This is an early draft. Your comments are greatly appreciated.

AN EMPIRICAL STUDY OF STATUTORY INTERPRETATION IN TAX LAW

95 N.Y.U. L. REV. (forthcoming 2020)

Jonathan H. Choi[1]

A substantial academic literature considers how agencies *should* interpret statutes. But few studies have considered how agencies *actually do* interpret statutes, and none has empirically compared the methodologies of agencies and courts in practice. This Article conducts such a comparison, using a newly created dataset of all Internal Revenue Service ("IRS") publications ever released, along with an existing dataset of court decisions. It applies natural language processing, machine learning, and regression analysis to map methodological trends and to test whether particular authorities have developed unique cultures of statutory interpretation.

It finds that, over time, the IRS has increasingly made rules on normative policy grounds (like fairness and efficiency) rather than merely producing rules based on the "best reading" of the relevant statute (under any interpretive theory, like purposivism or textualism). Moreover, when the IRS does apply interpretive criteria, it has grown much more purposivist over time. In contrast, the Tax Court has not grown more normative and has followed the same trend toward textualism as most other courts. But although the Tax Court has become more broadly textualist, it prioritizes different interpretive tools than other courts, like *Chevron* deference and holistic-textual canons of interpretation. This suggests that each authority adopts its own flavor of textualism or purposivism.

These findings complicate the literature on tax exceptionalism and the judicial nature of the Tax Court. They also inform ongoing debates about judicial deference and the future of doctrines like *Chevron* and *Skidmore*. Most broadly, they provide an empirical counterpoint to the existing theoretical literature on statutory interpretation by agencies.

[1] Fellow, Classical Liberal Institute, New York University School of Law. Thanks to Anne Alstott, Aaron Bruhl, Richard Epstein, Bill Eskridge, Ryan Fackler, Jacob Goldin, Hayes Holderness, Ariel Kleiman, David Louk, Joshua Macey, Florencia Marotta-Wurgler, Julian Nyarko, the participants in the Harvard Law School Caselaw Access Project Summit, the participants in the New York University School of Law Lawyering Scholarship Colloquium, and many others at the Yale Law School and the New York University School of Law.

Article Contents

INTRODUCTION

After decades of debate over statutory interpretation by courts, scholars have more recently turned to the interpretive practices of agencies. Many have argued that agencies have relatively greater expertise in assessing statutory purpose, concluding that they ought to be more purposivist[2] than courts.[3] More fundamentally, many have suggested that judicial deference regimes, like *Chevron*,[4] empower agencies to make rules based on normative

[2] Textualists generally emphasize the plain meaning of statutory text and eschew legislative history. Purposivists generally look to all available evidence, including legislative history. The methodological distance between purposivists and textualists is often overstated, since all sides generally attempt to reconstruct statutory purpose and merely differ in the tools that they use to do so. For instance, although textualists are often presented as the foil to purposivists, modern textualists will also generally consider nontextual indicia of statutory purpose when statutory text is unclear. *See, e.g.*, John F. Manning, *What Divides Textualists from Purposivists?*, 106 COLUM. L. REV. 70, 84–85 (2006) ("[T]extualists generally forgo reliance on legislative history as an authoritative source of [the statute's apparent overall] purpose, but that reaction goes to the reliability and legitimacy of a certain type of evidence of purpose rather than to the use of purpose as such. . . . [W]hen semantic ambiguity creates the necessary leeway, textualists will try to construct a plausible hypothetical purpose [if possible].").

[3] See, e.g., William N. Eskridge, Jr., *Expanding Chevron's Domain: A Comparative Institutional Analysis of the Relative Competence of Courts and Agencies to Interpret Statutes*, 2013 WIS. L. REV. 411, 420–27, 434 ("[A]gencies interpret statutes purposively, and that is on the whole a good impulse in the modern regulatory state. A consequence of a purposivist approach to statutes is that the interpreter will read the statute dynamically, to reach beyond the original problems that were the basis of congressional deliberation."); Michael Herz, *Purposivism and Institutional Competence in Statutory Interpretation*, 2009 MICH. ST. L. REV. 89, 93 ("In general, my conclusion is that agencies make more respectable and less problematic purposivists than do judges."); Jerry L. Mashaw, *Norms, Practices, and the Paradox of Deference: A Preliminary Inquiry into Agency Statutory Interpretation*, 57 ADMIN. L. REV. 501, 511 (2005) ("In some instances, only the skillful deployment of legislative history will permit agencies to fulfill their constitutional role as faithful agents in the statute's implementation."); Cass R. Sunstein & Adrian Vermeule, *Interpretation and Institutions*, 101 MICH. L. REV. 885, 928 (2003) ("[A]gencies are likely to be in a better position to decide whether departures from the text actually make sense.").

