV. 113, 126 (1998); Aditya Bamzai, *The Origins of Judicial Deference to Executive Interpretation*, 126 YALE L.J. 908, 913 (2017).

12. *See* Bamzai, *supra* note 11, at 913.

13. *See, e.g.*, *id.* at 965 (referring to the "new jurisprudence of deference that emerged in the early 1940s").

14. As suggested by its high citation counts in both caselaw and academic articles, *Chevron* is taken as a doctrinal marker by many observers. *See, e.g.*, Merrill, *supra* note 5, at 254 (reviewing the citation counts and noting that the case "[took] the legal world by storm").

the period after 1875 one of *de novo* review, with limited exceptions, or one in which a norm of deference had already taken root?—and this analysis will regard them as candidate doctrinal periods, roughly hewing to the characterization in *Loper Bright* itself.

There, Justice Roberts described the period from 1875 to the judicial revolution as one in which courts often granted "respect" to administrative views, at least when an "interpretation was issued roughly contemporaneously with enactment of the statute and remained consistent over time."[15] However, though the views of the executive might "inform the judgment of the Judiciary, [they] did not supersede it."[16] Thus, unlike on questions of fact, where courts might be partially bound by executive fact-finding, courts would exercise their own judgment on questions of law.[17] This period may be referred to as the *early* deference regime.

A second period begins shortly after the judicial revolution of 1937 and runs to *Chevron*. During this period, jurists and scholars observe a strand of cases in which courts regard agency legal determinations as binding. The most salient case is perhaps *Gray v. Powell*, in which the Court said where "a determination has been left [by Congress] to an administrative body, this delegation will be respected and the administrative conclusion left untouched."[18] What is contested is how dominant that strand of cases was, and what conditions need to apply before courts would defer to agency determinations. Some observers see the Court fundamentally, if incompletely, altering its approach to judicial review shortly after the judicial revolution. In *Loper Bright*, the Court acknowledged this line of cases, but it limited their importance: They were "cabined to factbound determinations" and they did not "refashion" the approach to judicial review.[19] This period may be referred to as a *transitional* deference regime—one in which courts deferred to agencies, if incompletely and episodically.

The third period runs from *Chevron* itself to *Loper Bright*. Under *Chevron*, if a court determines that "the statute is silent or ambiguous with respect to the specific issue," it must defer to the agency's interpretation if "based on a permissible construction of the statute."[20] That form of

15. Loper Bright Enters. v. Raimondo, 603 U.S. 369, 386 (2024).

16. *Id.*

17. *Id.* at 387 (noting that during this early period Congress could give agencies the "power to make *findings of fact* which are conclusive" (citation omitted)).

18. Gray v. Powell, 314 U.S. 402, 412 (1941). Another commonly cited case is *NLRB v. Hearst Publications, Inc.*, 322 U.S. 111 (1944). These cases receive more attention in Part I.

19. *Loper Bright*, 603 U.S. at 389.

20. Chevron U.S.A. Inc. v. Nat. Res. Def. Council, Inc., 467 U.S. 837, 843 (1984), *overruled by Loper Bright*, 603 U.S. at 412–13.

deference, the Court maintained in *Loper Bright*, represented a "marked departure" from what came before it and was inconsistent with the language of the APA.[21] Unlike earlier regimes, as the Court saw it, the *Chevron* deference regime "demands that courts mechanically afford *binding* deference to agency interpretations, including those that have been inconsistent over time."[22]

A main task of this analysis is to examine agency litigation outcomes in these three putative periods of deference law. The fourth, *Loper Bright* period is a question mark. This historical analysis cannot tell us precisely how deference will evolve in the post-*Chevron* era, but it can tell us something of the constants and variables in the relationship between agency action and judicial review. It can inform how agencies fared in courts before the advent of *Chevron* and limn a portrait of institutional constraints in the *Loper Bright* era. It can point us toward deference realities.

Several challenges impede a straightforward historical assessment. To start, there is no existing measure of agencies' litigation outcomes that covers the Supreme Court historically, let alone the lower courts. A significant aspect of this project was to develop and validate such a measure based on expert annotation and the recent generation of transformer-based natural language models. Generating this new measure required months of human labor by a team of experts, in addition to the deployment of recent models.

Even with this measure, however, daunting challenges remain. The first challenge is that the deference regime may be correlated with other features of the political or legal environment. For example, the move from the traditional *de novo* default regime to the transitional regime was occasioned by a significant change in personnel in the Supreme Court—conservative jurists were replaced by liberal ones.[23] A change in litigation outcomes might reflect the shifting deference regime, or instead the shifting composition of the Court. The second challenge is even more troubling: underlying primary behavior and litigation strategies may themselves be influenced by deference regimes. For instance, a more deferential regime may induce agencies to adopt more bold interpretations of statutes. If so, litigation outcomes might not, in fact, change after a shift in deference regimes—but that is because underlying behavior shifted, not because deference regimes do not matter. Both classes of challenges present serious obstacles to any straightforward interpretation of the resulting caselaw.

---

21. *Loper Bright*, 603 U.S. at 396.
22. *Id.* at 399.
23. *See* Daniel E. Ho & Kevin M. Quinn, *Did a Switch in Time Save Nine?*, 2 J. LEGAL ANALYSIS 69, 92 (2010) (noting "the dramatic change in the Court's membership during this time period").

A first step of this study, therefore, is to lay out what can and cannot be learned about the operation of deference regimes from litigation outcomes using a theoretical model.[24] The basic lesson from this exercise is that more might be learned than it initially seems—but that durable, quite possibly irresolvable concerns will remain.

The fundamental empirical strategy is to account for concerns that can be accounted for and to identify and calibrate residual concerns. The analysis examines progressively demanding filters that address, for instance, compositional changes in the judiciary. For concerns that cannot be directly addressed, the analysis uses bounding exercises,[25] which assume best- and worst-case scenarios. The true effect will almost surely be inside the best- and worst-case bounds, allowing us to calibrate the magnitude of the concern.

The results, in brief, do not reveal consistent evidence of an association between deference regime and agency win-rates. Agencies do not seem systematically to fare better following increases in deference. This is true for the increase in deference following the judicial revolution, and even more clearly following the *Chevron* revolution. And it is true for both the Courts of Appeals and for the Supreme Court.

Given the likelihood that deference law bears little relationship with agency litigation outcomes, the later parts of this Article argue for a re-orientation toward deference realities. Consider two prominent realities insufficiently attended to in debates over deference. First, courts are highly limited institutions, and simple realities of judicial administration mean that courts will be inclined to defer to agencies on the mine run of cases.[26] They do not have the capacity to review most cases *de novo* and, regardless of the formal standard of review, in practice will face incentives to defer to agencies.[27] Second, in the minority of cases where jurists care, there will be sufficient pliability in any formal deference regime to reach the desired outcome.[28] Even in a pristine *Chevron* world, jurists may resolve a case at step one rather than two, effectively reviewing the case *de novo*. And, of course, even formally *Chevron* was subject to many exceptions, and it was

24. This theoretical part of the study is considered in more detail in Edward H. Stiglitz, *Public Law Litigation: Legal Standards, Observables, and Inferences* (Cornell L. Sch., Legal Stud. Rsch. Paper No. 25-14, 2025), http://ssrn.com/abstract=5266449 [https://perma.cc/SAX2-2JG9].

25. *See infra* Section IV.D.

26. To take the example of *Chevron* itself, Merrill noted that Justice Blackmun wrote "Whew!" on the first page of the draft opinion. That remark, Merrill suggests, may reflect "a sense of relief that the opinion handled the complicated issue in a way that absolved Justice Blackmun of any further engagement with the matter." Merrill, *supra* note 5, at 274.

27. *See infra* Section V.A.

28. *See infra* Section V.A.

not always clear when *Chevron* would apply versus another formal deference regime.[29]

These deference realities suggest a different path forward in the debate over deference law. To understand what drives deference, more attention ought to lay with realities of decision-making. The deference that agencies enjoy, for instance, may be more shaped by judicial bureaucracy than deference regimes. Do rising caseloads drive deference? Do expansions of judicial-bureaucratic capacity reduce deference? The later parts of this Article consider case-studies of changes to the judicial-bureaucracy—the expansion of law clerks in 1930 and of judges in 1978—as a window into the role of judicial-bureaucratic factors in litigation outcomes. As we head into the young *Loper Bright* period, this analysis suggests that focus should be on institutional levers that might affect deference realities. They may be able to explain historical trends in deference—and they may hold more promise for affecting the relationship between courts and agencies in the post-*Chevron* era.

The balance of this Article proceeds as follows. Part I discusses the broad historical contours of deference law since 1875. Part II explains the methodology used to measure litigation outcomes for relevant cases from 1875–2020. Part III considers the theoretical question of what it is possible to learn about deference law from litigation outcomes and concludes that more might be learned than initially seems. Part IV contains the results from the historical-empirical exercises. Part V turns to two proposed case studies on the relationship between judicial-bureaucratic capacity and litigation outcomes. Part VI speculates about the young *Loper Bright* era, including a call for more attention to focus on judicial-bureaucratic features of the legal system. The conclusion follows.

I. DEFERENCE AND ITS HISTORICAL REGIMES

How much deference, if any, should courts give to agency interpretations of statutes? Courts and scholars regard this as a high-stakes question, implicating both pragmatic policy and constitutional law concerns.

On the policy side, whether administrative agencies or courts determine the meaning of a statute will often influence the policies that can be implemented under the statute. A broadly interpreted statute will allow agencies significant discretion in the policies it pursues; a narrowly interpreted statute will often limit what the agency may accomplish. Though

29. These exceptions emerged over time, including perhaps most prominently the various flavors of step zero inquiry. *See, e.g.*, United States v. Mead Corp., 533 U.S. 218 (2001); King v. Burwell, 576 U.S. 473 (2015).

it was not always this way, recently the conflict over which institution arbitrates statutory meaning tended to pit Democratic administrations against a conservative judiciary.[30] The Supreme Court's docket is littered with such disputes over an agency's interpretive authority: the authority of the Environmental Protection Agency (EPA) to regulate greenhouse gas emissions;[31] the authority of the Department of Education to reduce or eliminate student loan debt;[32] the authority of the Occupational Health and Safety Administration (OSHA) to require certain employers to ensure their workers were vaccinated against COVID-19;[33] or the Center for Disease Control's authority to impose an eviction moratorium on landlords.[34] Some of the most high-profile and pressing policy problems thus come down to the question of whether an agency possesses the relevant statutory authority, the resolution of which often turns on how much deference courts give to agencies.

Aligned with the policy question is one of constitutional law. The first case that most students read in law school is *Marbury v. Madison*, which tells us that it is "emphatically the province and duty of the judicial department to [s]ay what the law is."[35] Binding deference, runs the concern, threatens that understanding of constitutional roles by allowing agencies to say what the law is. It is common to label *Chevron* as counter-*Marbury* or anti-*Marbury*.[36] Justice Roberts's opinion in *Loper Bright* harkened to these concerns,[37] though ultimately rooted his decision in the language of the APA.[38]

Chevron was not the first word in deference law, nor was it the last. The historical landscape of deference is contested on many points, but consider

30. In other periods, the arrangement was roughly reversed. *Chevron* deference itself, for instance, arose as the Reagan administration sought to deregulate, and deference plausibly afforded the administration space to do so against often-liberal lower courts. That simple ideological story would explain, for instance, why Justice Scalia was one of *Chevron* deference's greatest supporters in the initial years. *See* Merrill, *supra* note 5, at 278–79.

31. West Virginia v. EPA, 597 U.S. 697 (2022) (holding that the EPA did not have the authority to regulate greenhouse gases as implemented).

32. Biden v. Nebraska, 600 U.S. 477 (2023) (holding that the Department of Education did not have authority to negate student loans as implemented).

33. Nat'l Fed'n of Indep. Bus. v. Dep't of Lab., OSHA, 595 U.S. 109 (2022) (holding that OSHA did not have authority to "mandate" vaccines).

34. Ala. Ass'n of Realtors v. Dep't of Health & Hum. Servs., 594 U.S. 758 (2021) (holding that HHS did not have authority to require a moratorium on evictions).

35. Marbury v. Madison, 5 U.S. (1 Cranch) 137, 177 (1803).

36. Cass R. Sunstein, *Beyond* Marbury*: The Executive's Power to Say What the Law Is*, 115 YALE L.J. 2580, 2589 (2006) (referring to *Chevron* as a counter-*Marbury*, with approval); Paul R. Verkuil, *The Wait Is Over:* Chevron *as the Stealth* Vermont Yankee *II*, 75 GEO. WASH. L. REV. 921, 925 (2007) (referring to *Chevron* as an anti-*Marbury*).

37. *See* Loper Bright Enters. v. Raimondo, 603 U.S. 369, 385 (2024).

38. *Id.* at 393.

a basic periodization that appears to have been embraced by the Court. First, there was a period that runs from the start of federal-question jurisdiction in 1875 to the judicial revolution of 1937;[39] second, a period that runs from the judicial revolution to *Chevron*;[40] and, third, a period that runs from *Chevron* to *Loper Bright*.[41] We now live in a nascent *Loper Bright* period, the contours of which can scarcely be made out. An objective of this historical-empirical assessment is to help understand the current, fourth period of judicial review post-*Loper Bright*. The past cannot tell us what lies in the future, but it may inform our views of the constants and variables in the relationship between courts and agencies in statutory interpretation.

In an article that appears to have influenced the Court's characterization of historical deference law, Aditya Bamzai regarded the first period of deference as one in which courts reviewed agency interpretations *de novo* by default, giving deference or "respect" to agency positions only when at least one of two canons of construction was relevant.[42] The first canon became relevant when the agency's position reflected a long-standing or customary interpretation.[43] The second canon became relevant when the agency's position was adopted contemporaneously with the passage of the statute.[44] Refracted through those canons, courts may defer to agency interpretations, but this did not reflect a view that agencies deserved deference generically in consequence of their administration of the statute. In this account, there was no "general" deference to agencies during this period.

The story is not entirely uniform in this period. It is possible to identify instances in which courts deferred to agencies without mentioning either the canon of customary or of contemporaneous interpretation. In the 1904 case of *Bates & Guild Co. v. Payne*, Justice Brown summarized his understanding of existing law, writing for the Court that, where Congress commits fact-finding to agency discretion, that discretion carries over to

39. *See id.* at 388–89 (noting several cases from the early 1940s that applied "deferential review"). As noted earlier, this article dates the period from the 1937 revolution, which we find marks a break in historical jurisprudence. *See* Thalken et al., *supra* note 10, at 9259–60.

40. *See Loper Bright*, 603 U.S. at 396 (noting that *Chevron* represented a "marked departure" from earlier deference law).

41. *Id.* at 412 (overturning *Chevron*).

42. Bamzai, *supra* note 11, at 916–19. As noted later, this article has been contested on various grounds by scholars. Though not directly relevant to the present empirical analysis, an important claim Bamzai makes is that the APA represented an effort to roll back deference law to its pre-1940s version—that the APA was a backlash against permissive deference law of the early 1940s. Even if Bamzai's descriptive historical account of deference law holds, it is another step to show that of the APA. For a forceful critique of this aspect of the article, see Ronald M. Levin, *The APA and the Assault on Deference*, 106 MINN. L. REV. 125, 170–74 (2021).

43. Bamzai, *supra* note 11, at 930.

44. *Id.*

law, such that even in questions "of law alone, [an agency's] action will carry with it a strong presumption of its correctness, and the courts will not ordinarily review it, although they may have the power, and will occasionally exercise the right of so doing."[45] The Court did not mention canons of construction as a limiting feature of the decision.[46]

Bamzai regards *Bates & Guild* as anomalous for the period, a "crack in the glass" of the existing deference regime, but not representative of it.[47] Though one can find other examples of the Court giving "weight" to agencies' views without reference to the canons,[48] there is some evidence that at least the *Bates & Guild* Court recognized the decision as a departure. Justice Brown engages in extended throat-clearing before producing his summary of existing doctrine, noting that "although the question is largely one of law . . . there is some discretion left [to the agency],"[49] and observing pragmatically that a different view would leave courts "flooded" by cases.[50] Were his subsequent summary of the existing doctrine uncontroversial, those features of his otherwise economical opinion would be unnecessary. In dissent, Justice Harlan is more explicit: "The rule of construction which this court has recognized for more than three-quarters of a century is now overthrown."[51]

Dissents tend to exaggerate, and Justice Harlan's appears to be no exception. Though *Bates & Guild* was sometimes later cited for the proposition that Courts would defer to agencies on questions of law,[52] it was referred to in only ten cases by the Court between 1904 and 1937,[53] and often for the conclusiveness of agency findings of fact.[54] One case, for example, affirmatively cites *Bates & Guild* and also explicitly carves out

45. Bates & Guild Co. v. Payne, 194 U.S. 106, 109–10 (1904).

46. *Id.*

47. Bamzai, *supra* note 11, at 966. *But see* Levin, *supra* note 42, at 170 (contesting the uniqueness of *Bates & Guild*).

48. *E.g.*, United States v. Reynolds, 250 U.S. 104, 109 (1919) (noting that the agency position is "entitled to weight as an administrative interpretation of the act," though also noting that "it comports with our impression of the natural meaning of the language employed by Congress").

49. 194 U.S. at 107.

50. *Id.* at 108.

51. *Id.* at 111 (Harlan, J., dissenting).

52. Pub. Clearing House v. Coyne, 194 U.S. 497, 508 (1904) (citing *Bates & Guild* and observing that, "in many cases . . . the action of the department is accepted as final by the courts, and even when involving questions of law this action is attended by a strong presumption of its correctness").

53. Crowell v. Benson, 285 U.S. 22, 89 n.26 (1932); Silberschein v. United States, 266 U.S. 221, 225 (1924); Leach v. Carlile, 258 U.S. 138, 139–40 (1922); Brougham v. Blanton Mfg. Co., 249 U.S. 495, 499–500 (1919); Houston v. St. Louis Indep. Packing Co., 249 U.S. 479, 484 (1919); Smith v. Hitchcock, 226 U.S. 53, 58 (1912); Cent. Tr. Co. v. Cent. Tr. Co. of Ill., 216 U.S. 251, 261 (1910); Nat'l Life Ins. Co. of U.S. v. Nat'l Life Ins. Co., 209 U.S. 317, 325 (1908); *Pub. Clearing House*, 194 U.S. at 508 (same case as *National Life Insurance Co.*); Dismuke v. United States, 297 U.S. 167, 172 (1936); United States *ex rel.* Chi. Great W. R.R. Co. v. Interstate Com. Comm'n, 294 U.S. 50, 63 n.11 (1935).

54. *E.g.*, *Leach*, 258 U.S. at 139–40; *Houston*, 249 U.S. at 484.

questions "wholly dependent upon a question of law."[55] A reasonable interpretation of this pattern is that the case did not overthrow the existing doctrinal arrangement and that, as Bamzai put it, "courts in the first few decades of the twentieth century generally hewed to the traditional interpretive formulations,"[56] with courts tending to review questions of law *de novo*.[57] I refer to this initial, largely *de novo* period after 1875 as the *early* regime of deference law.

It is common to pin a change in deference as occurring in the 1940s.[58] In earlier empirical work, however, I find with co-authors that the judicial revolution of 1937 marked a dramatic shift in jurisprudence,[59] and that event will serve as a period transition in the present analysis. It is difficult to cleanly summarize the scope of review during the period between the judicial revolution and *Chevron*, but the idea of deference on questions of law became more familiar in this time.[60] That understanding was facilitated by a view that legal questions were often intrinsically linked to factual questions. As deference on questions of fact was established even in the period before the judicial revolution,[61] the insight that fact was bound to law called for a doctrinal reckoning. Courts tended to resolve that tension by giving deference to agencies on questions of law, at least where they detected interwoven factual questions. In practice, that characterizes a wide class of disputes.[62]

Gray v. Powell is the most discussed case marking a post-revolution turn to deference.[63] The case centered on whether a railway consumer of coal who also produced coal through a mining operation could be a coal "producer" under the Bituminous Coal Act and therefore eligible for statutory price-fixing exemptions.[64] The Department of Interior determined that they were not producers, and they challenged that determination. The court of appeals rejected that determination, but the Supreme Court upheld it, significantly holding that "[w]here, as here, a determination has been left to an administrative body, this delegation will be respected and the administrative conclusion left untouched."[65] The Court's position was

55. *Silberschein*, 266 U.S. at 225.
56. Bamzai, *supra* note 11, at 969.
57. *Id.*
58. *E.g.*, *id.* at 976; Levin, *supra* note 42, at 161; Bernard Schwartz, Gray vs. Powell *and the Scope of Review*, 54 MICH. L. REV. 1 (1955).
59. *See* Stiglitz & Thalken, *supra* note 10, at 91–92; Thalken et al., *supra* note 10, at 9259–60.
60. *E.g.*, Loper Bright Enters. v. Raimondo, 603 U.S. 369, 388 (2024).
61. Crowell v. Benson, 285 U.S. 22, 51 (1932).
62. *E.g.*, NLRB v. Hearst Publ'ns, Inc., 322 U.S. 111, 130–31 (1944).
63. Gray v. Powell, 314 U.S. 402, 412–13 (1941).
64. *Id.* at 403.
65. *Id.* at 412.

remarkable because, as the dissent noted, "there is not a single disputed fact" in the record,[66] making this a matter of pure law or statutory interpretation. *NLRB v. Hearst Publications, Inc.* is another case commonly regarded as marking a shift to deference, there on the meaning of "employee."[67] A third commonly cited case from this period involved allowable tax deductions, *Dobson v. Commissioner*, which held that "the judicial function is exhausted when there is found to be a rational basis for the conclusions approved by the administrative body," and noted that the agency must "consider both law and facts."[68]

Even as the Court signaled more deference in these cases, however, they were not necessarily consistently followed. Kenneth Culp Davis observed in his 1951 administrative law treatise, "[t]he one statement that can be made with confidence . . . is that sometimes the Supreme Court applies it and sometimes it does not. . . . Many cases defy explanation except in terms of judicial discretion."[69] In *Loper Bright*, Justice Roberts noted this apparent inconsistency, and minimized the new deference cases by saying they "did not purport to refashion the longstanding judicial approach to questions of law," and were "cabined to fact-bound determinations."[70] The extent to which these cases refashioned law or were fact bound is disputed. Bamzai, for instance, writes that these cases "effectively abandon[ed] the traditional interpretive methodology,"[71] a conclusion that other scholars and the dissenting justices in *Loper Bright* support.[72] As Matthew Stephenson notes, Davis's quote, indeed, may be misleading, as apparent inconsistency may be due to inconsistency in the invocation of *Gray* rather than inconsistency

66. *Id.* at 418 (Roberts, J., dissenting).

67. 322 U.S. at 111–12.

68. Dobson v. Comm'r, 320 U.S. 489, 501 (1943) (internal quotation omitted).

69. Schwartz, *supra* note 58, at 2 (quoting KENNETH CULP DAVIS, ADMINISTRATIVE LAW 893 (1951)). Breyer et al. characterized the period after the judicial revolution and before *Chevron* as one in which:

> [L]eading cases supported the view that great deference must be given to the decision of an administrative agency applying a statute to the facts and that such decisions can be reversed only if without rational basis. . . . On the other hand, the Court sometimes refused to 'defer' to an agency interpretation of a statute . . . [and] sometimes [courts] distinguished between a pure question of law, which was for the courts to decide, and 'mixed questions of law and fact,' where . . . deference was more appropriate.

BREYER ET AL., *supra* note 5, at 235–36.

70. Loper Bright Enters. v. Raimondo, 603 U.S. 369, 370 (2024).

71. Bamzai, *supra* note 11, at 976–77.

72. Levin, *supra* note 42, at 170–74; 603 U.S. at 468 (Kagan, J., dissenting) (citing to Davis for the idea that *Gray* was "the leading case" of the time on deference, and that it "established what is known as 'the doctrine of *Gray v. Powell*'" (quoting DAVIS, *supra* note 69, at 882).

in the application of its principles.[73] In this sense, Justice Roberts's opinion appears to be the minority one. I refer to the period between the judicial revolution and *Chevron* as the *transitional* regime.

Chevron itself marks the start of the third putative regime of deference law. Though the justices involved in the case did not appear to see that they were revolutionizing deference law, the decision soon took life in the D.C. Circuit and lower courts, credited often to the efforts of then-Judge Scalia.[74] The decision earned 81 references in 1985, 109 in 1986, and over 150 by 1987—roughly half of these references were D.C. Circuit cases.[75] Doctrinally, the decision established the famous *Chevron* two-step, in which courts were first to ask if the statute spoke to the precise issue, and if not, then simply to ask if the agency's construction was "permissible" or "reasonable."[76]

To some observers, including Justice Roberts in *Loper Bright*, *Chevron* represented a "marked departure" from earlier deference law.[77] In this telling, *Chevron* was remarkable because, unlike earlier forms of deference law, "[i]t requires a court to *ignore*, not follow, 'the reading the court would have reached' had it exercised its independent judgment."[78] It was also a general form of deference, in the sense that on its terms it did not require complex intermingling of factual and legal questions, nor did it require that the agency's interpretation be long-standing or adopted near the passage of the statute. *Chevron* deference did not condition on the presence of factual questions, nor on the timing of the agency's interpretation—the interpretation in *Chevron* itself represented a revision of a position adopted closer to the act's passage.[79] This type of deference, Justice Roberts wrote, was inconsistent with the APA, which called on courts to "decide all relevant questions of law, interpret constitutional and statutory provisions, and determine the meaning or applicability of the terms of an agency action."[80] *Chevron* represented a requirement for courts to let agencies decide questions of law, rather than the courts.

---

73. Matthew C. Stephenson, The *Gray* Area: Finding Implicit Delegation to Agencies After *Loper Bright* 34 (Aug. 25, 2025) (unpublished manuscript), https://ssrn.com/abstract=5328964 [https://perma.cc/P5J7-X2J9].

74. See Merrill, *supra* note 5, at 277.

75. These figures represent results from searches in Westlaw's federal Courts of Appeals database.

76. Chevron U.S.A. Inc. v. Nat. Res. Def. Council, Inc., 467 U.S. 837, 842–43 (1984), *overruled by Loper Bright*, 603 U.S. at 412–13.

77. 603 U.S. at 396.

78. *Id.* at 398–99 (citing *Chevron*, 467 U.S. at 843 n.11).

79. 467 U.S. at 863 (noting that the agency "changed its interpretation").

80. 5 U.S.C. § 706.

Loper Bright overturned *Chevron*, setting us down another path on deference law. Under *Loper Bright*, the APA requires courts to decide statutory questions.[81] The Court left open the possibility of courts granting "respect" to agency interpretations, but those agency interpretations would not be binding, and courts must exercise "independent judgment."[82] Open questions in the *Loper Bright* era include what "respect" means in this context, when it is owed and when it is not, and how clear Congress must be when it "expressly" delegates discretion to agencies.[83] *Loper Bright* promises to generate a rich family of cases. Already at the time of this writing, less than a year after *Loper Bright* was issued, Westlaw indicates that over 100 Courts of Appeals decisions cite to the case. That rapid citation uptake meets or exceeds that of *Chevron*.[84] Also as with *Chevron*, the doctrine's full contours will take time to develop, as courts refine the doctrine in application, and compositional changes in the judiciary bear on its understanding.

Many of the historical points in the foregoing are contested.[85] Did courts defer prior to the judicial revolution or review *de novo*? Did *Gray* refashion deference law in a broad way, or was it instead an inconsistent doctrinal signal? But the basic contours of at least the Court's understanding of the history can be made out: a progression from less to more deference over time, with possible breaks at the judicial revolution and *Chevron*, with *Loper Bright* representing the most significant deviation from that trend.[86]

Taking those as candidate regimes of deference law, it is fair to ask—do they matter? The fortunes of agencies before courts no doubt change over time, but it is not clear that those changes should be attributed to doctrinal shifts.

A core challenge in any empirical study is to isolate, to the extent possible, the changes in doctrine from these other, confounding influences. Both non-doctrinal influences—institutional pressures to defer, and policy

81. 603 U.S. at 398–99 (noting that the APA commands that the reviewing court decide all relevant questions of law).

82. *Id.* at 394, 404.

83. *Id.*

84. As of this writing, a total of 117 Courts of Appeals decisions cite to *Loper Bright*; a total of 104 Courts of Appeals decisions cited to *Chevron* in the year after it was issued.

85. *Compare* Bamzai, *supra* note 11, *with* Levin, *supra* note 42, *and* Stephenson, *supra* note 73.

86. Of course, part of the story of *Chevron* was its whittling away, from *Mead*, to the first versions of the Major Questions Doctrine, to the New Major Questions Doctrine—*Loper Bright* is simply the starkest example of the turn against deference. *See* United States v. Mead Corp., 533 U.S. 218 (2001); King v. Burwell, 576 U.S. 473 (2015) (applying an earlier version of the major questions doctrine); West Virginia v. EPA, 597 U.S. 697 (2022) (developing the new major questions doctrine); *see also* Daniel T. Deacon & Leah M. Litman, *The New Major Questions Doctrine*, 109 VA. L. REV. 1009 (2023); Cass R. Sunstein, *There Are Two "Major Questions" Doctrines*, 73 ADMIN. L. REV. 475 (2021).

preferences—pose serious threats to straightforward inferences. Changes in caseload, for instance, might roughly coincide with changes in doctrine. Do courts defer more because of increasing caseload? Or the doctrinal change? Likewise, changes in policy preferences might roughly coincide with changes in doctrine. Do courts defer more because the judiciary became more liberal? Or because of changes in doctrine? The empirical assessments attempt to manage these concerns.

II. MEASURING LITIGATION OUTCOMES AND CASE CHARACTERISTICS

A. Existing Approaches

An initial obstacle to understanding how agencies fare under various deference regimes is that there is no readily available measure of litigation outcomes. Existing scholarship on agency outcomes is limited with respect to jurisdictions and years, as well as the method of identifying the relevant sample within those years and jurisdictions.

The largest-scale recent study is Barnett and Walker's study of *Chevron* in the Courts of Appeals, in which they examined agency performance in 1,327 cases decided between 2003 and 2013.[87] They selected cases based on references to the *Chevron* decision,[88] a sample that included all circuits and agencies. They then hand-coded the cases for whether the court accepted the agency's interpretation, applied deference regime, and other relevant variables. An earlier study of similarly impressive scope by Schuck and Elliott considered how agencies fared in roughly 2,500 decisions of the Courts of Appeals in the periods before and after the *Chevron* decision.[89] They examined any case of direct review of agency action, and hand-coded the decisions for whether the court affirmed or remanded the agency action, and other relevant variables. Another prominent study by Miles and Sunstein examined all Courts of Appeals decisions that applied *Chevron* deference and involved the NLRB or the EPA between 1990 and 2004.[90] They again hand-coded litigation outcomes and other relevant variables.

Notice that each of these notable studies is limited by years, courts, or agencies. An equally important limitation of the studies is that the sample

87. Kent Barnett & Christopher J. Walker, Chevron *in the Circuit Courts*, 116 MICH. L. REV. 1, 5 (2017).

88. They further filter cases to include only those that refer to: agency, ALJ, order, formal adjudication, rule, and § 553. *Id.* at 22.

89. Peter H. Schuck & E. Donald Elliott, *To the* Chevron *Station: An Empirical Study of Federal Administrative Law*, 1990 DUKE L.J. 984, 1003.

90. Thomas J. Miles & Cass R. Sunstein, *Do Judges Make Regulatory Policy? An Empirical Investigation of* Chevron, 73 U. CHI. L. REV. 823, 825 (2006).

of cases tends to condition on references to the *Chevron* decision. If an opinion does not cite to *Chevron*, Barnett and Walker or Miles and Sunstein will not identify that case in their sample. That sample selection feature is important because jurists often exercise discretion regarding what cases to cite or deference regime to apply. Suppose that courts always accept agency interpretations when they cite *Chevron*, but they only cite *Chevron* in cases in which they accept the agency interpretation; when they disagree, they cite another standard of review or simply proceed without clearly applying a standard of review. In this scenario, it will appear that *Chevron* results in a one-hundred percent agency win-rate. But of course, that metric means little because they only cite *Chevron* when they accept the agency's interpretation. This is an extreme example designed to illustrate a point. Yet evidence abounds that citing to specific standards is partially discretionary. Empirical efforts demonstrate the decision of whether to use *Chevron* deference is inflected by political considerations;[91] and at the Supreme Court level, citations to *Chevron* were declining even before *Loper Bright* arrived, consistent with the Court's souring mood on the doctrine.[92] Other approaches to selection exist, but they too are limiting. The Schuck and Elliott study avoids citation-based selection, though examines only a narrow interval of time.[93] A related study by Eskridge and Baer, likewise, avoids citation-selection: They consider all cases involving statutory interpretation, covering all agencies and cases decided between 1983 and 2005.[94] But they examine only the Supreme Court, where the (relatively) small number of decided cases makes this feasible.

Existing studies make reasonable research choices—scholars have limited time and resources, so they must be strategic about which cases to examine. A goal of this study is to escape those limits. The study seeks to examine the full run of cases involving review of agency statutory interpretation, from the birth of general federal question jurisdiction to approximately the present. It aims to cover all circuits, agencies, and years over that period. Moreover, it does so without relying on citation-based sample selection.

---

91. Kent Barnett, Christina L. Boyd & Christopher J. Walker, *The Politics of Selecting* Chevron *Deference*, 15 J. EMPIRICAL LEGAL STUD. 597, 599 (2018).

92. Linda Jellum, Chevron*'s Demise: A Survey of* Chevron *from Infancy to Senescence*, 59 ADMIN. L. REV. 725, 730 (2007); *see also infra* Section V.A. Eskridge and Baer, moreover, find at the Supreme Court level that over half of relevant decisions do not engage in any formal deference test, opting instead for "ad hoc judicial reasoning." William N. Eskridge, Jr. & Lauren E. Baer, *The Continuum of Deference: Supreme Court Treatment of Agency Statutory Interpretations from* Chevron *to* Hamdan, 96 GEO. L.J. 1083, 1100 (2008).

93. Schuck & Elliott, *supra* note 89, at 990–91.

94. Eskridge & Baer, *supra* note 92, at 1094.

B. Measuring Litigation Outcomes

Previously impossible, that task is now feasible due to the advent of transformer-based language models.[95] Only a few years old,[96] the architecture of these models allows them to understand context, as was not possible with most earlier models.[97] Thus, as in ordinary human understanding of language, the meaning of a word depends on the words that come before and after it in these transformer-based models.[98] The attention mechanism in the architecture, moreover, allows the models to be trained in parallel, greatly expanding the scope of what the models can be trained on.[99] This enlarged scope of training data is what gives the models their common label of large language models. That core architecture powers the Generative Pre-trained Transformer (GPT) model and other generative language models, though not all large language models are generative. In earlier work, indeed, I find with a team of co-authors that a fine-tuned, lighter-weight, non-generative large language model outperforms larger generative language models on complex text classification tasks.[100] The pipeline for the present study involves both generative and fine-tuned, lighter-weight language models.

The raw data for this project comes from Harvard's Caselaw Access Project, which produced machine-readable versions of the federal reporters.[101] After filtering out short orders and decisions issued before 1875, the raw data consist of approximately 465,000 Courts of Appeals decisions and 26,000 Supreme Court decisions, totaling some 1.4 billion words.[102] With that data in hand, two basic tasks remain: First, identify

95. For another novel application of these models, see Stiglitz & Thalken, *supra* note 10 and Thalken et al., *supra* note 10.

96. For the seminal paper, see Ashish Vaswani et al., Attention Is All You Need (Aug. 2, 2023) (unpublished manuscript), https://arxiv.org/abs/1706.03762 [https://perma.cc/ZV2K-SZ27].

97. *Id.* at 10.

98. *Id.* at 5 (explaining that "[e]ach position in the encoder can attend to all positions in the previous layer of the encoder," meaning each token's representation is computed with reference to its full document context within the model's maximum sequence length).

99. *Id.* at 2.

100. Thalken et al., *supra* note 10, at 9260.

101. *Caselaw Access Project: Our Data*, HARV. L. SCH.: CASELAW ACCESS PROJECT, https://case.law/ [https://perma.cc/BGB2-FX3K]. Note that relatively few agency cases go unreported. For example, searching for agency cases on Westlaw decided between 1980 and 2000 (advanced: (("rule" OR "regulation") & ("agency" OR "board" OR "commission") & ("statute" OR "act" OR "legislation" or "law" or "statutory"))) reveals 130 unreported cases and 9,244 reported cases.

102. The series starts in 1875 because that denotes the birth of federal question jurisdiction. However, the Courts of Appeals did not exist as we understand them until the Evarts Act of 1891. *See* JON O. NEWMAN, HISTORY OF THE ARTICLE III APPELLATE COURTS, 1789–2021: THE EVOLUTION OF THEIR GEOGRAPHIC SCOPE, NUMBER OF JUDGESHIPS, AND JURISDICTION 2 (2021), https://www.fjc.gov/content/363614/history-article-iii-appellate-courts-1789-2021 [https://perma.cc/7Y6N-7KWD]. For simplicity of exposition, I refer to the appellate courts as the "Courts of Appeals" in this Article.

decisions that involve agency interpretations of statutes; and second, identify relevant case characteristics, most importantly the litigation outcome.

The first task involves filtering out the bulk of decisions that do not involve agency statutory interpretation, a critical step in the analysis. Earlier studies perform this task either by selecting the sample based on citations to important cases, such as *Chevron*, or by reading the cases in a narrow jurisdictional domain, such as the Supreme Court docket.[103] However, selection based on citations is problematic due to the possibility of strategic or selective citation and doctrinal application; moreover, it is not clear what case citations to focus on, given the long series studied in this analysis. Reading all cases is infeasible. At an average reading pace, it would take a human over ten years to read the corpus in this study, assuming they read constantly without breaks or sleep.[104]

The approach of this paper is to fine-tune a large language model to classify the cases as involving agency statutory interpretation or not. To fine-tune the model, we need examples of cases that do and do not involve agency statutory interpretation—the model learns from these examples the patterns in language that characterize each of the classes.[105] Among the many features of a case that Westlaw Headnotes characterize is whether they involve the "Administrative Construction of Statutes," a sub-feature of Administrative Law and Procedure in their classification system.[106] Their coverage of cases is not entirely comprehensive—there are cases that involve statutory interpretation that the Headnotes do not identify. But the cases with that heading cover a wide range of agencies, courts, and time periods, and therefore provide a solid foundation for the model to learn the relevant patterns in judicial language. A (stratified) random sample of cases that do not fit within that Headnote represent the examples that do not involve agency statutory interpretation.[107] As reported in more detail in the

103. *Supra* Section II.A.

104. The average human reads about 250 words per minute. *E.g.*, Mark D. Jackson & James L. McClelland, *Sensory and Cognitive Determinants of Reading Speed*, 14 J. VERBAL LEARNING & VERBAL BEHAV. 565, 568 (1975) (noting average readers in the 200–300 words per minute range). At that pace, a corpus of 1.4 billion words would take about 10.7 years to read if they read without a break for that period.

105. *See* Thalken et al., *supra* note 10 (fine-tuning language models to produce a measure of jurisprudence).

106. I use the "general" category in this label: "15A-k2201." Westlaw was used only to identify relevant cases, with case identifiers collected manually using a standard academic account. Headnote text itself was not used, and case text derives from the Caselaw Access Project.

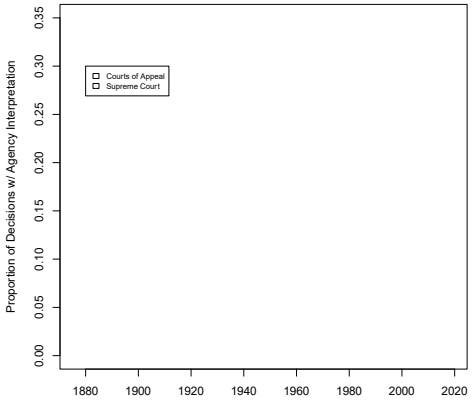
107. The negative cases are stratified so that they correspond in proportion to the jurisdictions of the positive cases.

Appendix, these examples serve to fine-tune, or teach, a large language model how to identify cases with statutory interpretation.

The model performs remarkably well at classifying cases with statutory interpretation. The standard metric for classification success is the F1 statistic, which averages performance over both recall (the fraction of true positives identified) and precision (the fraction of identified positives that are true positives). A perfect F1 score is 1, and the model returns with a score of 0.92, which is high by any standard, and indicates low levels of false positives and false negatives. Applying this model to the entire corpus, it identifies approximately 28,000 cases that involve agency statutory interpretation: roughly, 3,400 at the Supreme Court level and 25,000 at the Courts of Appeals level. Those figures represent roughly thirteen percent of Supreme Court cases and six percent of Courts of Appeals cases.

It is worth pausing to examine changes in the number of cases with agency statutory interpretation over time. This serves both as a validation check on the model results and helps to contextualize the results to follow. Figure 1 plots the proportion of decisions with agency statutory interpretations at the Supreme Court and Courts of Appeals levels. The figure shows that, at both the Supreme Court and Courts of Appeals levels, agency statutory cases were relatively rare before the early 1900s and progressive era legislation. At both levels, statutory decisions increased markedly between about 1920 and 1940, after which they approximately equilibrated at five to ten percent of decisions for the Courts of Appeals and fifteen to twenty-five percent for the Supreme Court. These patterns fit with basic expectations: Agency interpretations were rare before the progressive era legislation and accompanying agencies; they increased through the 1920s and 1930s, as progressive era agencies and statutes combined force with New Deal agencies and statutes; those issues peaked, were partially resolved, and settled at a mid-century stasis; that stasis was upset by the third major wave of regulatory legislation and agencies in the early 1970s, resulting in increasing caseloads through that decade; those issues peaked, were partially resolved, and we settled into a late-century stasis. As many of the issues presented by regulatory statutes and agencies were novel, difficult, and consequential, it is natural that they represent a relatively large fraction of the docket in the Supreme Court.