But see Richard J. Pierce, *How Agencies Should Give Meaning to the Statutes They Administer: A Response to Mashaw and Strauss*, 59 ADMIN. L. REV. 197, 202 (2007) ("[T]he agency should use the same 'traditional tools of statutory construction' that it expects a reviewing court to use. If the agency uses a different method of interpretation—for example, if it relies on legislative history to a greater extent than a reviewing court as Strauss urges—it increases significantly the risk of judicial reversal without good reason."). Pierce's suggestion that agencies should follow the interpretive practices of courts only applies to interpretation carried out in *Chevron* step one. With respect to *Chevron* step two, Pierce believes (as do many others) that agencies ought to select the best policy rather than relying on any conventional interpretive norms. *See infra* note 5 and accompanying text.

[4] Chevron U.S.A., Inc. v. Nat. Res. Def. Council, Inc., 467 U.S. 837 (1984).

policy concerns, rather than merely seeking the "best reading" of a statute (using purposivism, textualism, or any other methodology).[5]

But despite a large theoretical literature on how agencies ought to interpret statutes, little scholarship has considered how they *actually do* interpret statutes.[6] Past work has focused on agency practice within a relatively narrow period,[7] making it impossible to evaluate how agency practice differed over time (especially before and after *Chevron*). Moreover, no empirical work has compared how agencies and courts differ while interpreting the same statutes.

[5] *See infra* Section I.A; *see also Chevron*, 467 U.S. at 842, 843 & n.9 (holding that an agency's interpretation of an ambiguous statute warrants deference so long as it represents a "reasonable policy choice"); *cf.* Covad Commc'ns v. FCC, 450 F.3d 528, 537 (D.C. Cir. 2006) (requiring the agency to "articulate a satisfactory explanation for its action including a rational connection between the facts found and the choice made"); Pierce, *supra* note 3, at 200 (arguing that, under *Chevron*, agencies can choose among permissible interpretations of a statute "only by engaging in a policymaking process"). *But see* Aaron Saiger, *Agencies' Obligation to Interpret the Statute*, 69 VAND. L. REV. 1231, 1231 (2016) ("An agency that commands deference bears a duty to adopt what it believes to be the best interpretation of the relevant statute.").

One could theoretically defend normative rulemaking by arguing that a reasonable legislator would have preferred the normatively best policy to prevail, and that therefore the best means for the agency to act as the "faithful agent" of the legislator is to prioritize policy concerns. This might be considered a particularly expansive form of purposivism, reminiscent of T. Alexander Aleinikoff's "nautical" approach, which "understands a statute as an on-going process (a voyage) in which both the shipbuilder and subsequent navigators play a role." T. Alexander Aleinikoff, *Updating Statutory Interpretation*, 87 MICH. L. REV. 20, 21 (1988). That said, scholars have generally accepted the distinction between the pursuit of the "best reading" of a statute and the "best policy." *See infra* Section I.A.

[6] *See* Christopher J. Walker, *Inside Agency Statutory Interpretation*, 67 STAN. L. REV. 999 (2015) (surveying attitudes toward statutory interpretation among agency administrators); Amy Semet, *An Empirical Examination of Agency Statutory Interpretation*, 103 MINN. L. REV. 2255 (2019) (considering statutory interpretation in decisions by the National Labor Relations Board). However, Walker's survey did not consider any actual decisions by agencies, and Semet's empirical work may be specific to the NLRB, due to its unusually intense partisanship. *See* Semet, *supra*, at 2280 ("Board voting is highly ideological . . . . Often, the Board reverses many of the decisions of the prior administration when a new partisan majority takes gains control of the Board."); Ronald Turner, *Ideological Voting on the National Labor Relations Board*, 8 U. PA. LAB. & EMP. L. 707, 712 (2006) (arguing the same point). More broadly, without considering comparable judicial practice, it is difficult to say how much her results were driven by the NLRB's status as an agency, and how much they were driven by issues unique to labor relations law.