FIGURE 1. PROPORTION OF DECISIONS WITH AGENCY INTERPRETATIONS



The set of cases identified by this first stage of the pipeline as involving agency statutory interpretation constitutes the primary sample analyzed in the second step of the pipeline. The objective of this second step is to identify case characteristics based on the raw opinion data. The most important characteristic is whether the court accepts or rejects the agency's interpretation of the statute at issue. This is a challenging characteristic to measure. Annotating each decision by hand, as is done in existing work, is impossible due to the scope of the study. Moreover, the examples we use to train or evaluate the model should map onto the primary sample of this study. They should cover the years, jurisdictions, and substantive domains represented in the sample. Yet the samples used in earlier research do not match on at least one of those criteria: None reach much earlier than 1980; most focus only on the Supreme Court or Courts of Appeals, but not both; they often generate samples based on citations to specific cases. As such, existing annotated data is inapt for the present analysis, and new human-annotated data must be collected.

To develop new training data, a team of four upper-year research assistants from Cornell Law School was tasked with annotating hundreds of federal cases over the course of a semester. The pool of cases is the same pool used to train the large language classifier in the first step of the pipeline, those identified by Westlaw as involving the administrative construction of statutes. The annotation team's cases were randomly drawn from that pool each week, with more weight placed on earlier time periods due to the smaller number of cases. As detailed in the Appendix, a random sub-sample of cases was assigned to all members of the team and used to assess inter-coder reliability, which varied by week but was generally strong. For each

case the team reviewed, they were asked to determine whether the court "favored" or "accepted" the agency's interpretation, or instead "rejected" the interpretation. The task was modeled closely on the codebook from Eskridge and Baer's seminal study.[108]

The goal is to take this annotated sample and project it onto the mass of cases for which we do not have human-annotated labels. This is a difficult task because many judicial opinions tend to be lengthy, and a jurist may reveal their position on the agency interpretation at almost any place in the opinion. Moreover, jurists often discuss the pros and cons of accepting an agency position, semantic mottling which can confound classifiers. I experimented with various approaches: simply asking a generative large language model to classify the decisions; fine-tuning a generative language model; fine-tuning a lighter-weight non-generative language model. They tended to perform poorly. A strategy that was remarkably effective, however, broke the task into two sub-steps: First, ask a sophisticated generative language model to summarize the decision with respect to the agency interpretation;[109] second, fine-tune a light-weight model using those summaries and the human annotations as inputs. This two-step approach yielded a model that returned with a strong F1 performance of 0.84.

To provide a sense of the model's predictions, consider several cases in which agencies won and lost in various historical periods. Table 1 reports six decisions within each period, three of which favor the agency and three of which run against the agency. The last column of the table reports the model's predicted probability that the court favors the agency. As can be seen from the table, the model predictions tend to track the litigation outcomes closely. Decisions that favor the agency tend to return with a predicted probability of that outcome of around ninety percent; by contrast, decisions that go against the agency tend to return with a predicted probability that the court accepts the interpretation of closer to twenty percent.[110] Notably, of the eighteen reported cases, only three were in the

108. Eskridge & Baer, *supra* note 92, at 1203.

109. In this draft, I use Gemini 2.0 Flash, which performs about as well as GPT-4o on the LegalBench benchmark, is inexpensive, and has low latency.

110. Most of the model's missteps appear to relate to errors in case summarization (rather than in the classifier). For instance, the summarizer might confuse whether the Supreme Court sides with the lower court or with the agency. For example, in *Gray v. Powell*, 314 U.S. 402 (1941), the summarizer declared in one sentence that the Court rejected the agency's position, but then in another said that the Court reversed the lower court, "thereby supporting the agency's original determinations." The classifier finds this mixed language confusing, and hence the middle-ground classification probability. Despite such occasional challenges, the model performs well overall, showing strong F1 performance and accuracy.

training data: *Norwegian Nitrogen Products Co. v. United States*;[111] *Social Security Board v. Nierotko*;[112] and *FTC v. Bunte Bros.*[113]

TABLE 1. EXAMPLE DECISIONS FROM EACH PERIOD

	Decision	Outcome	Model: Probability Win
Early (1875-1937)	*Norwegian Nitrogen Prods. Co. v. United States,* 288 U.S. 294 (1933)	Wins	0.97
	Bates & Guild Co. v. Payne, 194 U.S. 106 (1904)	Wins	0.96
	Louisville & Nashville R.R. Co. v. United States, 238 U.S. 1 (1913)	Wins	0.97
	ICC v. Cincinnati, New Orleans & Texas Pacific Ry. Co., 167 U.S. 479 (1897)	Loses	0.1
	FTC v. Gratz, 253 U.S. 421 (1920)	Loses	0.08
	FTC v. Eastman Kodak Co., 274 U.S. 619 (1927)	Loses	0.05
	Gray v. Powell, 314 U.S. 402 (1941)	Wins	0.54
	NLRB v. Hearst Publ'ns, Inc., 322 U.S. 111 (1944)	Wins	0.97
Transitional (1938-1983)	*Train v. Nat. Res. Def. Council*, 421 U.S. 60 (1975)	Wins	0.97
	Soc. Sec. Bd. v. Nierotko, 327 U.S. 358 (1946)	Loses	0.1
	FPC v. Sierra Pac. Power Co., 350 U.S. 348 (1956)	Loses	0.09
	FTC v. Bunte Brothers, 312 U.S. 349 (1941)	Loses	0.07
	Chevron U.S.A., Inc. v. Nat. Res. Def. Council, Inc., 467 U.S. 837 (1984)	Wins	0.97
	Nat'l Cable & Telecomms. Ass'n v. Brand X Internet Servs., 545 U.S. 967 (2005)	Wins	0.97
Chevron (1984-2024)	*City of Arlington v. FCC*, 569 U.S. 290 (2013)	Wins	0.97
	FDA v. Brown & Williamson Tobacco Corp., 529 U.S. 120 (2000)	Loses	0.1
	MCI Telecommunications Corp. v. AT&T Co., 512 U.S. 218 (1994)	Loses	0.07
	Christensen v. Harris County, 529 U.S. 576 (2000)	Loses	0.12

Note. Three of these cases were in the training data: Norwegian Nitrogen Prods. Co. v. United States, Soc. Sec. Bd. v. Nierotko, and FTC v. Bunte Brothers.

C. Other Case Attributes

The empirical application requires several case attributes other than the litigation outcome. These case features derive from three main sources.

First, basic case descriptors, such as the case citation, majority opinion author, and date of decision come directly from the Caselaw Access Project JSON files. Straightforward parsing of those files also produces the number of words in the majority opinion.

Second, many more involved features, such as the agency relevant to the litigation, can be identified using large language models. Though the generative models performed poorly when tasked with determining whether the agency won or lost the case, they tend to perform well at tasks such as

111. Norwegian Nitrogen Prods. Co. v. United States, 288 U.S. 294 (1933).
112. Soc. Sec. Bd. v. Nierotko, 327 U.S. 358 (1946).
113. FTC v. Bunte Bros., 312 U.S. 349 (1941).

summarization[114] (used earlier) and named entity recognition (which agency is in this case?),[115] even without fine-tuning or prompting with examples.[116]

Finally, some of the tests call on measures of the judge's ideology. Scholars adopt a variety of approaches to measuring judicial ideology.[117] Most of the existing approaches, however, cannot be used for this study because they require data that does not exist in the jurisdictions or years covered. For instance, one approach requires campaign finance data;[118] that is not available except for recent decades. Other approaches trade on voting data;[119] however, those measures cannot generally be compared across jurisdictions or levels of judicial hierarchy. The study, therefore, turns to an older measurement strategy: attributing the ideology of the judge to the president who appoints them.[120] More recent strategies hold advantages over this president-attribution strategy—it is not the case that all judges appointed by a president share the same ideology, let alone his specific ideology.[121] Yet, though imperfect, the ideology of the appointing president carries information about the ideology of a judge—it is an estimate—and

114. Tianyi Zhang et al., *Benchmarking Large Language Models for News Summarization*, 12 TRANSACTIONS ASS'N FOR COMPUTATIONAL LINGUISTICS 39 (2024).

115. Mingchen Li & Rui Zhang, How Far Is Language Model from 100% Few-Shot Named Entity Recognition in Medical Domain (May 5, 2024) (unpublished manuscript), https://arxiv.org/abs/2307.00186 [https://perma.cc/2BBY-F7D7]; Shuhe Wang et al., GPT-NER: Named Entity Recognition via Large Language Models (Oct. 7, 2023) (unpublished manuscript), https://arxiv.org/abs/2304.10428 [https://perma.cc/MNV6-QH5D].

116. Some empirical exercises, moreover, involve the ideological direction of the agency decision or judicial decision, liberal or conservative. That task may be more difficult than extracting the relevant agency, but the models produced aligned predictions when provided sufficiently detailed examples of what we mean by "liberal" and "conservative." Those in-context examples relevant to ideological labels derive from the codebook for Songer's database of Courts of Appeals decisions. Donald R. Songer, *U.S. Courts of Appeals Databases*, SONGER PROJECT (Nov. 19, 2010), http://www.songerproject.org/us-courts-of-appeals-databases.html [https://perma.cc/95KZ-MZ9R]. To validate the bot responses, I provided a research assistant with the same task for a random sample of 200 cases. Using that human judgment as the ground truth, the model performed well with an F1 score of 0.82. To contextualize that F1 score, I tasked a second research assistant to independently label a random sample of the same subset. The second research assistant performed very similarly to the bot. If we again suppose the first research assistant is the ground truth, the second research assistant returned with similarly strong classification performance, an F1 score of 0.83. There may be difficult border line cases on which humans disagree, but on this task at least human performance is very similar to model performance if properly prompted.

117. Adam Bonica & Maya Sen, *Estimating Judicial Ideology*, 35 J. ECON. PERSPS. 97 (2021).

118. *Id.* at 105.

119. *Id.* at 101.

120. The ideology of the appointing president, in turn, comes from Poole and Rosenthal's widely used measure of ideology. Keith T. Poole & Howard Rosenthal, *Patterns of Congressional Voting*, 35 AM. J. POL. SCI. 228, 230 (1991). Epstein et al.'s approach is the closest to the one adopted in this study. Lee Epstein, Andrew D. Martin, Jeffrey A. Segal & Chad Westerland, *The Judicial Common Space*, 23 J.L. ECON. & ORG. 303, 306–09 (2007). That measure, however, extends back only to 1953, and this study starts in 1875.

121. For example, President Reagan appointed both Justice Kennedy and Justice Scalia, who often disagreed, as in *Mistretta v. United States*, 488 U.S. 361 (1989).

this strategy holds the great advantage that it reaches all years and jurisdictions in this study.

III. LEARNING ABOUT DEFERENCE LAW

What can we hope to learn from win-rates about deference law? A straightforward view is that, if deference law matters, agency win-rates ought to increase if deference increases. Yet that view is not necessarily correct. It is not correct because it is very unlikely that all features of the world remain static as deference law changes.

The threats to learning about deference law from win-rates are many. A central difficulty is that the nature of agency interpretations may change in response to changes in deference law.[122] Told that courts will defer to their interpretations, runs the concern, agencies will adopt bolder interpretations that more aggressively pursue their ends. Because they adopt bolder, less plausible interpretations, an increase in deference may cash out with no increase in win-rates. But that is not because deference law does not matter—it is because agencies adapt their behavior to deference law. A related concern is that deference law may create or augment a zone of interpretations that fit within an effective litigation safe harbor, such that they do not face litigation because would-be challengers view it so unlikely that they will win.[123] That may affect win-rates by changing the composition of litigated cases, typically by censoring the least-bold interpretations and therefore likely depressing observed agency win-rates.

Other clear threats to straightforward inference include changes in the composition of the judiciary and other changes temporally proximate to changes in deference law. Both major changes in deference law—around the judicial revolution of 1937 and the *Chevron* decision—were occasioned by changes in the judges deciding agency cases. In the first instance, the judiciary became markedly more liberal, as President Roosevelt's appointees filled the bench;[124] in the latter instance, the judiciary plausibly became more friendly to President Reagan, as his judicial appointees took office. If there is an increase in agency win-rates following the judicial revolution, is that deference law induced higher win-rates, or because the judges deciding cases were friendlier to President Roosevelt? A creeping concern, likewise, is that the stock of statutes subject to interpretation

122. Yehonatan Givati, *Strategic Statutory Interpretation by Administrative Agencies*, 12 AM. L. & ECON. REV. 95, 96 (2010); Stiglitz, *supra* note 24, at 6.

123. Givati, *supra* note 122; Stiglitz, *supra* note 24, at 13.

124. Ho & Quinn, *supra* note 23.

changes over time, and variably confining statutory language potentially also influences agency win-rates.

A. A Matter of Interpretation

A core difficulty in understanding the relationship between deference law and agency win-rates is that agencies may adapt their interpretations in response to changes in deference law. A more deferential regime may tempt an agency to adopt a bolder interpretation. If that is right, even if a deference regime changes how judges decide agency cases, win-rates may not change because the nature of the agency cases changes. These changes in agency interpretations will not generally be observable or possible to measure credibly.

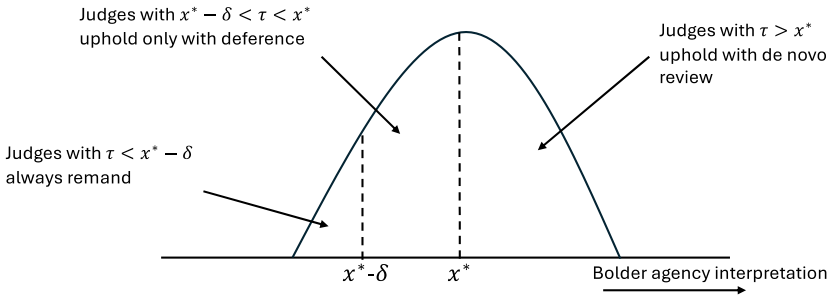
Elsewhere, I examine this issue more formally,[125] and in this Article I highlight the main intuitions. Suppose an agency takes an action that involves an interpretation of a statute, and that interpretation is represented by a point on an interval from zero to one, with higher numbers indicating a "bolder" interpretation. Let $x \in [0,1]$ denote this interpretation. Larger values of x indicate a bolder interpretation of the statute. After the agency takes its action, the courts review it. Many different judges exist in the court system, each with an τ that indicates their tolerance for bold interpretations: A value of $x > \tau$ means that the judge will vacate and remand the agency action. Suppose that the distribution of judges' τ is F, continuous and strictly increasing, with support over [0,1]. A judge from this distribution is selected to hear the case after the agency decides on x.[126] The agency prefers large values of x: e.g., a Democratic EPA might want to aggressively interpret the Clean Air Act to allow for climate-related regulation; or a Republican EPA might want to aggressively interpret the same Act to loosen regulations. The agency payoff is x if the action is upheld and 0 if not.

Suppose initially that the agency sets $x = x^*$ and that judges review *de novo*. If so, the probability that the agency wins is $1 - F(x^*)$. The agency wins if the selected judge has a threshold greater than x and otherwise loses, that is, and the expression reveals the probability the selected judge has a τ greater than x^*. This basic setup is depicted in Figure 2. The figure plots the distribution of judge types or τ with reference to the possible agency interpretations of the statute, with x^* denoting the agency's chosen interpretation. As noted in the figure, only those judges with $\tau > x^*$ uphold

125. Stiglitz, *supra* note 24.
126. This could equally be thought of as a panel of judges being selected.

the interpretation under *de novo*. Judges with lower values of τ reject the interpretation under *de novo* review. The probability the courts uphold the interpretation, therefore, is simply the probability that $\tau > x^*$, which again is given by $1 - F(x^*)$. Notably, if deference regimes do not matter, win-rates will not change—judges will continue to decide cases in the same manner, and the probability an agency wins remains $1 - F(x^*)$.

FIGURE 2. JUDICIAL DECISIONS WITH *DE NOVO* AND DEFERENTIAL REVIEW



Now let us introduce the possibility of effective deference. Suppose initially that agencies do *not* adapt interpretations in response to changes in deference. Effective deference changes judicial review by allowing the agency position to be upheld even if the adopted interpretation exceeds the judge's intrinsic threshold of the acceptable reading of the statute. The regime of *de novo* review can be adapted to reflect deference in the following way: A judge will uphold agency interpretations so long as they fall below $x < \tau + \delta$. Interpretations greater than τ and less than $\tau + \delta$ will be upheld, even though they exceed the judge's intrinsic reading of the statute. This is not to say that the agency has a free pass. An interpretation $x > \tau + \delta$ will be vacated and remanded by the court. The probability that the agency wins the case is now $1 - F(\ x^* - \delta)$. Thus, deference changes the probability that an agency wins a challenge to its interpretation. It does so by altering the votes of a class of judges: those with intrinsic thresholds less than x^* and greater than $x^* - \delta$. These judges would vote to reject the

interpretation according to their best reading of the statute, but they uphold it due to deference.[127]

So far, the discussion takes x^* as a given and assumes that agencies do not adapt their behavior to changes in deference. That is in effect a possible corner solution in a richer account that allows agencies to change their positions. To start that more involved account, recall that the agency prefers bolder, more aggressive interpretations, but is constrained by what a court will uphold. Its payoff is x if the courts uphold the interpretation and zero if the courts reject the interpretation. These payoffs are incorporated into the following utility function, $U(x) = x(1 - F(x))$, which accounts for the likelihood that a given interpretation is accepted by the courts.

The agency's task is to find the interpretation, or value of x, that results in the best tradeoff between boldness and likelihood of acceptance by courts. As detailed in a companion piece, it can be shown that with the earlier assumptions regarding F, the best interpretation is $x^* = \frac{1-F(x)}{F\prime(x)}$, which is a unique interior solution under common assumptions. This interpretation balances the benefits of increasing the boldness of the interpretation against the likelihood that bolder interpretations will be more likely to be set aside by the reviewing court. For example, in the special case where judges are distributed uniformly over the interval, the optimal policy is simply $x^* = \frac{1}{2}$. That is, the agency sets the policy at the midpoint of the distribution of judges.

The question now is what happens under a deferential regime if agencies do adapt their interpretations. As noted earlier, deference operates to enlarge the set of interpretations that a judge will uphold. Instead of upholding only those interpretations with $x < \tau$, the deferring judge will now uphold any interpretation with $x < \tau + \delta$, where δ operationalizes the strength of the deference regime. One easy intuition is that the agency will simply convert or consume the margin of δ with a bolder interpretation. However, that turns out not to be the case. Elsewhere, I show that deference under this setup typically induces the agency to adopt a bolder interpretation,[128] but that it does not *fully* consume the deference in terms of a bolder interpretation. The agency appreciates both bolder interpretations and higher win rates, and it

127. Deference changes votes and agency win-rates in this perspective. This view also embraces the idea that deference regimes operate to de-politicize judging, at least in part. It does so by partially homogenizing the votes of those with divergent intrinsic thresholds, which can be analogized to ideological positions: Judges with τ in the interval of $[x^* - \delta, x^*]$ change their votes from reject to uphold. As I observe in a companion piece, this uniformity in ideology may or may not translate to more uniformity in voting across partisanship, depending on the distribution of ideology over partisanship. Stiglitz, *supra* note 24, at 16–17.

128. *Id.* at 12.

consumes at least part of the deference in terms of a high probability of prevailing during litigation.

An implication of this result is that, in principle, the potential of strategic agency interpretations does not foreclose the possibility of detecting changes in win-rates following changes in deference law. The shift in win-rates will be attenuated relative to the shift without strategic behavior, but there will in theory still be *some* movement in win-rates attributable to the change in deference law. Additional results from this companion theoretical analysis include that agencies with greater appetites for bold interpretations will experience smaller increases in win-rates after an increase in deference.[129] This analysis and set of results indicate the importance of conducting an empirical assessment at the agency-level, if not at a finer level of detail. The effects of a change in deference regimes will depend on how the interpreter values the trade-off between increasing boldness and increasing win-rates. The coarsest level of data that would plausibly hold that tradeoff fixed is the agency level.

The analysis so far considers strategic interpretations and win-rates, but it assumes that there is no litigation selection. A robust literature in the private law space studies how selection affects litigation outcomes, with Priest and Klein's seminal article arguing that changes in legal standards may not affect observed litigation outcomes due to selection. The legal standard may affect which cases are weak, but conditional on the weakness of the case, legal standards do not affect the incentives to finally litigate claims.[130] Regardless of legal standard, it is always the close cases that experience litigation. This focus on selection makes good sense in private law, where the number of potentially litigated disputes is nearly uncountable—selection must be the main story in understanding observed litigation.

As I argue, however, the public law space is very different. There are relatively few possible disputes.[131] Over the decade starting in 2013, Congress produced an average of only one hundred and seventy-seven public laws per year.[132] And many of those public laws merely conveyed

129. *Id.* at 13. This follows from the fact that they convert more of the deference into a bolder interpretation, with less left over for an increase in win-rates. Agencies may have a taste for boldness because they feel that half-measures would be ineffective, or because they are mission-oriented, or because they are ideologically extreme. Similarly, we would expect to see more uniformity in judicial voting over ideology. As a corollary to the first setpoint, the increase in uniformity should be less for decisions involving mission-oriented agencies.