[7] Walker's article relies on a single survey conducted in 2013. Walker, *supra* note 6, at 1015. Semet's article considers NLRB decisions between 1993 and 2017. Semet, *supra* note 6, at 2282. *Chevron* was decided in 1984. *See* Chevron U.S.A., Inc. v. Nat. Res. Def. Council, Inc., 467 U.S. 837 (1984).

This Article contributes to this conversation by studying a fertile area for agency-court comparisons: federal tax law. Because the IRS is one of the largest government agencies[8] and because its Internal Revenue Bulletin has been published so consistently (weekly) for so long (since 1919), it provides ample material for a longitudinal study of interpretive methodology over time. Similarly, the Tax Court handles the vast majority of federal tax cases (roughly 97%)[9] and has operated since 1942,[10] again producing a large amount of source material.

It was previously difficult or impossible to analyze such large bodies of documentation, not least because they were not readily accessible by researchers. This Article solves this problem by creating a new dataset of all Internal Revenue Bulletins ever published, which it analyzes along with a dataset launched by Harvard Law School's Caselaw Access Project less than a year ago.[11] Between these two sources, this Article analyzes 182,535 pages of Internal Revenue Bulletins and 470,099 court opinions.[12]

Broadly, this Article asks four main questions. First, how have interpretive methods evolved at the Tax Court and the IRS *within* each institution? Second, what is the difference *between institutions*—do agencies interpret statutes differently from courts? Third, what is the difference *between subject areas*—does the Tax Court interpret statutes differently from other federal courts (both Article I and Article III courts)? Fourth, what are the implications of interpreters' choices *between methods*—do they vary by party, or are particular methods associated with particular outcomes (either pro- or anti-taxpayer)?

To answer these questions, this Article uses "natural language processing" (algorithmic analysis of large bodies of text[13]) to assess how the IRS, the Tax Court, and other courts have used different tools in their

[8] The IRS had 74,454 employees as of fiscal year 2019, forming the vast majority of the Treasury Department's staff. DEP'T OF THE TREASURY, CONGRESSIONAL BUDGET JUSTIFICATION AND ANNUAL PERFORMANCE REPORT AND PLAN 1 (2019).

[9] Elizabeth Chao & Andrew R. Roberson, *Overview of Tax Litigation Forums*, TAX CONTROVERSY 360 (Apr. 21, 2017), https://www.taxcontroversy360.com/2017/04/overview-of-tax-litigation-forums.

[10] The U.S. Board of Tax Appeals, the predecessor to the Tax Court, was founded by the Revenue Act of 1924. Revenue Act of 1924, Pub. L. No. 68-176, § 900, 43 Stat. 253, 336 (1924). The Board of Tax Appeals was restructured and renamed the U.S. Tax Court by the Revenue Act of 1942. Revenue Act of 1942, Pub. L. No. 753, § 504(a), 56 Stat. 798, 957 (1942).

[11] *See infra* Appendix Section A.

[12] *See infra* Appendix Section A.

[13] *See* CHRISTOPHER D. MANNING & HINRICH SCHÜTZE, FOUNDATIONS OF STATISTICAL NATURAL LANGUAGE PROCESSING xxxi (1999).

decisions over time: interpretive versus normative,[14] and textualist versus purposivist. It measures the frequency with which authorities cite these tools—for example, textualists citing dictionaries, or purposivists citing legislative history—to map methodological trends. It then uses machine learning for more granular analysis,[15] by training algorithms to distinguish between court opinions based on interpretive methodology alone. This allows the algorithm to identify which specific terms, if any, are most strongly associated with the Tax Court and with the District Courts, providing a more nuanced account of the kind of purposivism or textualism each court applies. Finally, the Article uses regression analysis to test whether methodology can be predicted based on certain case characteristics, such as the party of the trial judge or the case outcome.