130. George L. Priest & Benjamin Klein, *The Selection of Disputes for Litigation*, 13 J. LEGAL STUD. 1, 4–6 (1984).

131. Stiglitz, *supra* note 24, at 2.

132. *Id.* at 4.

honorifics.[133] Administrative agencies produce more law—on the order of 4,000 rules per year—but many of those rules are minor, technical, or correcting in nature. They produce only about seventy-five "major" rules per year, which have an annual effect of $100 million or more on the economy or otherwise present significant effects for consumers or businesses.[134] Likewise, though they produce many orders, likely only a small fraction of those orders involve meaningful statutory interpretation.[135] It is well within the capacity of the federal judiciary to hear cases on every public law and likely every important agency action. A series of studies, moreover, assess that large fractions of important rules are in fact litigated: An internal EPA audit found that eighty percent of its rules were challenged;[136] other scholars determine that "virtually every" EPA and OSHA rule is challenged.[137] Moreover, perhaps reflecting this same assessment, it is common in studies of interactions between courts and agencies to assume that the agency action will face judicial review.[138] Of course, this does not mean that every rule is, in fact, challenged. Nor does it mean that agencies only adopt interpretations in rules. But as a baseline assumption, it is reasonable to invert the assumption from private law: instead of assuming that almost no dispute is litigated, it is more reasonable to assume that almost every dispute is litigated.[139]

The reason for this is only partially about the number of possible disputes. The nature of the disputes also differs from private law. The disputes tend to be high-stakes, to implicate a diverse array of parties, and to concentrate costs on particular industries or firms. That combination of

---

133. *Id.*

134. *Id.* The "major" designation is legally defined as a rule likely to result in:

> (A) an annual effect on the economy of $100,000,000 or more;
>
> (B) a major increase in costs or prices for consumers, individual industries, Federal, State, or local government agencies, or geographic regions; or
>
> (C) significant adverse effects on competition, employment, investment, productivity, innovation, or on the ability of United States-based enterprises to compete with foreign-based enterprises in domestic and export markets.

5 U.S.C. § 804(2).

135. Relative to rules, data on orders is more fragmented and difficult to account for. As noted later, the core population of interest is agency actions that advance noteworthy interpretations. Those will often but not always be in rules; not every rule or order satisfies that condition.

136. William D. Ruckelshaus, *Environmental Protection: A Brief History of the Environmental Movement in America and the Implications Abroad*, 15 ENV'T L. 455, 463 (1985).

137. KAY LEHMAN SCHLOZMAN & JOHN T. TIERNEY, ORGANIZED INTERESTS AND AMERICAN DEMOCRACY 367 (1986).

138. *E.g.*, Matthew C. Stephenson, *Evidentiary Standards and Information Acquisition in Public Law*, 10 AM. L. & ECON. REV. 351 (2008); Ethan Bueno de Mesquita & Matthew C. Stephenson, *Regulatory Quality Under Imperfect Oversight*, 101 AM. POL. SCI. REV. 605 (2007); Sean Gailmard & John W. Patty, *Formal Models of Bureaucracy*, 15 ANN. REV. POL. SCI. 353 (2012).

139. *See also infra* note 177 and accompanying text.

features makes it likely that someone wishes to challenge an agency rule. Typically, they will challenge for economic reasons: If the compliance costs of a new rule run over $100 million annually, as is possible for a major rule, it will be in the affected industry's interest to challenge the rule, even with very low odds of success.[140] The thin line between administration and politics means that if a challenger is not attracted by economic incentives, they may be by political or ideological incentives. Moreover, unlike private litigation, settlement in cases involving matters of noteworthy interpretation is a dim prospect. An administration is unlikely to withdraw a finalized regulation due to the threat of litigation—more likely, the agency will litigate the claim and defend the regulation.[141] Combined, this means both that a rule is likely to be challenged and that it is likely to be finally litigated.[142]

None of this is to say, again, that every interpretation is, in fact, challenged and finally litigated. Only that, as a baseline assumption, it is a reasonable place to start, whereas it would not be reasonable to assume that every potentially litigated private dispute is litigated. Moreover, even if the assumption is violated, and it is not possible to characterize the features of the rules that do not experience litigation, it is still possible to bound the effects of a doctrinal change with more modest assumptions—that is, to develop a lower and upper bound of plausible effects. The bounding exercise considered below and elaborated on in the Appendix follows the spirit of Manski bounds.[143] The lower bound on win-rates is given by adjusting the win-rate under the assumption that all rules not involved in litigation would have been lost by the agency; the upper bound is given by adjusting under the assumption that all rules not involved in litigation would have been won by the agency. The lower and upper bounds thus provide a sense of the worst and best case scenarios for the agency, and the true result

140. One Government Accountability Office study found that, when compelled to pay plaintiff attorney fees, the EPA paid on the order of $100,000 per payment. U.S. GOV'T ACCOUNTABILITY OFF., GAO-11-650, ENVIRONMENTAL LITIGATION: CASES AGAINST EPA AND ASSOCIATED COSTS OVER TIME 25 (2011). Even at ten times that cost of litigation, challenging a major rule might be worthwhile even if the probability of success is below one percent.

141. This is not to say that agencies fail to settle cases—industry groups often do negotiate and settle with the agency. Cary Coglianese, Challenging the Rules: Litigation and Bargaining in the Administrative Process 127–31 (1994) (Ph.D. dissertation, University of Michigan) (ProQuest). However, though it is difficult to generate an overall assessment, many of these settled cases appear to be ones involving technical issues rather than the matters of statutory interpretation of present interest. As one EPA attorney noted, "[m]y experience is that mostly when we are actually settling issues . . . the issues that were settled were the technical ones." *Id.* at 126. Another difficulty is that administrative data does not directly tie to settlement. The literature regards "voluntary dismissal" as settlement, but it is also possible that these represent instances of a party unilaterally withdrawing a case. *Id.* at 134.

142. Stiglitz, *supra* note 24, at 5–6.

143. Charles F. Manski, *Nonparametric Bounds on Treatment Effects*, 80 AM. ECON. REV. 319 (1990).

is almost sure to be in the interior of those two bounds. The main additional assumption required for this assessment is the number of rules excluded from litigation. In private law, that number would be almost infinite, and so would swamp the number of finally litigated disputes, making the bounded estimate uninformative. In public law, it is reasonable to assume that the number of meaningful unlitigated disputes is a fraction of the litigated disputes, facilitating a helpful bounding exercise. At the agency level, another check on selection involves examining the rate at which they experience litigation. If the volume of litigation decreases, say, after a change in deference law, this may be a flag that selection is at play.[144]

B. Docket Threats

Even if there is little selection into litigation, or even little selection in appeals, there is selection at docketing. The Supreme Court gained substantial docket discretion in the Judiciary Act of 1925,[145] in the earlier part of the period studied and before any of the landmark deference law transitions. This means that docket selection at the Supreme Court level is a constant threat in this analysis. Once granted discretion over its docket, some theory must be derived to understand which cases the justices select to hear. The literature suggests various (typically non-exclusive) possibilities, for example: resolving circuit splits,[146] error correction and reversing lower courts,[147] agreement within the Court on how to resolve a dispute,[148] or opportunities to advance a policy agenda.[149]

How these theories overlay on the relationship between deference law and agency win-rates is complicated and unclear. To consider one strand of this complicated context, after an increase in deference, the agency win-rate at the Supreme Court may decrease, if the lower courts become too permissive and the Court intervenes to fine-tune understanding of the doctrine. Or the agency win rates may increase, if lower courts embrace the new deference law too tepidly, and the Court wants to fine-tune in the other

144. For more on this point, see Stiglitz, *supra* note 24.

145. Judge's Bill, ch. 229, 43 Stat. 936 (1925) (requiring a writ of certiorari for most appeals to the Supreme Court).

146. H. W. PERRY, JR., DECIDING TO DECIDE: AGENDA SETTING IN THE UNITED STATES SUPREME COURT 270 (1991).

147. *E.g.*, Saul Brenner & John F. Krol, *Strategies in Certiorari Voting on the United States Supreme Court*, 51 J. POL. 828, 828–29 (1989); Adam Bonica, Adam Chilton & Maya Sen, *The "Odd Party Out" Theory of Certiorari*, 87 J. POL. 31, 32–33 (2025).

148. *E.g.*, Ryan J. Owens & David A. Simon, *Explaining the Supreme Court's Shrinking Docket*, 53 WM. & MARY L. REV. 1219, 1224 (2012).

149. Robert A. Dahl, *Decision-Making in a Democracy: The Supreme Court as a National Policy-Maker*, 6 J. PUB. L. 279, 280–81 (1957); Kevin T. McGuire & Barbara Palmer, *Issue Fluidity on the U.S. Supreme Court*, 89 AM. POL. SCI. REV. 691, 691–92 (1995).

direction. The relationship between deference law and win-rates may, similarly, be complicated through the docket by the policy agenda of the Court, intra-Court changes in composition and dynamics, or other factors.

There is no obvious way to isolate these complications from the connection between deference law and agency win-rates. Consequently, though the results include Supreme Court estimates, they should be caveated and interpreted with substantial caution. Even if the empirical application is largely successful at connecting deference law and agency win-rates, the strategic joint of Supreme Court docket control introduces an element that cannot be readily accounted for. An increase in agency win-rates following a deferential turn in the law may indicate more about the relationship between the lower courts and the Supreme Court, to continue with the example, than the relationship between deference law and agency win-rates.

C. Temporal Threats

Even if agencies did not adapt their interpretations to deference law, it would still not be straightforward to estimate the effect of deference law on win-rates: For example, over time, the composition of the judiciary changes, the issues being presented changes, the statutes being interpreted change. Any of those changes might explain a difference in agency win-rates before and after a transition in deference law.

Ideally, of course, it would be possible to randomize the deference regime that judges considered—we could then observe the effect of the *Chevron* decision, for example, without worrying about case or court confounders. That is not possible. Where explicit randomization is not possible, the next best alternative is to find some natural experiment. Here, that would require finding a context in which the deference regime did not apply to some subset of courts or cases; those courts or cases could then be used as a "control" group for other courts or cases to which the deference regime did apply. That strategy, too, appears unavailing in this context. The Supreme Court drives the changes in deference law, and those changes apply to, or treat, all cases and courts. There is no obvious control group of cases or courts not affected by the *Chevron* decision.

As a result, it is difficult to design a study that produces a clean estimate of the effect of one deference regime versus another on agency win rates. Differences between two deference regimes may be due to differences in the statutes considered, the agency interpretation, judicial preferences, or the political and economic environment, to name just a few possibilities. Despite these challenges, the importance of the question—do deference

regimes matter?—justifies an effort to try to answer it. In the absence of experiments or natural experiments, the basic strategy is to try to minimize and marginalize concerns by carefully selecting the sample to analyze and by adjusting regressions to account for confounders.

Consider the concerns that can be reliably addressed in this study. A sharp concern is that the composition of the judiciary changes around the time of the doctrinal transitions. The judges deciding cases post-1937, for example, were not only working with more agency-friendly caselaw—many new judges were also appointed by President Roosevelt and may have been predisposed ideologically or personally to support the president who appointed them.[150] This concern can be addressed by including jurist fixed effects. Under this approach, the estimate for the doctrinal effect will be identified on variation in behavior within a judge or justice—that is, it focuses on changes in behavior within the jurist, before and after a transition. Relatedly, the composition of agencies litigating cases may change, and that can be reliably accounted for using agency fixed effects.

Another set of threats relates to confounders that evolve over a long time horizon. A sensible objection is that it is unreasonable to compare agency win rates in 1880 and 1980 because so many features of our legal, political, and economic environment changed over that century. The stock of statutes that agencies interpret changes, the nature of what they regulate changes, the agencies doing the interpretation change, political culture changes, and so on. These changes tend to be slow-moving, but influential, and over a long period, they can jeopardize inferences. Any of these characteristics may plausibly explain changes in agency win rates between 1880 and 1980, rather than deference law.[151]

A standard solution is to include time (say, year) fixed effects, which absorb unobserved year-level variation in these factors. However, that approach is not viable in this doctrinal setting because all judicial decisions within a year would be affected by the doctrinal shift, and it is therefore not possible to identify a separate effect for the doctrinal transition.[152] Failing

150. Ho & Quinn, *supra* note 23.

151. This is the concern that drove Schuck and Elliott strategy: They consider selected nearby years before and after *Chevron*, for instance 1975 and 1988. The advantage of the dataset in this study is that we have the full set of possible years, and it is not necessary to select one or two years, which may not in fact be representative of the before or after period. Both this study and Schuck and Elliott's own data show considerable variation from year to year, elevating the concerns about inferences based on a few isolated years of decisions.

152. One partial exception to this limit is that the doctrinal innovation may apply, in effect, to only certain domains. Choi exploits tax exceptionalism and *Chevron* to show that the doctrine shifted attention that the Department of Treasury devoted to statutory interpretation (as opposed to normative matters). Jonathan H. Choi, *Legal Analysis, Policy Analysis, and the Price of Deference: An Empirical Study of* Mayo *and* Chevron, 38 YALE J. ON REGUL. 818 (2021).

that solution, the study adopts a second-best approach: It examines varying bandwidths of time around putative transitions in deference law. The spirit of this approach resembles that of Schuck and Elliott,[153] though that analysis selects only a few years before and after the reform, years which may or may not be representative of the trends. Here, instead, the measure developed permits examination of various bandwidths of time before and after a transition, inclusive of all years within those bandwidths. By examining only decisions in a caliper around the transition, we limit the threat that slow-moving threats pose to inference. Any characteristic that is invariant within the calipers will be accounted for by the sample restriction. If we regard judicial culture as slow-moving, to take an example, it may be reasonable to assume that it is effectively fixed within that period. And if so, the sample restriction will control for judicial culture. Because more features remain fixed over shorter time periods, the narrower the caliper the more time-invariant features a restriction will capture.

The tradeoff of a narrow bandwidth, however, is that the estimates will be based on a smaller sample of data.[154] For example, if five-year bandwidths are used, the estimate for the *Chevron* regime will be based on decisions from the 1980s. It may be that a wider window of comparison that sweeps in more cases and allows the regime to develop is more appropriate. Many significant decisions of a deference regime, for instance, emerge only years after the marker for the transition—e.g., *NLRB v. Hearst Publications, Inc.* would be outside a five-year window for the judicial revolution,[155] and *Brand X* would be well outside a five-year window for the *Chevron* decision.[156] That must be balanced against the fact that the wider the window, the more unobservable features encroach on the validity of the estimate. Thus, though neither extremely narrow nor wide windows hold much promise, it is unclear what bandwidth is most appropriate. The analysis acknowledges this point by experimenting with a range of possible periods, from five years to forty years.

Even within narrow periods of time, quick-moving changes can threaten inferences. For instance, the party of the president may change around a transition, suggesting that the nature of agency preferences might shift even with a narrow caliper. Looking at the ten years before and after the *Chevron* decision, for example, would group interpretations from the Carter Administration with interpretations from the early Reagan Administration. Changes in win rates before and after the transition might be due to courts'

153. Schuck & Elliott, *supra* note 89, at 990–91.
154. For other periods, the regime is absorbed by the time fixed effect.
155. NLRB v. Hearst Publ'ns, Inc., 322 U.S. 111 (1944).
156. Nat'l Cable & Telecomms. Ass'n v. Brand X Internet Servs., 545 U.S. 967 (2005).

posture regarding the various administrations, rather than the consequences of the deference regime itself. Win-rates may be driven by ideological alignment, rather than deference regimes. Threats of this type can be mitigated through careful use of controls.

D. Learning in Sum

In brief, there is more to learn than might first appear. Theory suggests that agencies do not fully consume increases in deference through bolder interpretations: Part of the deference dividend goes to increased likelihood of winning litigation. This means that it is theoretically possible to detect operative changes in deference law in win-rates. Moreover, litigation selection is not as salient in public law as in private law, and to the extent it remains a concern it is possible to create bounded estimates. Numerous threats remain, but they can be mitigated by examining changes in behavior within jurists and agencies, imposing restrictions on the sample, and controlling for possible confounders. The fundamental empirical philosophy of the study is to account what can be accounted for, and to calibrate understanding of what cannot be accounted for. In this respect, it is likely we can learn more about deference law and agency win-rates through the Courts of Appeals than the Supreme Court, as inferences relating to the latter institution will be clouded by its discretionary docket.

IV. EMPIRICAL APPLICATION: WIN RATES

The guiding strategy of this empirical application is to move from wide to narrow aperture in the analysis. The initial assessments provide a broad overview of litigation outcomes over the last 150 years; subsequent analyses provide narrower, more tightly controlled assessments of the relationship between deference regimes and litigation outcomes, guided by the earlier theory.

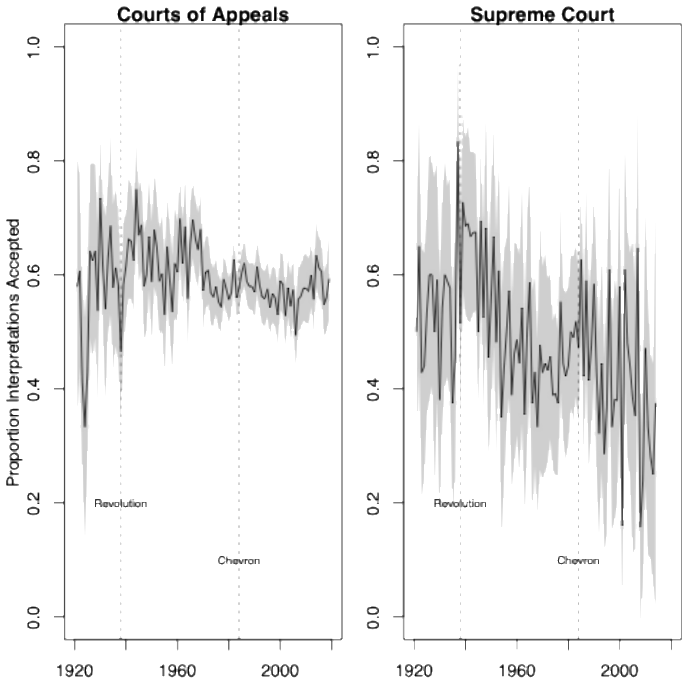
A. Overview of Litigation Outcomes

As a first pass at the data, consider the average agency win-rates by year at the Courts of Appeals and Supreme Court levels. Figure 3 plots the proportion of decisions that favored the agency's statutory interpretation, from 1920 to 2020.[157] The left panel shows the proportion at the Courts of Appeals level; the right panel the Supreme Court level. The dark line

157. The series is truncated because many years prior to 1920 had very few decisions (see also Figure 1).

represents the proportion won by agencies in that year, and the lighter grey regions around the line denote the ninety-five percent confidence interval. The dashed vertical lines mark the points of transition between major deference regimes at the judicial revolution (1937) and the *Chevron* decision (1984).

FIGURE 3. HISTORICAL AGENCY WIN-RATES



At a broad level, the figure shows that the agency win rate tended to hover between forty and seventy percent over the last one hundred years, for both the Courts of Appeals and the Supreme Court. Over the entire series, the average win-rate is fifty-eight percent for the Courts of Appeals and fifty percent for the Supreme Court. Though no existing study examines win-rates over this long of a period, this range of estimates roughly aligns where they overlap in years with other studies. Barnett and Walker find that the Courts of Appeals upheld seventy-one percent of agency interpretations between 2003 and 2013.[158] In the longest-run study, Schuck and Elliott find

158. Barnett & Walker, *supra* note 87, at 1.

that the Courts of Appeals affirmed agencies at rates of fifty-five percent (1965), sixty-one percent (1975), seventy-six percent (1984), eighty-one percent (1985), seventy-six percent (1988).[159] Eskridge and Baer examine Supreme Court decisions between 1984 and 2006 and find, on average, an agency win rate of sixty-nine percent.[160] The algorithmic win-rate is substantially in line with earlier estimates, with considerable year-to-year variability reflected in Figure 3.