The main results are as follows. First, over time, the IRS has increasingly made rules based on normative rather than interpretive principles.[16] In contrast, the Tax Court has used roughly the same proportion of normative and interpretive tools since it was founded in 1942.[17]

Second, the IRS became much more purposivist and less textualist from the 1920s to approximately 1950, but has retained the same relatively purposivist posture since then.[18] On the other hand, the Tax Court has followed the general judicial movement of the past four decades away from purposivism and toward textualism.[19] The combination of these first two results suggests greater methodological cohesion among courts than among tax specialists.

[14] This Article describes decisionmaking as "interpretive" when it reflects a decisionmaker's attempt to act as a "faithful agent" of the legislature, archaeologically discerning a statute's true meaning while abstaining from value judgments. An interpretive approach, under this definition, may follow any interpretive method, including textualism, purposivism, or pragmatism. In contrast, a "normative" approach reflects a decisionmaker's attempt to create rules de novo based on its own policy preferences. There is a broader sense in which any decision by a court or agency could be described as "interpretive" if it concerns a statute; this Article does not use the term in that sense. The interpretive and normative perspectives will often overlap and often be considered simultaneously, especially since the normative desirability of a particular interpretation might be considered a factor in favor of its interpretive validity. *See also infra* note 5 (observing that a nonstandard view of interpretation might hold that the interpretive and normative viewpoints are identical).

[15] Machine learning uses computer algorithms in order to accomplish a particular task without human instructions. This Article primarily uses machine learning based on statistical inference. *See infra* Section II.B.

[16] *See infra* Section III.A.

[17] *See infra* Section III.B.

[18] *See infra* Section III.C.

[19] *See infra* Section III.D.

Third, the machine learning results reveal that Tax Court opinions can be distinguished from those of District Courts and the Court of Federal Claims based on the specific interpretive tools each employs. As compared to District Courts, the Tax Court favors congressional reports (especially reports from the Congressional Budget Office and the Joint Committee on Taxation) over hearings, holistic-textual canons (those emphasizing a cohesive reading of the tax code) over language canons, and *Chevron* deference over constitutional canons.[20] This complicates the conventional story that all courts have become more textualist—while they have in broad terms, the precise flavor of each court's interpretation differs in the details.

Fourth, regression analysis indicates that Tax Court judges appointed by Democratic presidents are more likely to use purposivist terms and less likely to use textualist terms than Republican appointees.[21] However, substantive outcomes (whether the court rules for or against the taxpayer) do not have a statistically significant relationship with interpretive methodology.[22]

Apart from theoretical interest, the findings in this Article have important practical implications. By underscoring agencies' shift toward normative decisionmaking, this Article is consistent with the widespread belief that *Chevron* permits agencies to make their own policy judgments rather than merely rediscovering Congress's. Some scholars view this as a feature of judicial deference, and some view it as a bug. Either way, this finding informs the positions taken by *Chevron*'s critics and its supporters.

The findings also suggest that tax exceptionalism—the widespread belief that tax statutes are or ought to be interpreted differently[23]—may be overstated in some respects and understated in others. Overstated, in that the Tax Court methodologically hews closer to other courts than to the IRS, despite their shared subject matter. So the conventional story that tax experts are exceptional because they are more purposivist may be incorrect. Understated, at the same time, in that the Tax Court does differ in its particular selection of textualist tools, suggesting that a more nuanced form of exceptionalism may apply.

Finally, the findings support controlling but controversial case law indicating that the Tax Court plays an "exclusively judicial role."[24] The

[20] *See infra* Section III.E.

[21] *See infra* Section III.F.

[22] *See infra* Section III.G.

[23] *See infra* notes 62–65 and accompanying text.

[24] Freytag v. Comm'r, 501 U.S. 868, 892 (1991). *But see* Kuretski v. Comm'r, 755 F.3d 929, 932 (D.C. Cir. 2014) (appearing to reach the opposite conclusion); Brant J. Hellwig, *The Constitutional Nature of the United States Tax Court*, 35 VA. TAX REV. 269, 326 (2016) ("The exercise of attempting to definitively locate the United States Tax Court

conclusion in this Article that the Tax Court interprets statutes more like other courts than like the IRS undermines the claims of some scholars that Tax Court opinions should be subject to judicial deference, much like agency pronouncements[25]—instead, at least on the key dimension of interpretive methodology, the Tax Court behaves like other courts, suggesting that de novo review may be appropriate.[26]

Part I discusses the key questions that this Article seeks to answer. Part II describes data and empirical methods. Part III presents results and explanations for those results. Part IV conducts robustness checks to provide assurance that these results are correct. The Conclusion considers possible implications of the results. The Appendix provides additional detail on methods and data.