This measure can be used to start to break down agency litigation outcomes based on deference regime. Recall the three primary regimes under consideration: the early regime (1875–1937); the transitional regime (1938–1983); and the *Chevron* regime (1984–2024).[161] Figure 4 shows the average agency win-rate within each of those basic deference regimes, for the Courts of Appeals and Supreme Court. The brackets around each bar denote the ninety-five percent confidence intervals for the period. The story this figure tells is one of—on average—basically unchanging agency win rates over time at the Courts of Appeals and declining win rates at the Supreme Court level. The Courts of Appeals win rate stays at about sixty percent; the Supreme Court win rate declines from about sixty percent to fifty percent. Deference progressively expanded through these regimes, from the traditional, nearly *de novo* posture, to the general form of deference embodied in *Chevron*—but if more deference supposedly translates to an easy pass for agencies, it is difficult to see that in this initial run of the data.[162]

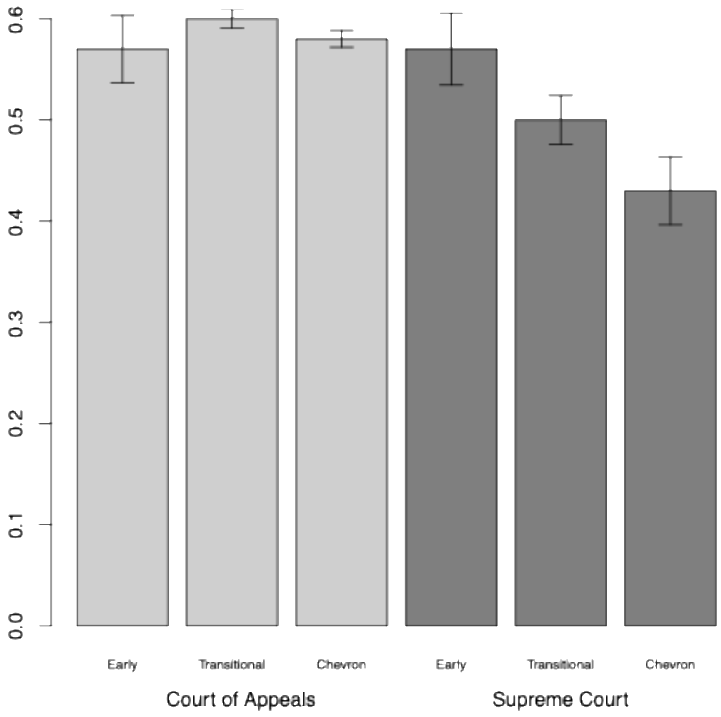
159. Schuck & Elliott, *supra* note 89, at 1003, 1038–39, 1057.

160. Eskridge & Baer, *supra* note 92, at 1129.

161. The data do not extend to 2024, but conceptually the *Chevron* regime runs until *Loper Bright*.

162. The largest jump is at the Supreme Court level, from the traditional to the transitional regime. The increase in win-rates is about three percentage points, with a p-value of 0.17.

FIGURE 4. AGENCY WIN-RATES BY DEFERENCE REGIME



Though suggestive, these win rate estimates should be treated with caution. Ultimately, we want to be able to say something about the relationship between deference regimes and win-rates. Yet the aperture in this initial assessment is too broad to support claims of this type. Even if agencies did not adapt their interpretations in response to deference regimes, other confounders abound: The preferences of the agencies shift over time; Congress creates new agencies, such as the EPA and OSHA, and discards others, such as the ICC; the stock of statutes that agencies interpret changes; the preferences of the courts change; the regulatory issues change; and the political and economic environment changes. All of this means that, though agencies won more during the early regime than the *Chevron* regime, we cannot attribute that change in win-rate to the change in deference regimes. It may be that the difference is due to any of these (or perhaps other) confounders.

B. Average Win-Rates

More reliable estimates can be produced by accounting for confounders. This strategy aims to mitigate concerns by narrowing the sample and statistically adjusting for factors that threaten inferences regarding the relationship between deference law and agency win rates. Broadly understood, this strategy aims to ease concerns related to two primary threats: slow-moving changes in the regulatory, political, or economic environment, which may or may not be observable; and quicker-moving, observable changes, often involving ideological alignment between the agency position and the courts. As noted earlier, restricting the sample to bandwidths around transitions aims at the slow-moving threats; controls for alignment aim at the fast-moving threats to inference within those bandwidths.

With the bandwidths of relevant years, the basic form of the estimated regressions is,

$$w_{iatj} = \gamma_a + \tau_{ji} + \beta Regime_t + \epsilon_{iatj},$$[163]

where w_{iat} is the litigation outcome in case i involving agency a in time period t, and γ_a is an optional fixed effect for agency a, τ_j is an optional fixed effect for jurist *j* (or jurist-direction fixed effect for case *i*). Our main interest is in the β parameter. That parameter tells us the estimated effect of moving from the early to the transitional deference regime or the effect of moving from the transitional to the *Chevron* regime, depending on the sample.

1. Judicial Revolution

Consider first the results related to the judicial revolution of 1937. As reported in Figure 5 for the Courts of Appeals, the results tend to be non-significant, suggesting that the revolution had little effect on the probability that an agency won litigation. Each dot on the figure is the point estimate for a specification, and the lines emanating from that dot represent the ninety-five percent confidence intervals. The dashed horizontal line is at zero, and if the confidence intervals overlap with that line, the estimate is not statistically significant. The grid below the figure indicates the specification relevant to an estimate. For instance, the left-most estimate

163. We can also estimate an event study specification, of the form $w_{iat} = \gamma_a + \sum_{k \neq o} \beta_k D_{iat}^k + \epsilon_{iat}$. The advantage of the event study specification is that it allows tracing of dynamics. However, it is more difficult to present results for many specifications under that approach, and if the average effect is near-zero, dynamics likely hold little interest.

derives from a model with a forty-year bandwidth, and no agency fixed effects; the right-most estimate comes from a model with a five-year bandwidth, agency fixed effects, and judge-direction fixed effects.

Of the twenty models reported, only one is close to statistically significant: that with a thirty-year bandwidth and no fixed effects. The estimate is, with expectations, positive. Read in the context of the other specifications, however, it is difficult to credit this estimate. All other models return with coefficients not close to statistically significant. The main story appears to be that the judicial revolution had little effect on litigation outcomes for agencies at the Courts of Appeals level.

At the level of the Supreme Court, the story is similar—for the most part, there seems to be little effect of changing deference law. The exception to this story is for the narrower bandwidths without fixed effects (on the left of the figure). These estimates indicate substantial positive effects of changing deference law in the period immediately around the transition. These positive estimates most likely reflect the rapidly changing composition of the Supreme Court around 1937. Once we include judge fixed effects, the results recede. This is of a piece with earlier research that finds justices changed their voting behavior and jurisprudence following the judicial revolution.[164] Moreover, as noted, Supreme Court results should be regarded with substantial caution due to the institution's discretionary docket.

164. Edward H. Stiglitz & Rosamond Thalken, Understanding Change in Jurisprudence (Oct. 31, 2025) (unpublished manuscript) (on file with author); Ho & Quinn, *supra* note 23, at 69, 72.

FIGURE 5. WIN RATES FOR JUDICIAL REVOLUTION (COURTS OF APPEALS)

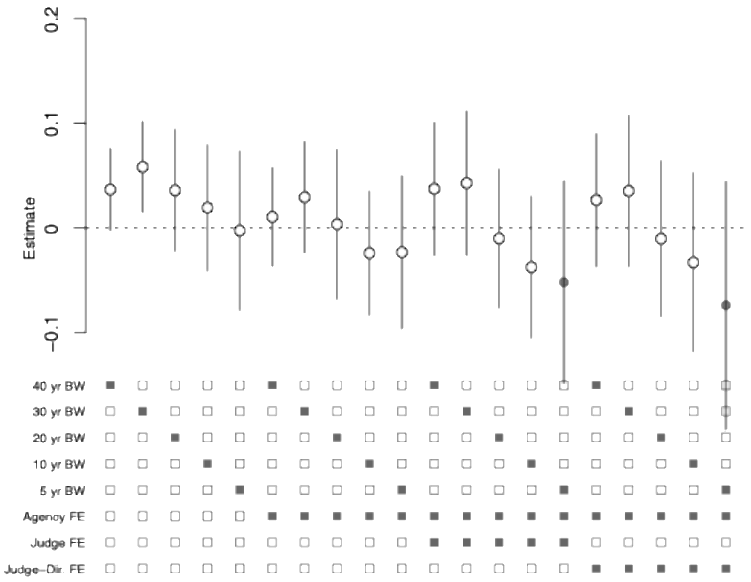
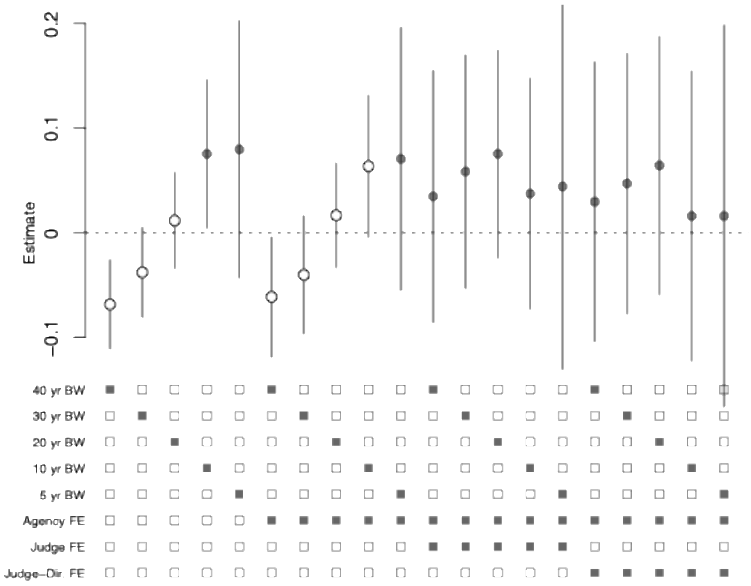


FIGURE 6. WIN RATES FOR JUDICIAL REVOLUTION (SUPREME COURT)



Though the Courts of Appeals coefficients tend to be statistically insignificant, a valid concern with these results is that they are underpowered—the true effect is, say, positive, but we just cannot distinguish it from zero because the estimate is imprecise due to low power. Equivalence tests represent a way to invert this issue of power: The researcher specifies a priori the smallest effect size of interest and then conducts one-sided tests of the null that the effect is at least as large as the effect of interest (EOI).[165] If the test is rejected, one can conclude that the effect size is smaller than the effect of interest. Notably, an underpowered setup will make it more difficult to reject that hypothesis—the null hypothesis is that the effect is outside the EOI. Specifying the EOI is an art and often depends on community expectations. Barnett and Walker, for instance, find that agencies win about twenty percent more often when *Chevron* applies than when *Skidmore* applies.[166] Most would likely agree an effect size of twenty percentage points is of substantive interest—but would an effect of half that size be? A quarter? Rather than speculate on community expectations, this analysis applies an intuitive data-driven approach to determining an effect of interest: It uses the expected absolute difference in win probabilities across two random circuits. This tells us, if we randomly switched circuits, how would our fates change on average?[167] If the measured effect of the doctrinal change is smaller than the geographic variation under the old regime, perhaps the effect is not noteworthy.

Applying this idea to the judicial revolution, the EOI for the Courts of Appeals in the judicial revolution is 0.12. Returning to Figures 5 and 6, the estimates represented in white dots reflect those for which the equivalence test is rejected—meaning that we cannot say that they meet the effect size of interest, or more colloquially that they can be regarded as substantively null. As can be seen from the Courts of Appeals Figure 5, all but two of the estimates are substantively null. About half of the Supreme Court estimates are substantively null, though again, results for the Court should be regarded with caution.

165. This is typically done with two one-sided tests: One tests whether the estimate is larger than the effect size of interest, and another tests whether the estimate is smaller than the negative of the effect size of interest. Though we have theoretical expectation of a positive estimate, this analysis follows that two-test convention. An estimate will be regarded as equivalent only if it rejects both one-sided tests. Daniël Lakens, *Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses*, 8 SOC. PSYCH. & PERSONALITY SCI. 355 (2017).

166. Barnett & Walker, *supra* note 87, at 30.

167. That is, we calculate $EOI = E[|p_j - p_i|]$ for all circuits $j \neq i$.

2. Chevron *Revolution*

Turning to the *Chevron* revolution, the estimates for the Courts of Appeal suggest that the decision had little effect on agency litigation outcomes, as reported in Figure 7. The confidence intervals for virtually all estimates overlap with zero. The results are similar for the Supreme Court, as reported in Figure 8. On the whole, the results from this exercise suggest modest effects, if any, for the *Chevron* revolution.

Using the random-circuit approach as for the judicial revolution, the smallest effect size of interest for the *Chevron* revolution is 0.05. That effect size of interest produces substantively null estimates for all but one of the Courts of Appeals estimates, as shown by the white dots in Figure 7. This supports the idea that the revolution had little effect in the Courts of Appeals. We cannot generally reject the equivalence results in the Supreme Court, however.

FIGURE 7. WIN RATES FOR *CHEVRON* REVOLUTION (COURTS OF APPEALS)

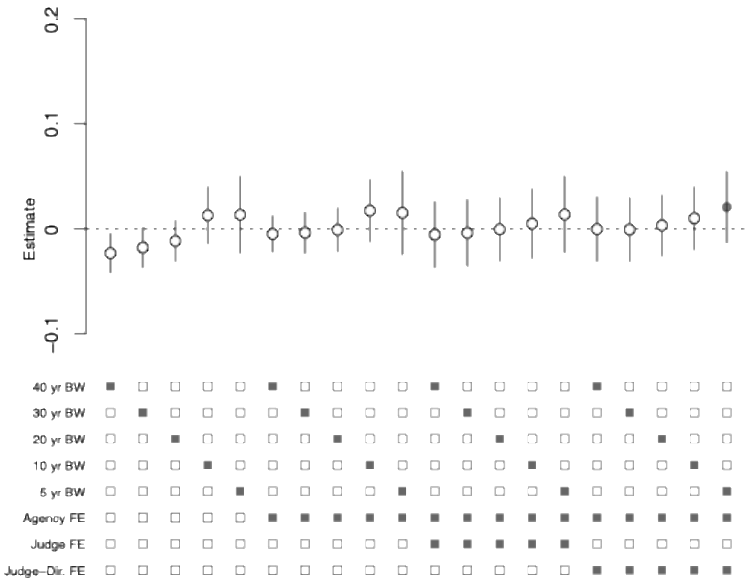
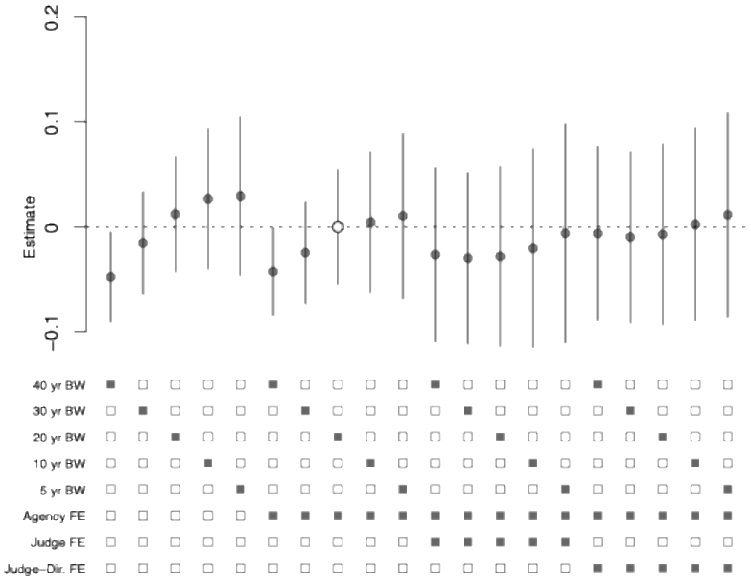


FIGURE 8. WIN RATES FOR *CHEVRON* REVOLUTION (SUPREME COURT)



3. Summary: Average Win-Rates

Agency win rates did not appear to change substantially for the Courts of Appeals following either the judicial revolution or the *Chevron* revolution. Examining a wide range of windows around these transitions, agencies tended to do about equally well before and after the changes in deference law. The pattern is more varied for the Supreme Court. For example, some coefficients relating to the judicial revolution of 1937 return positive in the short term, suggesting a Roosevelt-friendly Supreme Court, a finding that is consistent with earlier research on the judicial revolution.[168] But as noted, results for the Supreme Court should be treated with caution due to their discretionary docket.

C. Agency-Specific Win-Rates

One of the lessons from the theoretical exercise is that inferences may be on surer footing when examining agency-specific results. This follows from the fact that agencies trade off increases in the probability of affirmance

168. *See* Stiglitz & Thalken, *supra* note 10; Stiglitz & Thalken, *supra* note 164; Ho & Quinn, *supra* note 23.

with bolder interpretations differently. The results from the previous Section partially address this concern by including agency fixed effects in some specifications, allowing the estimates to derive from within-agency variation in win rates. Disaggregating the patterns by agency, however, allows for a closer inspection of how deference regimes may affect agency litigation outcomes. It may be that deference regimes detectably change litigation outcomes for some agencies but not for others. The theoretical discussion indicates that there ought to be the largest increases in win rates for the least mission-oriented agencies. An agency-specific analysis, moreover, allows us to probe whether the rates at which agencies engage in litigation change before and after a regime transition. Such a change would constitute a red flag and would suggest that selection into litigation may be complicating inferences.

This agency-specific approach also presents challenges. Because each agency has its own estimate, the number of possible results to report grows greatly. To account for this, the exercise focuses on the window that appeared to show the most likely results (at least for the *Chevron* transition), the five-year window, and on a single, fuller specification. The specification is,

$$w_{iatj} = \gamma_a + \beta Regime_t + \boldsymbol{\Theta} Regime_t * \gamma_a + \epsilon_{iatj},$$

where all remains as earlier, though the specification now includes interactions between the deference regime and the issuing agency.[169] For most agencies, the agency-specific effects will be given by adding the regime coefficient, β, to the agency-specific parameter, Θ_j.[170] The analysis further screens agencies with two filters: An agency must have at least ten decisions in both the pre- and post-transition periods, and the number of agency decisions in the pre- and post-periods must be statistically balanced. The first filter ensures that there are observations involving that agency on both sides of the transition. The second filter aims at identifying agencies for which selection into litigation or a highly dynamic regulatory agenda might be a concern.[171] I report results for agencies that fail the second filter but mark their estimates with asterisks.

As shown in Figures 9 (Courts of Appeals) and 10 (Supreme Court), there is little evidence of widespread increases in agency-specific win-rates

169. Notably absent from this specification are jurist fixed effects. Although in principle it would be possible to include jurist fixed effects, the relatively small number of observations in the agency-specific context means that few judges decide multiple cases in the dataset.

170. That is true for all agencies, except the reference agency; for that agency the estimate will be given simply by β.

171. The balance between periods is tested by chi-square tests.

following the judicial revolution. For every agency that improved in performance before the courts, there is another that fared less well. The FTC, for instance, experienced marginally better litigation outcomes following the judicial revolution, but the FCC was worse off. At the Supreme Court, the USDA appears to perform better post-revolution. As noted, results for the Supreme Court are additionally complicated by the Court's discretionary docket, which introduces another strategic and difficult-to-account-for joint in the assessment.

FIGURE 9. AGENCY-SPECIFIC WIN RATES FOR THE JUDICIAL REVOLUTION (COURTS OF APPEALS)

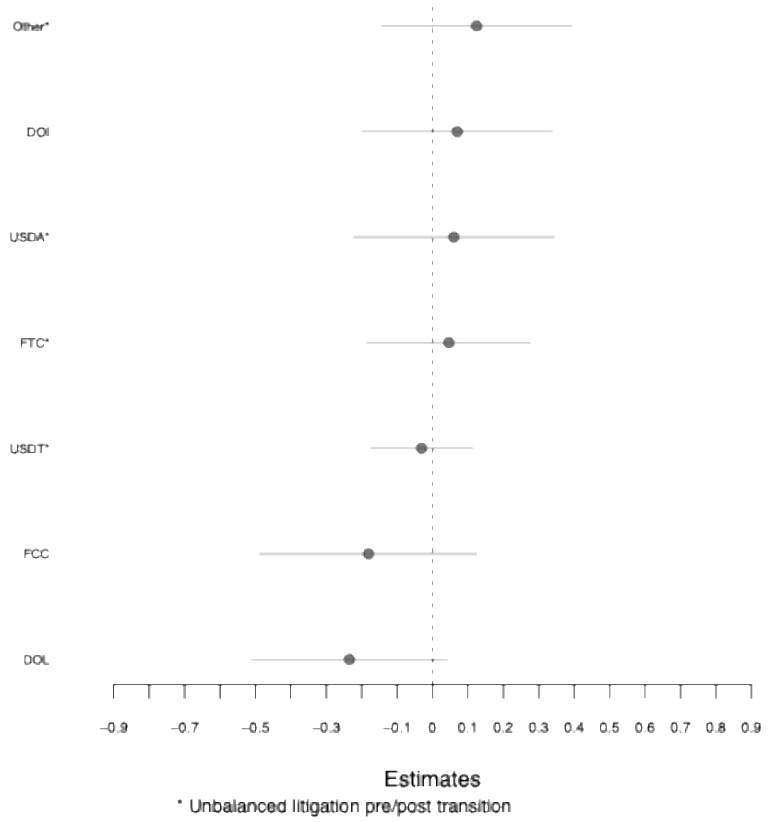
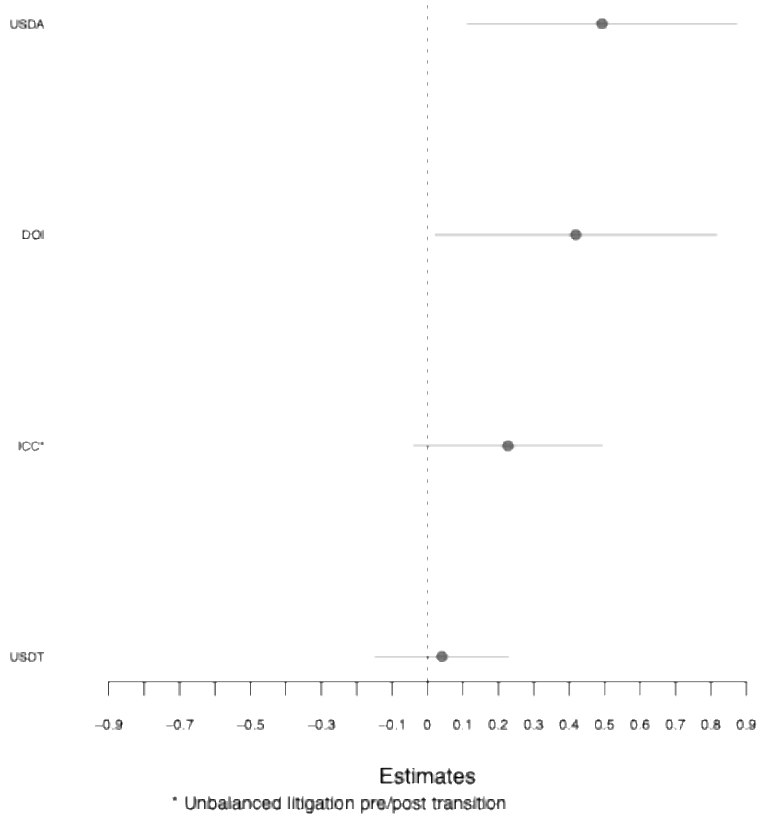


FIGURE 10. AGENCY-SPECIFIC WIN RATES FOR THE JUDICIAL REVOLUTION (SUPREME COURT)



The same basic pattern is evident for the *Chevron* revolution—no strong evidence of widespread increases in agency win-rates, as noted in Figures 11 and 12. At the Courts of Appeals, for every agency that sees an increase in win-rates, another agency sees a decrease in win-rates. At the Supreme Court level, the same story again.

FIGURE 11. AGENCY WIN RATES FOR THE *CHEVRON* REVOLUTION (COURTS OF APPEALS)

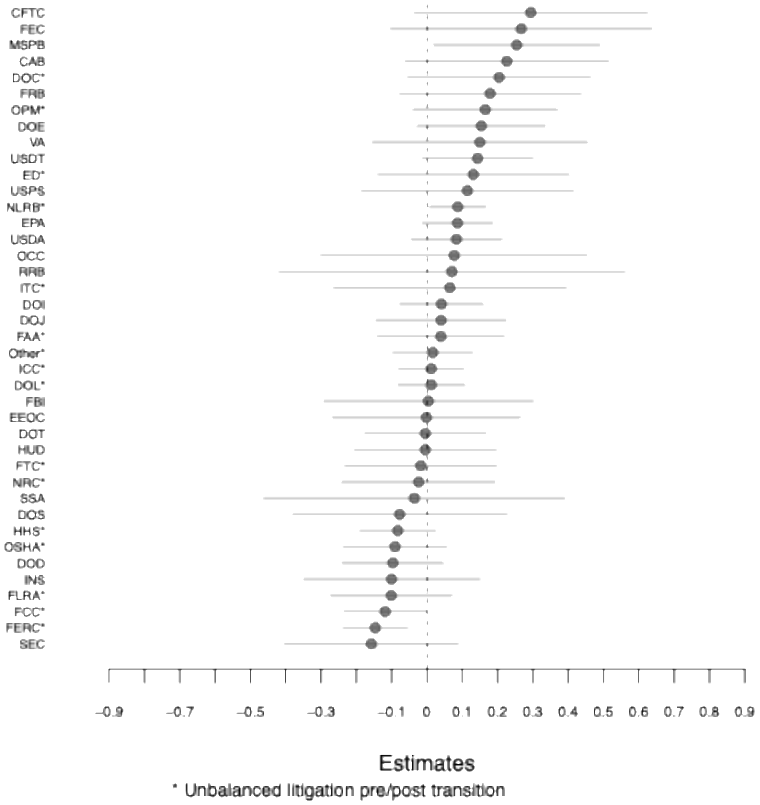
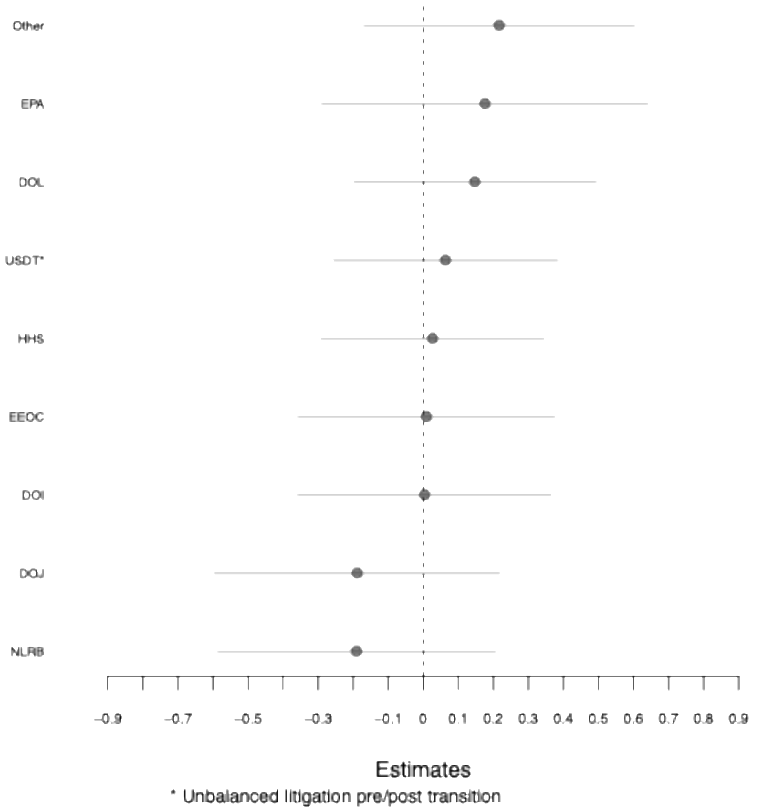


FIGURE 12. AGENCY WIN RATES FOR THE *CHEVRON* REVOLUTION (SUPREME COURT)



D. Calibrating for Selection through Bounds

So far, the analysis has reported roughly one hundred results for different eras, levels of court, and agencies. The thrust of the results is that there is little consistent evidence of a relationship between deference law and agency win rates. Though selection is less of a concern in public law than in private law,[172] it cannot be ignored. Part of the response to selection in the previous Section was to flag as a concern those agencies that experienced different rates of litigation before and after a transition. A bounding exercise is a more general response.

172. Stiglitz, *supra* note 24, at 2.

The idea of a bounding exercise is to make best- and worst-case assumptions about the judicial decisions not observed due to selection and then evaluate win-rates adjusted for these assumptions. The assumptions, to be sure, will be unrealistic. That is the point. They reflect the worst or best possible worlds for agencies. The true estimate, under this approach, is almost sure to fall somewhere between the upper and lower bound estimates. The worst-case scenario for the agency win-rate assumes that the agency would have lost all unobserved litigation. The best-case scenario is that the agency would have won all unobserved litigation. Besides these stark assumptions, the exercise's additional required assumption is the number of decisions selected out of litigation before and after a transition.[173] To probe sensitivity, the exercise examines a range of possible selection rates.

Rather than generate bounds for each of the roughly one hundred results so far, the exercise focuses on two main results to calibrate understanding of selection: the average win-rates for the judicial revolution and the *Chevron* revolution, at the Courts of Appeals level.[174] The Appendix details the bounding approach and derives the best- and worst-case estimates for the agency. Figure 13 shows the bounds for the judicial revolution. The dashed horizontal line in the figure denotes the observed difference between the post- and pre-agency win-rates, an estimate that closely resembles those reported earlier.[175] The solid lines represent bounding estimates. For example, the thinnest lines reflect the bounding estimates when it is assumed that one percent of pre-transition cases are missing. The worst-case bound declines from the starting point, as the proportion of post-transition missing cases increases; the best-case bound increases from the starting point, as the proportion of post-transition cases increases. Thus, at the point

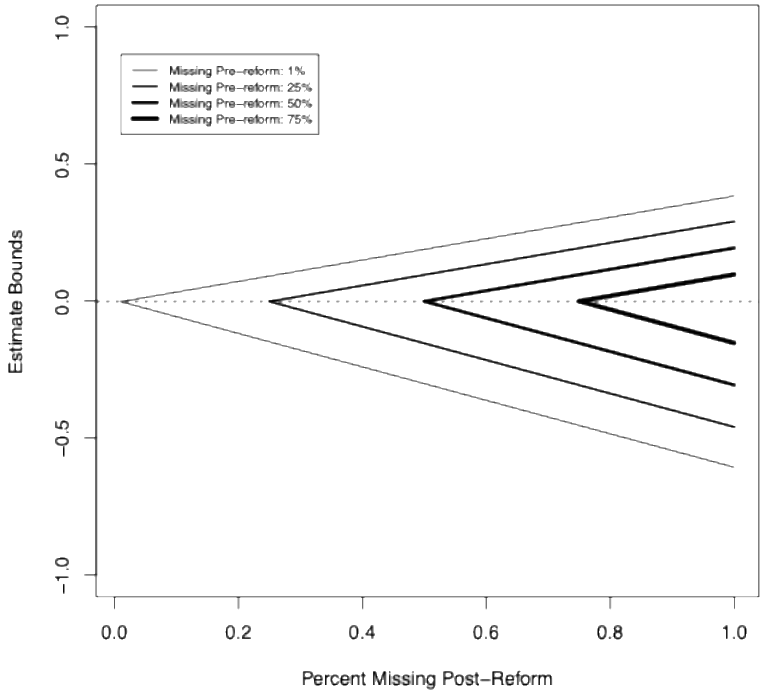
173. Selection is driven by would-be challengers deciding that the costs of litigation cannot be justified given the value and odds of success. The weight of theoretical considerations points toward an increase in selection following an increase in deference. An increase in deference decreases the odds of success for would-be challengers and so enlarges the fraction of agency actions that do not undergo litigation, all else equal. Moreover, because deference enlarges the scope of "safe harbor" interpretations that do not face litigation, it encourages agencies to modify their interpretations to fit into the zone effectively protected by deference. (A safe harbor exists because the presence of litigation costs means that there will be some very modest interpretations not worth litigating, even without deference. It is ambiguous which interpretations would be modified to fit in the safe harbor. On the one hand, moderate interpretations close to the safe harbor can be protected with only modest modification by the agency, suggesting that deference selects out the strongest agency cases. On the other hand, the safe harbor is also attractive for extreme interpretations with a low probability of victory for the agency—the change in interpretation must be large, but the resulting increase in litigation certainty is also large. *See* Stiglitz, *supra* note 24, at 14–15.)

174. Results for the Supreme Court should be interpreted with more caution due to their discretionary docket. *See supra* Section III.B.

175. The estimate for this exercise is based on a five-year bandwidth around the transition. See Appendix for details.

that nearly one hundred percent of the post-transition cases are missing, the best- and worst-case estimates span a large range. As noted in the figure legend, the bolder lines reflect varying assumptions about the extent of missingness in the pre-transition period.

FIGURE 13. BOUNDS ON ESTIMATES FOR THE JUDICIAL REVOLUTION

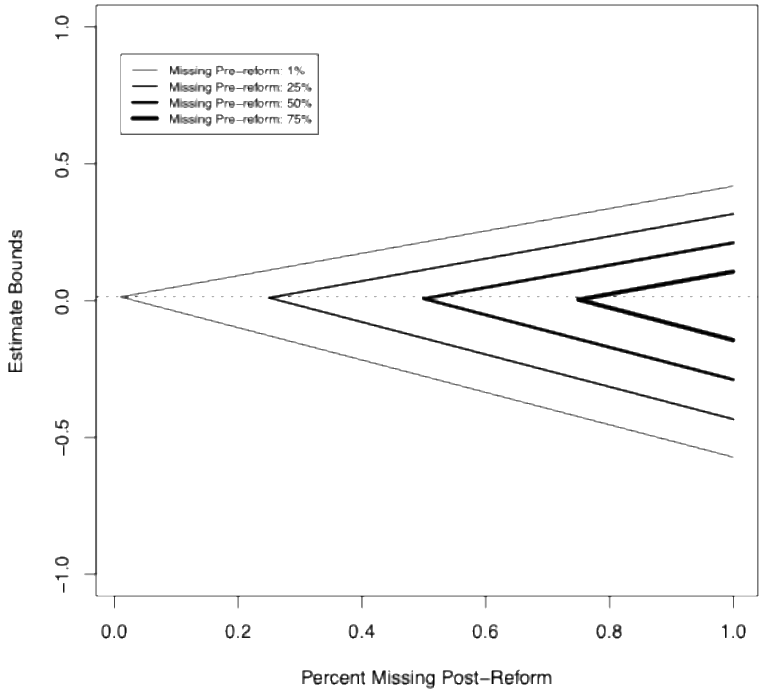


Several observations fall out of this figure and exercise. One lesson is that, if missingness is balanced in the pre and post periods, and unobserved cases go either for or against the agency, missingness is unlikely to substantially influence the estimate.[176] Even if seventy-five percent of cases are missing from the pre- and post-observations, the best and worst case estimates closely resemble the observed estimate. A corollary of this point is that missingness can matter substantially if it is unbalanced in the pre and post periods. One partial intuition, for instance, is that deference makes non-

176. The Appendix considers a more extreme bounding exercise in which it assumes that the missing cases flip for the agency at the time of the transition: E.g., the upper bound will be given by assuming that the agency loses all pre-transition cases and wins all post-transition cases. The exercise reported in the body of the article assumes that all unobserved cases go for (best case) or against (worst case) the agency, in both the pre- and post-transition periods.

litigation more attractive, as it both reduces the odds a challenger wins, and encourages (some) agencies to modulate interpretations so they fit within an effective safe harbor. If right, this implies that missingness may be larger post-transition than pre-transition. The figure illustrates how estimates vary under best- and worst-case scenarios for the agency as rates of missingness diverge in the pre- and post-reform periods. For example, if the pre-transition rate of missingness is twenty-five percent, and the post-transition rate is forty percent, the best-case estimate is an eleven percent effect, and the worst-case estimate is a negative fourteen percent effect. Figure 14 shows the corresponding figure for the *Chevron* revolution.

FIGURE 14. BOUNDS ON ESTIMATES FOR THE *CHEVRON* REVOLUTION



What is a reasonable guess at the level of missingness? A common figure in the literature, which often focuses on the EPA, is that about eighty percent of agency rules face a legal challenge.[177] This figure appears to derive from

177. *See* Lisa Schultz Bressman, *Procedures as Politics in Administrative Law*, 107 COLUM. L. REV. 1749, 1781 n.173 (2007) (noting that "[m]ajor policy decisions . . . rarely evade judicial challenge" and collecting relevant references). However, determining the numerator (number of challenged rules)

a report by President Reagan's head of the EPA, who wrote: "I asked our general counsel's office to do a review of the rules, regulations, and standards which the agency had put out in the last ten years and determine how many of those resulted in litigation. The answer was eighty percent."[178] If that is right, a reasonable baseline for missingness might be twenty percent.

That twenty percent baseline should be regarded as a reasonable guess rather than an estimate or a certain claim. The true denominator of interest is more closely captured by the number of substantive or important interpretations adopted by agencies. The eighty percent figure is premised on a denominator of rules issued, which is likely both under and over inclusive. It is under-inclusive because agencies adopt many important interpretations through orders, especially in the earlier parts of the series analyzed.[179] And it is over-inclusive because the legal category of "rules" is capacious, including corrections, technical amendments, summary approval of state plans, and so on, which often do not adopt meaningful interpretations of statutes.[180]

Another key assumption involves how selection changes via deference law—is there more (or possibly less?) selection after an increase in deference? How much more? If the history of *Chevron* is a guide, the answer is, not much. In fact, there appear to be more Courts of Appeals cases involving agency interpretation of statutes after *Chevron* than before. A search indicates 1,710 Courts of Appeals cases decided in 1983, the year before *Chevron*, and 1,942 cases decided in 1985.[181] If the number of notable interpretations produced in those periods remains relatively stable, this suggests that selection before and after *Chevron* is roughly at parity.

---

and denominator (number of rules or actions) is difficult and contested. In a valuable study, Coglianese and Walters search Westlaw for EPA rules and judicial decisions, and they suggest a lower rate of litigation. Cary Coglianese & Daniel E. Walters, *Litigating EPA Rules: A Fifty-Year Retrospective of Environmental Rulemaking in the Courts*, 70 CASE W. RSRV. L. REV. 1007, 1018–25 (2020). A main problem in this exercise is filtering out rules that consist of mere technical corrections or clarifications, rather than policy decisions or interpretations of statutes. For example, many of the 500-plus final rules for the EPA published in 2011 appear to be almost summary approvals of state implementation plans, or revisions thereof. This comes from a search of the Federal Register database, available at federalregister.gov, filtering for a publication year of 2011, document category of "rule," and agency set to Environmental Protection Agency. The agency issued only forty "significant" rules that year under Executive Order 12,866. The correct denominator likely depends on the research question of interest. Here, the interest is in policy decisions involving statutory interpretation, and so the lower denominator is likely more appropriate.

178. Ruckelshaus, *supra* note 136, at 463.

179. At the same time, most orders likely do not adopt notable interpretations of statues, so simply including them in the denominator would be an over-correction.

180. *See supra* note 177.

181. For 1985, the Westlaw search string was: advanced: ("agency" and (rule or regulation) and (legislation or statute or act)) & DA(aft 12-31-1984 & bef 01-01-1986).

And at least judging by the number of rules issued,[182] the denominator appears relatively stable. Selection may not be substantially distorting the reported win-rate estimates.

E. Limitations

This analysis is difficult and it is important to recognize the limits of this study. Three limits stand out: one relating to measurement and two relating to theory.

First, measurement is difficult and prone to errors and disagreements, whether by human or machine. Previous studies employ humans to determine agency win-rates.[183] This study employs humans augmented by machines for scale. Either way, there will be errors and disagreements in identifying the correct population of cases—is this an agency interpretation case?—and in determining whether the court supported or opposed the agency.[184] It is possible for a court to side with an agency on some issues and not others. The annotation team was instructed to consider the case holistically and to determine whether the agency primarily won or lost. But cases will sometimes be unclear, and in such mixed cases, there will often be room for disagreement even among experts. The same features that confuse humans, moreover, will often confuse machines.[185]

Second, though selection is less relevant in public law than in private law, it remains a concern. The bounding exercises provide useful guardrails, but the correct bounding assumptions are not entirely clear. How much selection is reasonable to assume? Based on historical studies, I propose that roughly twenty percent of notable interpretations select out of litigation, though that and likely any precise figure is contestable. And how much change in selection should we expect after a change in deference law? Based on Westlaw-search empirics that compare litigation rates for rules, I suspect relatively little change in selection. But agencies may adopt interpretations through guidances or orders, and those interpretive vehicles are more difficult to track; if they shift interpretive vehicles contemporaneously with

182. The Westlaw search string was: advanced: (action /5 ("final rule" or "final regulation")) & DATE(aft 12-31-1982 & bef 01-01-1984) % action /5 ("technical amendment" or "correction" or "clarification"). The number of results was 3,902 in 1983 and 3,363 in 1985. It is difficult to closely connect issued rules with final judicial decisions, however, because final decisions lag substantially behind rule issuance.

183. *See supra* Section II.A.

184. The primary consequence of measurement error is to inflate standard errors, a concern that equivalence tests aim to respond to.

185. Rosamond Thalken & Edward H. Stiglitz, Measuring Jurisprudence 18–19 (2025) (unpublished manuscript) (on file with author).

a change in deference law, moreover, the exercise becomes more difficult still.

Another limit relates to the other compositional unobservable—the nature of agency interpretations. Elsewhere, I argue that a theory of public law litigation ought to focus more on strategic primary behavior than selection.[186] When gifted deference, agencies can consume it either by adopting more aggressive interpretations or through greater odds of litigation victory. I show there via a fairly general model that agencies do not fully consume the benefit of deference through more aggressive interpretations; they face incentives to consume part of it in terms of higher win-rates, hedging against the steep losses associated with an adverse litigation outcome.[187] But even so, it remains that agencies consume *some* of the gift via more aggressive interpretation and that the difference in win-rates will be attenuated relative to what it would be without strategic primary behavior.[188] Changes in win-rates will be in principle detectable, in other words, but they will be more difficult to detect. One response to the results of this Section is that they reflect that difficulty, not the immateriality of deference law. Given that underlying agency interpretations cannot be easily quantified or compared, the concern cannot be easily dismissed. What can be said, however, is that we do not find evidence affirming an effect of deference law. Moreover, several features of the results suggest an underlying data generating process of small or null doctrinal effects with random shocks, rather than a story of strategic adaptation: even in the short term, plausibly before agencies have time to fully adapt their interpretations and mode of operations, the coefficients tend to be small; many estimates especially at the agency-level return not just with small but negative coefficients;[189] theoretically, under the strategic adaptation view, the agencies with the largest positive deltas in win-rates should be those with the most technocratic, least mission-oriented agendas, and that pattern is not obviously evident in the agency-level results.[190]

186. *See* Stiglitz, *supra* note 24, at 17–18.

187. *Id.* at 13.

188. How agencies split the gift of deference depends on the distribution of judges and their taste for boldness, neither of which can be directly observed. *Id.* at 14–15.

189. There is little theoretical reason to expect a negative estimate if deference law works as commonly thought.

190. *See* Stiglitz, *supra* note 24. The logic is simple: mission oriented agencies may feel that a half-win is worse than a smaller shot at a total win—for example, the EPA may feel that climate cannot be solved by half measures, justifying aggressive interpretations that are likely to be invalidated.

V. DEFERENCE REALITIES

It is worth asking why deference regimes may not increase agency win-rates—and if they do not, what might.[191] As a framing, consider two prominent non-doctrinal realities: judicial-bureaucratic capacity and constrained opportunistic judicial behavior.[192]

A. Framing Realities

Courts, first, are highly limited institutions. Each judicial chamber consists of a single generalist decision-maker, skilled in legal matters, but a novice in virtually any policy or scientific domain. In the modern court, supporting this decision-maker are one to four law clerks, typically recent law graduates, themselves diligent but less skilled in legal matters, and likewise novices in virtually any policy or scientific domain. Absent native expertise or the ability to conduct investigations, chambers rely on litigant briefing and the incentives of the adversarial system to produce information relevant to the dispute.[193] The adversarial system produces useful information, but parties disagree, and parsing, let alone resolving, those disagreements often requires substantial time and effort. A judicial chamber thus consists of a generalist decision-maker, supported by a thin judicial bureaucracy of generalist law clerks. This is not a high-capacity institution.

In this judicial-bureaucratic context, the simple realities of administration mean that courts may be inclined to defer to agencies in most cases. At inception, *Chevron* itself appears to have been one of these cases. Merrill's history of the decision portrays the justices as burdened by the labor of working through that complicated EPA case, untroubled by deferring to the agency as it relieved them of that burden, and essentially unaware that they were signing on to one of the most celebrated and

191. There is also a question of why so much time is devoted to deference law if it does not relate to agency win-rates. Two responses appear relevant. First, it is possible that deference law only matters little in an equilibrium in which pro- and anti-administration scholars and lawyers spend a great deal of time on those doctrines. In this view, doctrinal positions and arguments reflect a kind of arms race—they do not matter in equilibrium, but that is only because they work to a draw, and if one side defected and stopped devoting resources to the problem, they would lose. Second, it is possible that some doctrines do matter and others do not matter. It may not be possible to determine which is a priori, and as a result, all doctrines receive attention.

192. Both of these realities take a page from realist insights of a century ago. Realism is a high dimensional idea, but at its heart it opposes formalism, or the idea that rules of law drive judicial decision-making. Brian Leiter, *Rethinking Legal Realism: Toward a Naturalized Jurisprudence*, 76 TEX. L. REV. 267, 277–78 (1997). What fills in for these legal rules may be policy preferences, what the jurist views as fair given the facts of the case, or other social forces or institutional realities.

193. Mathias Dewatripont & Jean Tirole, *Advocates*, 107 J. POL. ECON. 1 (1999) (developing a model of information production in an adversarial system).

contested judicial decisions of their generation.[194] Justice Blackmun wrote on the draft opinion "footnotes!", a note that Merrill suggests "may reflect a sense of tedium in having to forge through these complex materials."[195] And on the first page of the draft, Justice Blackmun wrote, "Whew!", which Merrill likewise thought may reflect "a sense of relief that the opinion handled the complicated issue in a way that absolved Justice Blackmun of any further engagement with the matter."[196] Justice Blackmun's sentiments reflected his institutional position and the forces that almost every jurist would be subjected to.

Jurists do not have the time, resources, or bureaucratic support to review every case *de novo* meaningfully. Regardless of the prevailing formal doctrine, this suggests that as a pragmatic matter courts will defer to agencies in most cases. They will be pressed, that is, by institutional constraints into a relatively low-cost judicial decision. Siding with the agency is relatively low-cost because the agency typically is supported by a detailed record in a complex domain, is represented by competent counsel, and reflects the position of at least one political branch of government. Those factors may explain why, even in the early deference regime prior to the judicial revolution, a version of deference was still integrated into the law. The bias, if anything, was pro-agency, even in that more formally *de novo* era.[197] The precise language used to defer will vary over time and jurisdiction: They may "defer" in one era or court, they may give "respect" in another era or court, or "weight" in another court or era, or it may work through facially *de novo* language in another court or era.

But agencies' pragmatic fate before courts—do they win or lose—appears weakly tied to these doctrinal formulations. It may be relatively invariant because courts as institutions face approximately the same constraints on decision-making under the various doctrinal formulations. A substantial part of the reason that doctrinal formulations only very partially resolve agency outcomes is that they are incomplete and plastic. This, too, is an old realist insight that may not apply to all doctrinal domains but appears apt with respect to deference law.[198]

194. *See* Merrill, *supra* note 5.

195. *Id.* at 274.

196. *Id.*

197. This is unlike other areas of law, such as criminal law, where there is sometimes a thumb on the scale against the government. The rule of lenity, for instance, historically favors criminal defendants over the government. Zachary Price, *The Rule of Lenity as a Rule of Structure*, 72 FORDHAM L. REV. 885 (2004).

198. *See* Hanoch Dagan, *The Realist Conception of Law*, 57 U. TORONTO L.J. 607 (2007). As Dagan notes, the extent of indeterminacy is contested, with Hart notably arguing that it is less so than commonly taken.

A second framing impulse follows from that insight. When judges want to depart from the default of deference, they can usually do so, at a cost. This is true in highly deferential regimes. And it is also true in *de novo* regimes. Operating within the original *Chevron* formulation, for instance, a jurist will often be able to decide a case at step one or step two, which turns only on the perceived quantum of ambiguity in the statutory provision at issue. If they resolve the case at step one, the jurist effectively decides the case *de novo*;[199] if they decide at step two, they defer.[200] Even within the supposedly deferential regime of *Chevron*, courts can thus toggle to *de novo* review and readily decide against an agency interpretation. And within a *de novo* regime, the move is even simpler: They merely need to agree or disagree with the agency's position. This is not to say that statutes will be entirely indeterminate, of course, but only that there is usually a pathway available for the motivated jurist. The main constraint on this opportunistic behavior is likely the time and effort required to research the case and deliver a publicly defensible position. Opportunism is constrained, even if not by the deference regime. But where they care deeply, the path is viable.

Where do these two framing thoughts leave us? With a view that agencies will tend to win the cases brought against them. But that judges will opportunistically decide against agencies when they are motivated to do so. This is true across the spectrum of formal deference regimes.

B. Reality Adjustments and Agency Litigation Outcomes

Yet those framing thoughts do not mean that agency win-rates will be invariant. The institutional forces vary over time. The degree of mismatch between the demands on the judiciary and its capacity varies over time, and so too do the imperatives of deference. Likewise, there will be times when the judiciary is predisposed ideologically to agree with an administration's agencies, and times when it is predisposed to oppose agencies. As such, the incentives for jurists to opportunistically manipulate deference law vary with time. Variation in these realities can be used to probe the extent to which they drive litigation outcomes.

---

199. *E.g.*, FDA v. Brown & Williamson Tobacco Corp., 529 U.S. 120 (2000); MCI Telecomms. Corp. v. AT&T Co., 512 U.S. 218 (1994); Am. Bar Ass'n v. FTC, 430 F.3d 457 (D.C. Cir. 2005); Nat. Res. Def. Council v. EPA, 489 F.3d 1364 (D.C. Cir. 2007); Cath. Health Initiatives v. Sebelius, 617 F.3d 490 (D.C. Cir. 2010).

200. *E.g.*, Nat'l Cable & Telecomms. Ass'n v. Brand X Internet Servs., 545 U.S. 967 (2005); Mayo Found. for Med. Educ. & Rsch. v. United States, 562 U.S. 44 (2011); Verizon v. FCC, 740 F.3d 623 (D.C. Cir. 2014); Northpoint Tech., Ltd. v. FCC, 414 F.3d 61 (D.C. Cir. 2005); Nat'l Ass'n of Home Builders v. U.S. Army Corps of Eng'rs, 417 F.3d 1272 (D.C. Cir. 2005).

In the closing Section of this Article, I will point toward what appear to be promising sources of variation in judicial-bureaucratic capacity. The empirical strategy is to find discontinuous expansions in capacity, and to examine (i) whether those expansions change agency litigation outcomes, and (ii) whether that effect is conditional on ideological alignment between jurist and agency position. The intuition underlying these exercises on the expansion is that judicial-bureaucratic capacity increased discontinuously, and workload remained locally relatively stable or increased more continuously, thus providing insight into the relationship between judicial capacity and agency litigation outcomes. The theoretical framing discussion suggests that more judicial-bureaucratic capacity may, on average, tend to reduce the degree of deference to agencies, driving down agency win-rates. Part of the reason jurists defer is that, like Justice Blackmun apparently in *Chevron*, they find the cases bothersome and tedious.[201] Increasing the ability of jurists to handle complexity and tedium should negatively affect agency litigation outcomes. This decrease in agency win-rates ought to be larger in cases where there is misalignment between the jurist and the agency position.

Judgeships and clerkships represent two important levers of judicial-bureaucratic capacity. For a given aggregate caseload, the more judges that sit on the bench, the fewer cases they must each decide. This lower caseload means judges will have more time to deliberate and to write their opinions, which implies that they will be less constrained by the imperatives of deference and relatively free to pursue policy or ideological objectives. Clerkships represent a closely related, second judicial-bureaucratic lever. As a leading history of the institution of the clerkship put it, "[t]he clerkship represents the judiciary's response . . . to the press of the cases . . . . For better or worse, the clerkship has proved the invariable, now deliberate, response to the growth of appellate case loads throughout the country."[202] Law clerks do every task within a judicial chamber from gathering citations,

201. *See supra* Section V.A.

202. Paul R. Baier, *The Law Clerks: Profile of an Institution*, 26 VAND. L. REV. 1125, 1132 (1973); *see also* Todd C. Peppers, Michael W. Giles & Bridget Tainer-Parkins, *Inside Judicial Chambers: How Federal District Court Judges Select and Use Their Law Clerks*, 71 ALB. L. REV. 623 (2008); John Bilyeu Oakley & Robert S. Thompson, *Law Clerks in Judges' Eyes: Tradition and Innovation in the Use of Legal Staff by American Judges*, 67 CALIF. L. REV. 1286 (1979); John G. Kester, *The Law Clerk Explosion*, LITIG., Spring 1983, at 20; J. Daniel Mahoney, Foreword, *Law Clerks: For Better or For Worse?*, 54 BROOK. L. REV. 321 (1988); Chad Oldfather & Todd C. Peppers, *Judicial Assistants or Junior Judges: The Hiring, Utilization, and Influence of Law Clerks*, 98 MARQ. L. REV. 1 (2014).

to drafting and trouble-shooting draft opinions.[203] Clerks thereby expand judicial-bureaucratic capacity and operate to a similar effect as increasing the number of judges.

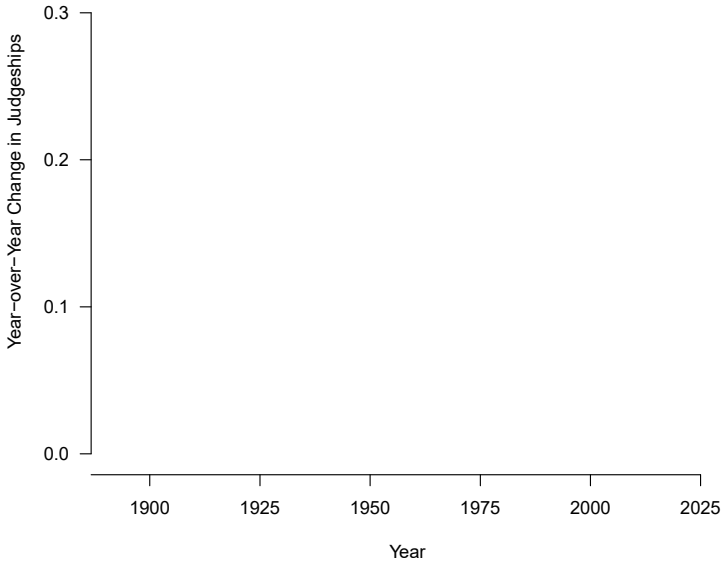
Consider first judgeships. The number of authorized federal judgeships has increased over time. Most of this historical increase is accretive: A circuit is given an additional judge in year 1, another circuit given an additional judge in year 2, and so on.[204] Figure 15 shows the year-over-year change in authorized judgeships from the creation of the U.S. Courts of Appeals in 1891 to the present. Thus, a value of, say, 0.10 indicates that the number of judgeships in that year was ten percent larger than the previous year. As can be seen from this figure, 1978 stands out as a year of remarkable change in the judiciary, with the number of authorized judgeships increasing by roughly thirty-five percentage points through the Omnibus Judgeships Act of 1978.[205] That increase is approximately twice the increase of any other year since 1891. The question of interest, then, is what happens to the relationship between courts and agencies when courts experience a sudden, thirty-five percent increase in their capacity?

203. Because clerks often partially supplant the judicial role—as by drafting opinions—they have been a controversial institution. *See, e.g.*, Mahoney, *supra* note 202; RICHARD A. POSNER, THE FEDERAL COURTS: CRISIS AND REFORM (1985).

204. *Chronological History of Authorized Judgeships - Courts of Appeals*, U.S. CTS., https://www.uscourts.gov/about-federal-courts/about-federal-judges/authorized-judgeships/chronological-history-authorized-judgeships-courts-appeals [https://perma.cc/PKK9-XF84].

205. *See* Omnibus Judgeships Act of 1978, Pub. L. No. 95-486, 92 Stat. 1629.

FIGURE 15. YEAR-OVER-YEAR CHANGE IN AUTHORIZED JUDGESHIPS



The history of clerkships is more complicated and less well-documented. Most of the increases to clerkships appear to be, like judges, accretive in nature, with specific circuits or judges receiving extra funding for law clerks as they petition either Congress or the Judicial Conference of the United States, which was delegated substantial control over clerkship numbers and compensation in 1983.[206] The appropriations bills tend to speak in general terms of law clerks and funding, rather than in terms of authorized clerkship positions, and the Judicial Conference does not appear to publish their changes in clerkship policy, which makes tracking clerkship numbers over time difficult relative to judgeships.[207] An exception to this obscured and

206. Act of Nov. 28, 1983, Pub. L. No. 98-166, 97 Stat. 1071, 1099–1100 (setting a total appropriation for judicial support staff, "*[p]rovided*, That the secretaries and law clerks to circuit and district judges shall be appointed in such number and at such rates of compensation as may be determined by the Judicial Conference of the United States."); *see also* Peppers et al., *supra* note 202, at 628.

207. To be sure, it is possible to speak in general terms of the trends historically. *E.g.*, POSNER, *supra* note 203, at 72 ("[in the 1930s] circuit judges each had one . . . in 1970 the circuit judges got a second . . . in 1980 the circuit judges got a third."); *see also* Kester, *supra* note 202, at 22 (remarking that "[f]ederal circuit judges were first authorized funds for law clerks in 1930; in 1970, they were allotted two, and in 1980, that grew to three each."); Donald P. Ubell, *Evolution and Role of Appellate Court Central Staff Attorneys*, 2 COOLEY L. REV. 157, 158 (1984) (noting that "all judges of the United States Court of Appeals had a law clerk by the 1930's. . . . The Court of Appeals' judge got his or her . . . second law clerk in 1969 and a third in 1979."). Those time markers may be generally accurate, but they were not accompanied by source citations, and they do not all appear to be traceable to

variegated pattern occurred in 1930, when Congress authorized for the first time a single law clerk for each circuit judge.[208] This change in authorization can be used to study the relationship between law clerks and agency litigation outcomes.

One advantage of the reforms to judgeships and law clerks is that, unlike the reforms to the deference doctrine, a plausible "control" group exists. For the main exercises, the Supreme Court serves as an indicator of trends in judicial behavior absent the reform. The number of Supreme Court justices remained fixed throughout the series;[209] likewise, the number of law clerks authorized for the Supreme Court did not change around 1930, though a fourth clerk might have been added in 1978.[210] Future work may examine these and other changes in judicial-bureaucratic capacity to study their effects on the relationship between courts and agencies.

VI. SPECULATIONS: *LOPER BRIGHT* AND AN AGENDA

This Article is an historical-empirical analysis. However, it was initially motivated by *Loper Bright* and the question of how that landmark decision might change the relationship between courts and agencies. What can we learn from history about this new era? The Article suggests a two-fold answer.

---

documentable, widely applicable changes in authorization by Congress, Judicial Conference, or other authority.

208. Act of June 17, 1930, Pub. L. No. 71-373, 46 Stat. 774 (revising the Judicial Code to provide that "[e]ach United States circuit judge is hereby authorized, with the approval of the Attorney General, to appoint a law clerk, whose salary shall not be in excess of $3,000 per annum; and the appropriation of such amount as is or may be necessary to pay the salaries and travel expenses of such law clerks is hereby authorized."). District courts received their first authorization in 1936. *See* Act of Feb. 17, 1936, Pub. L. No. 74-449, 49 Stat. 1140.

209. The most recent change in the number of justices was in the Judiciary Act of 1869, which stabilized the number of justices at nine. Judiciary Act of 1869, ch. 22, 16 Stat. 44.

210. The Supreme Court won authorization for a law clerk in 1919. *See* Act of July 19, 1919, ch. 24, 41 Stat. 163, 209 (providing funding "[f]or nine law clerks, one for the Chief Justice and one for each Associate Justice, at not exceeding $3,600 each, $32,400."). A second law clerk appears to have been authorized in 1947. Chester A. Newland, *Personal Assistants to Supreme Court Justices: The Law Clerks*, 40 OR. L. REV. 299, 314 (1961). A third clerk appears to have been regularized by 1970. *See* Kester, *supra* note 202, at 22; *see also* Baier, *supra* note 202, at 1133 (noting "an allotment of three law clerks to each Supreme Court Justice," and citing to a 1972 address by Chief Justice Burger). A fourth appears to have been authorized in 1978. *See* Kester, *supra* note 202, at 22. However, the articles claiming these increases to two, three, and four clerks did not cite to a statute or administrative decision, so it is difficult to fully assess these changes. As the control group received a "treatment" in 1978 that may move outcomes in the same direction as introducing more Courts of Appeals judges, the resulting estimate might be considered a conservative assessment.

A first lesson is that doctrinal change itself seems unlikely to result in a substantially different relationship between the courts and agencies.[211] They may lose more after *Loper Bright* than before. But that is unlikely to be due to the change in doctrine. Even on narrow doctrinal terms, the Court had been whittling away at *Chevron* for over twenty years, at least since *Mead* limited its reach and reasserted *Skidmore* deference as a central doctrine.[212] Not long before *Loper Bright*, the Court amplified the major questions doctrine to not just neutralize *Chevron* deference on important questions, but to put a thumb on the scale against agency authority in such cases.[213] Citations to *Chevron*, moreover, were declining well before *Loper Bright* in cases involving agency statutory interpretation. Figure 16 plots the proportion of agency cases in the Courts of Appeals that cite *Chevron*.[214] As can be seen there, the rate at which judges cited *Chevron* had been declining for at least ten years prior to *Loper Bright*. If this pattern persisted, *Chevron* may indeed have experienced something like a natural death within a few years, even without *Loper Bright*.[215]

More to the point, however, the results from the present analysis suggest that *Chevron* likely had little effect on litigation outcomes even in its youthful vigor. Judges did not appear to decide cases differently post-*Chevron*. Though framed as a deferential regime, there was significant room within the confines of the doctrine to decide cases against agencies when motivated to do so.[216] This is not again to say that all doctrinal formulations exert little influence on lower court judges—only that, in this instance, the doctrine was conceptually capacious and did not appear to constrain judges substantially. In this respect, *Chevron* keeps good company in deference law. There is likewise little evidence that earlier deference formulations notably influenced litigation outcomes for agencies.

---

211. For similar positions, see Adrian Vermeule, *The Old Regime and the* Loper Bright *Revolution* (Harvard L. Sch. Pub. L. Working Paper, Paper No. 25-02, 2024), https://ssrn.com/abstract=5049347 [https://perma.cc/5TX2-BYLV] (suggesting that this the ancient regime of judicial decision-making will survive, with changes in language); Lisa Shultz Bressman, *Lower Courts After* Loper Bright, 31 GEO. MASON L. REV. 499 (2024) (predicting that, if *Loper Bright* overrules *Chevron*, it will likely not influence lower court decisions substantially).

212. United States v. Mead Corp., 533 U.S. 218, 221 (2001).

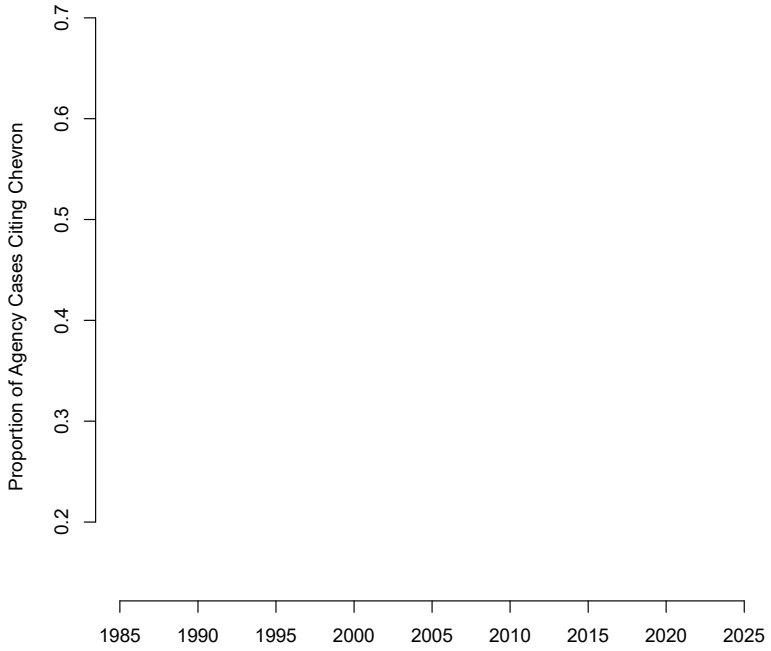
213. West Virginia v. EPA, 597 U.S. 697, 732 (2022) (requiring the government to "point to 'clear congressional authorization' to regulate").

214. This exercise is based on repeated searches of Westlaw, restricting attention to specific years and to the Courts of Appeals. The search string for *Chevron* citations: "adv: "467 U.S. 837"". The search string for agency cases: "advanced: (("agency" OR "board" OR "commission") & ("statute" OR "act" OR "legislation" or "law" or "statutory"))".

215. *See* Jellum, *supra* note 92.

216. *E.g.*, FDA v. Brown & Williamson Tobacco Corp., 529 U.S. 120 (2000).

FIGURE 16. PROPORTION OF AGENCY CASES CITING *CHEVRON* (COURTS OF APPEALS)



The later parts of this Article suggest another orientation: Deference realities may help understand agency litigation outcomes. With that perspective, caseloads remain high by historical standards, though they appear to be declining recently.[217] This suggests that even those judges predisposed to decide against agencies will face pressure to defer in the mine run of cases. Opposed judges will not have the time or bureaucratic resources to justify decisions against agencies. Whether through "respect" or "deference," the less costly and less risky option for judges will generally be to acquiesce to agency positions. At the same time, as before the demise of *Chevron*, opposed judges will opportunistically expend their limited resources to deny agencies' range of activity where they feel the issue is

217. The appellate caseload was about 50,000 in 2015 and 2020; in 2024, it was about 40,000. *See Federal Judicial Caseload Statistics 2015*, U.S. CTS. (Mar. 31, 2015), https://www.uscourts.gov/data-news/reports/statistical-reports/federal-judicial-caseload-statistics/federal-judicial-caseload-statistics-2001/federal-judicial-caseload-statistics-2015 [https://perma.cc/H625-QRS2]; *Federal Judicial Caseload Statistics 2020*, U.S. CTS. (Mar. 31, 2020), https://www.uscourts.gov/data-news/reports/statistical-reports/federal-judicial-caseload-statistics/federal-judicial-caseload-statistics-2020 [https://perma.cc/VSX9-MZQT]; *Federal Judicial Caseload Statistics 2024*, U.S. CTS. (Mar. 31, 2024), https://www.uscourts.gov/data-news/reports/statistical-reports/federal-judicial-caseload-statistics/federal-judicial-caseload-statistics-2024 [https://perma.cc/W8EJ-QDCL].

important. These realities point to continued deference in most cases, with judicial sniping in a minority of cases. That minority of cases will likely tend to be the most important and controversial subset of cases. But that is how it was prior to *Loper Bright*, too.

These realities further point to both a research and advocacy agenda. The legal literature is oriented to doctrinal formulations: Is *Chevron* a positive or negative? How can it be justified? How should we regard any of the various tweaks to the doctrine, from *Mead* to *West Virginia v. EPA*? The literature is helpful and insightful on these questions, but it neglects deference realities—the judicial-bureaucratic forces that may more substantially shape judicial behavior. To be sure, as an offshoot of the realist insight, legal scholars and social scientists have long recognized that judicial ideology bears on decision-making.[218] But less investigated is the descriptive relationship between judicial-bureaucratic features and decision-making. When the judiciary has more capacity, do judges behave differently? Do they defer more to agencies or snipe less opportunistically? Do clerks affect judges' tendency to defer or behave opportunistically?[219]

If those intuitions survive and grow, they also point to an advocacy agenda. Advocates, again, steep in judicial doctrine, but less in the details of the judicial-bureaucratic machinery. Yet it may be those details that influence decision-making and the relationship between courts and agencies more than doctrinal formulations. Those institutional levers would touch every aspect of judicial business, not only the agency-court relationship, so advocates would need to integrate over the full range of relevant legal issues. But debates and analysis of the relationship between legal determinations and judicial-bureaucratic levers, this analysis indicates, ought to be more central to the conversation.

CONCLUSION

What is the relationship between deference law and agency litigation outcomes? Based on this analysis, the answer is that it is hard to find much evidence of a robust effect. Exploiting a novel dataset of litigation outcomes that spans the late nineteenth century to almost the present, it is not easy to find evidence of a consistent or widespread influence of widely discussed doctrinal regimes on litigation outcomes. This is true of doctrinal changes

218. *E.g.*, Miles & Sunstein, *supra* note 90 and citations therein.

219. There is a small and innovative literature that examines how clerk ideology influences judicial behavior. *See* Adam Bonica, Adam Chilton, Jacob Goldin, Kyle Rozema & Maya Sen, *Legal Rasputins? Law Clerk Influence on Voting at the US Supreme Court*, 35 J.L. ECON. & ORG. 1 (2019). However, there is no quantitative literature to my knowledge on the question of how clerks affect judicial decision making through the expansion of judicial-bureaucratic capacity.

around the judicial revolution of 1937, and it is true of the *Chevron* revolution some half a century later.

Assessing the relationship between deference law and litigation outcomes is challenging for the reasons explained in the Article: E.g., it is not clear how to identify the relevant subset of decisions, no measures of litigation outcomes exist, and one must worry about judicial selection and unobservable strategic behavior. The fundamental empirical strategy of this Article is to use modern language models to develop data on a scale not previously feasible, to control and account for what can be accounted for, and to calibrate understanding of what cannot be accounted for. Under this approach, there will be a residual uncertainty about the relationship between deference law and litigation outcomes, but we try to bound it.

The later parts of the Article bid to re-orient attention toward deference realities, or the judicial-bureaucratic features that influence litigation outcomes. The analysis sketches for future analysis two case studies: the expansion of clerks to the Courts of Appeals in 1930 and the expansion of the judicial workforce in 1978. It is possible that much of the sweep of deference may be understood as the product of time-pressed, institutionally limited judges seeking to make their way through a crowded docket.

APPENDIX

A. Measurement Appendix

This Section details the methods and validation related to the measures analyzed in the study.

1. Identifying Cases with Agency Statutory Interpretation

Departing from earlier studies,[1] the analysis algorithmically identifies cases involving statutory interpretation. The algorithm requires annotated data to learn the patterns in language relevant to the cases. For this data, I rely on Westlaw Headnotes: I collect all cases that match the Headnote of "Administrative Construction of Statutes," a sub-feature of Administrative Law and Procedure in their classification system (15A-k2201).[2] That produced a population of roughly 1,600 District Court, Courts of Appeals, and Supreme Court decisions between 1875 and 2020. The cases concentrate in the later parts of the series, likely both because the volume of relevant cases increased over time, and Westlaw invested more resources in classifying recent cases. For this reason, I overweight the earlier periods: Twenty-three percent of the complete Headnote sample relates to cases decided before 1984; thirty-seven percent of the annotated sample relates to cases from this period. In total, the Westlaw sample used for fine-tuning includes just over 700 decisions, with 263 from before *Chevron* and 456 from the post-*Chevron* period. Entries in that sample serve as the affirmative examples of cases involving statutory interpretation. For the negative samples, I first restrict the full sample to decisions (a) not in the Westlaw Headnote list; (b) decided post 1875; (c) that do not mention terms related to statutory interpretation and an agency in the same paragraph;[3] and (d) contain more than 400 words in their majority opinion. I then create a random sample stratified by the level of court (District Court, Courts of Appeals, and Supreme Court), with sample size equal to twice the number of the sample from the Headnote sample. The negative cases are over-sampled so as to allow the model to learn the highly diverse nature of non-statutory cases. The first 512 tokens from the majority opinions of these decisions constitute the raw data for the algorithm.[4]

---

1. Earlier scholars either read the universe of cases in a narrow domain, or they key of judicial citations. Neither approach is viable in this study.

2. I rely on Westlaw only to identify the relevant cases—case text derives from Harvard's Caselaw Access Project.

3. The regular expression looks for (a) "(statute|legislation|\sact[\.|\s|,|;]).{0,50}(mean|constru|interpret|reading|understand)" and (b) "(constru|interpret|reading|understand|mean).{0,50}(statute|legislation|\sact[\.|\s|,|;])".

4. BERT models consider context windows of a maximum of 512 tokens.

Using that sample, I fine-tuned a version of legal-BERT,[5] which is a lightweight, domain-adapted large language model found to be highly capable in earlier research.[6] The model is fine-tuned for four epochs at a learning rate of 4e-5 and a warmup ratio of 0.1, using eighty percent of the annotation data for training, and twenty percent held out for validation.[7] The validation sub-sample was not used train the model and allows us to examine model performance—does it identify the cases that Westlaw identified as relating to (or not relating to) agency statutory interpretation?

Based on this exercise, the model performs remarkably well. It correctly identified the class of ninety-three percent of the cases in the validation data. The (unweighted) macro F1 score, which provides an overall sense of the false positives and negatives, is 0.92, high by virtually any standard. That high F1 score reflects strong performance in both classes: The model identifies agency statutory cases well; critically, it also identifies non-statutory cases well. Table A1 reports the classification performance table for this validation exercise. In total, these results indicate that the model competently scales Westlaw's relevant Headnote classification to the wider universe of judicial decisions examined in this study.

TABLE A1. MODEL PERFORMANCE: AGENCY STATUTORY INTERPRETATION

Class	Precision	Recall	F1-score	Support
Non-statutory	0.96	0.93	0.95	307
Statutory	0.86	0.92	0.88	130
Macro-average (unweighted)			0.92	

2. Identifying Litigation Outcomes

The second main measurement task is to determine whether the agency wins or loses a case involving its interpretation of a statute. This is a more challenging task because there is no obvious existing data that would be suitable for model training. Existing efforts to code for litigation outcomes consider sharply limited jurisdictions (e.g., only the Supreme Court),

5. Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras & Ion Androutsopoulos, *LEGAL-BERT: The Muppets Straight Out of Law School*, 2020 FINDINGS OF THE ASS'N FOR COMPUTATIONAL LINGUISTICS: EMNLP 2898.

6. Thalken et al., *supra* note 10.

7. Because the model achieved stable performance across random splits in preliminary tests, I report a single 80/20 train–validation split here. For the more challenging litigation-outcome task below, I use five-fold cross-validation to provide a more reliable estimate of generalized performance.

limited time periods (e.g., a few years before and after *Chevron*), or limit attention to cases based on case citations (or some combination of these three limits).[8] We want annotation data that will allow a model to learn relevant language that (a) covers many jurisdictions, both appellate and Supreme Court; (b) covers a long series, ideally reaching from 1875 to roughly the present; and (c) is not filtered through case citations, particularly as case citations will not be stable over a series that extends this deeply into history.[9]

To develop these novel annotated data, I assembled a team of upper-year Cornell law students with relevant coursework. The set of annotation cases is the same used to identify decisions involving agency statutory interpretation: A sample of the cases that Westlaw identifies using their Headnotes for agency statutory interpretation, discussed in the previous Section. At the start of the semester, the team was provided a codebook, closely adapted from that in Eskridge and Baer's study.[10] Team members were tasked with determining whether a judicial decision accepted or rejected the agency's interpretation of the statute. Each week, a random sample of the annotation cases would be assigned to each member, with a subset assigned to all members. The team did not know which cases were assigned solely to them and which were assigned in common to the team. The common cases allowed examination of inter-coder reliability, the extent to which different team members agree on the coding of the same case. Low inter-coder reliability may indicate that the construct is poorly defined or too difficult for humans to reliably assess, given the source material.[11] To make their determinations, team members were assigned a citation and instructed to make a holistic human assessment. They recorded their entries in a computer interface. During weekly meetings, the team would discuss difficult cases and refine the application of the codebook to caselaw material.

As with the task of identifying cases involving agency statutory interpretation, these annotated data can be used to fine-tune and validate a large language model to determine agencies' litigation outcomes. Though also a classification task, this turned out to be more difficult than the task of identifying cases of agency statutory interpretation. There, it was possible to achieve strong performance with the raw opinion data, and by simply

8. This latter approach introduces potential and difficult to assess biases in the sample due to jurists' strategic case citations.

9. To take an example, it is not possible to search for citations to *Chevron* before 1984.

10. Eskridge & Baer, *supra* note 92.

11. The Krippendorff's alpha was generally around 0.7 every week, suggesting reasonably strong inter-coder agreement. However, this is a difficult task—cases often return mixed results for an agency, and deciding whether the case primarily favored the agency or not can be challenging.

using the first approximately 500 tokens of majority opinion text. That straightforward approach yielded weak performance in this task.[12] The difference in performance is likely because, unlike the initial agency interpretation task, the language relevant to the appropriate class is often buried deep in the opinion rather than in the first pages of the opinion. Another difficulty is that judges often discuss the pros and cons of various possible holdings, linguistic variation that may confound straightforward application of the models.

It was possible to achieve much stronger performance by adopting a two-step strategy: first, prompt a high-parameter generative language model to summarize the decision with respect to the issue of agency interpretation;[13] second, fine-tune a domain-adapted large language model classifier using the summaries rather than the raw opinion data. This two-step strategy thus takes advantage of the large context window of the high-parameter language models and their impressive ability to summarize documents,[14] but also relies on the tight task alignment possible through fine-tuning a lighter-weight model.[15] Using an 80-20 training-test split, a legal BERT model was trained for three epochs, at a learning rate of 4e-5, with a warmup ratio of 0.1. The five-fold average model performance statistics appear in table A2. Overall, the model performs remarkably well on this difficult task. The unweighted macro average F1-score is 0.84, with strong performance with respect to both classes. The accuracy over all test observations is 84 percent.

12. I also experimented with other strategies, such as simply asking a large parameter model to classify the decision; this too yielded poor performance.

13. Summarization was done by Gemini 2.0 Flash. The prompt was "Review the text of a court decision below. The decision involves a federal agency's interpretation of a statute. Summarize whether the court's decision supports or rejects the federal agency's interpretation of the statute. Be sure to focus on the court's support or rejection of the *agency's* interpretation.\n\n" f"TEXT OF COURT OPINION: {text}".

14. The BERT-based classifiers used elsewhere in this study have a relatively small context window of 512 tokens.

15. The reason that more sophisticated models performed poorly on the task is likely due to poor task alignment.

TABLE A2. MODEL PERFORMANCE: AGENCY LITIGATION OUTCOME

Class	Precision	Recall	F1-score	Support
Lose	0.8	0.81	0.8	53
Win	0.87	0.86	0.87	79
Macro-average (unweighted)			0.84	

Randomly generated predictions can help to assess model performance. If we randomly assign predictions to the test data (weighted only by class balance in the training data), that is, and evaluate this "model" performance, it provides a sense of how effective the trained model is.[16] Using this random baseline, the accuracy is fifty-three percent, and the macro average F1 score is 0.49, with an F1-score of 0.37 for the losing class and 0.62 for the winning class. The applied model markedly improves on this performance baseline.

B. Selection-Bias Bounds Adjustments

Though selection is not as severe an issue in public law as in private law, it remains a concern. It is not possible to know what the agency win-rates would be if all actions were litigated—the missing decisions might be for or against the agency, or anywhere in between. Yet even if it is not possible to determine the "true" win-rate, it is still possible to estimate informative bounds around the true win-rate. For the "worst" case bound, the idea is to assume that all missing decisions went against the agency; and for the "best" case bound, assume that all missing decisions went for the agency.[17] This approach is not typically feasible in private law because the number of missing judicial decisions vastly outnumber the observed judicial decisions, and the bounding assumptions would dominate any estimate. In public law, however, selection is less of an issue, and it is more reasonable to assume that the missing cases constitute a fraction of the observed cases. This

16. Predictions derive from the DummyClassifier in scikit-learn under the stratified strategy.

17. These best- and worst-case assumptions reflect the view that the likelihood of selection moves monotonically with potential outcomes—e.g., that the weakest (or strongest) cases select out of litigation. This assumption makes some sense theoretically, as challengers are less likely to invest in litigation that is unlikely to result in a victory. For similar assumptions in the literature, see Charles F. Manski & John V. Pepper, *Monotone Instrumental Variables: With an Application to the Returns to Schooling* (Nat'l Bureau of Econ. Rsch., Working Paper No. 0224, 1998); David S. Lee, *Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects* (Nat'l Bureau of Econ. Rsch., Working Paper No. 11721, 2005). A more extreme bounding exercise that assumes monotonicity flips between pre- and post-transition periods is also examined below.

permits an informative bounding exercise, which borrows from Manski's non-parametric bounding approach.[18]

Start with an estimate of the change in win-rates based on observed litigation: $\Delta = \overline{W}_{post} - \overline{W}_{pre}$. This difference comes close to many of the reported regression estimates. For instance, consider the regression run with a five-year bandwidth on years, judge and agency fixed effects, at the Courts of Appeals level. With that specification, the β for the judicial revolution transition is -.06, and the β for the *Chevron* transition is 0.004. The corresponding estimates based on differences in observed win-rates in this window of data are -.04 and 0.009, respectively.

Aside from the best- and worst-case assumptions, the central additional assumption required is the fraction of cases not observed due to selection. Let P_{pre} be the missing rate in the pre-reform period and P_{post} be the missing rate post-reform. With that in hand, a worst-case bound of the estimated difference is,

$$\Delta_W = \Delta - \overline{W}_{post}P_{post} + \overline{W}_{pre}P_{pre}.$$

And a best-case bound on the estimated difference is,

$$\Delta_B = \Delta_{\mathrm{W}} + P_{post} - P_{pre}.$$

For the reasons articulated in the body of the Article, it is further reasonable to assume that $P_{post} > P_{pre}$, meaning that there is more missingness under a deferential regime. These bounding expressions can be used to calibrate understanding of doctrinal effects under different assumptions about missingness, as reported in the body of the Article.

A more aggressive bounding strategy is the following: For the lower bound, assume that the agency loses all post-transition unobserved cases and wins all pre-reform unobserved cases; for the upper bound, assume that the agency wins all post-transition unobserved cases and loses all pre-reform unobserved cases.[19] These assumptions are even more unrealistic than the assumption that the agencies win or lose all unobserved cases, but they provide extreme bounds on the possible doctrinal effects. Under these assumptions, the lower bound is given by,

$$\Delta_L = \Delta_{\mathrm{W}} - P_{pre}.$$

And the upper bound is given by,

$$\Delta_U = \Delta_{\mathrm{W}} + P_{post}.$$

18. Manski, *supra* note 143.
19. These positions reflect more context-free assumptions. *See id.*

Notice that the subscripts under these assumptions reflect lower and upper bounds (rather than best- and worst-case scenarios, as earlier). The corresponding figures, which follow the formatting of those presented in the Article body, appear below. The values of the bounds reflect the more aggressive assumptions.

FIGURE A1. UPPER AND LOWER BOUNDS (JUDICIAL REVOLUTION)

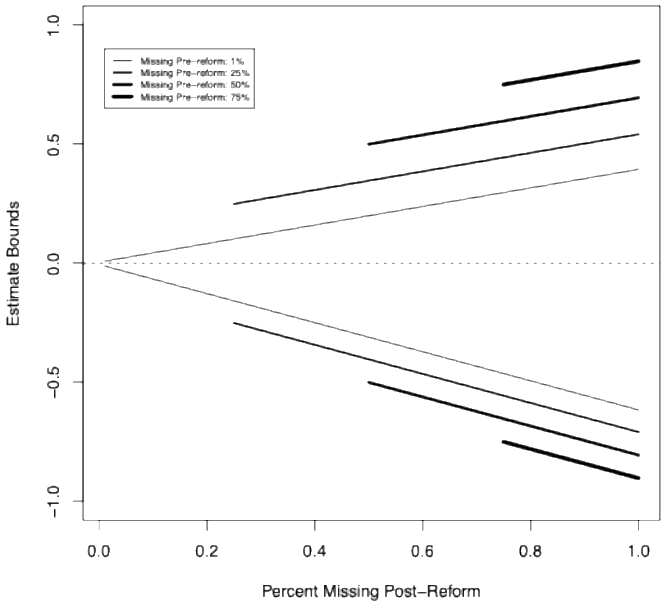


FIGURE A2. UPPER AND LOWER BOUNDS (*CHEVRON* REVOLUTION)