I. KEY QUESTIONS

A. Interpretive Judgments or Normative Policymaking?

Chevron famously held that an agency's interpretation of an ambiguous statute warrants deference so long as it reflects a "reasonable policy choice."[27] Many have concluded from this that agencies should make rules based on normative considerations, rather than merely aiming at the

in a particular branch of government proves difficult at best, and at times feels like a hopeless exercise.").

[25] Some scholars have argued that Tax Court opinions ought to be entitled to *Chevron* deference. *See* David F. Shores, *Deferential Review of Tax Court Decisions:* Dobson *Revisited*, 49 TAX L. REV. 629 (1996); David F. Shores, *Rethinking Deferential Review of Tax Court Decisions*, 53 TAX L. REV. 35 (1999); Andre L. Smith, *Deferential Review of the United States Tax Court: The* Chevron *Doctrine*, 37 VA. TAX REV. 75 (2017). Others have disagreed. *See* Steve R. Johnson, *The Phoenix and the Perils of the Second Best: Why Heightened Appellate Deference to Tax Court Decisions Is Undesirable*, 77 OR. L. REV. 235 (1998); Leandra Lederman, *(Un)Appealing Deference to the Tax Court*, 63 DUKE L.J. 1835, 1835 (2014) ("Contrary to some scholarship, this Article argues that, as a doctrinal matter, no vestige of the *Dobson* rule remains and that courts of appeals must apply the same standard of judicial review that they apply to district courts in nonjury cases."). As a practical matter, decisions of the Tax Court do not currently receive *Chevron* deference.

[26] This issue has a chicken-and-egg quality, in that the Tax Court likely uses textualist methodology at least in part to follow reviewing courts, since the Tax Court would risk reversal if it remained purposivist like the IRS. In contrast, if Tax Court decisions were to receive deference, the Tax Court would have more freedom to use purposivist methodology with less risk of reversal. So the Tax Court may presently behave like a court because it is treated like a court, without judicial deference. *See infra* notes 61–62 and accompanying text.

[27] Chevron U.S.A., Inc. v. Nat. Res. Def. Council, Inc., 467 U.S. 837, 842, 843 & n.9 (1984).

"best reading" of a statute. E. Donald Elliott recounts from his tenure at the Environmental Protection Agency's ("EPA") Office of General Counsel that, before *Chevron*, the EPA had treated each statute as a "prescriptive text having a single meaning, discoverable by specialized legal training and tools."[28] After *Chevron*, it treated statutes as creating "a range of permissible interpretive discretion," within which "[t]he agency's policy-makers, not its lawyers, should decide which of several different but legally defensible interpretations to adopt."[29]

Peter Strauss put forward an influential version of this view with his idea of "*Chevron* space." He argues that *Chevron* creates a zone of agency discretion for readings of the statute that are "permissible" but not "necessary" under ordinary rules of statutory interpretation. When confronted with several such plausible alternative readings, an agency may select among them, whether for normative policy reasons or interpretive reasons, without judicial interference.[30] A number of other scholars have created models following this approach, emphasizing the tradeoff between courts' interpretive goals and agencies' normative ones.[31]

Some have pushed back. In particular, Aaron Saiger has argued that agencies "must reject interpretations that it concludes are interpretively suboptimal, notwithstanding that an ethical, law-abiding reviewing court would acquiesce in those interpretations."[32] In his view, judicial deference to agencies requires those agencies to take on the mantle of the court, which has a duty to "reach the best account of what a statute means."[33]

This Article takes no position on whether a normative shift would be appropriate or not. It only remarks that a shift toward normative

[28] E. Donald Elliott, *Chevron Matters: How the Chevron Doctrine Redefined the Roles of Congress, Courts and Agencies in Environmental Law*, 16 VILL. ENVTL. L.J. 1, 11 (2005).

[29] *Id.* at 12; *see also* Mashaw, *supra* note 3, at 532–33 & nn.71, 73.

[30] Peter L. Strauss, *"Deference" Is Too Confusing—Let's Call Them "*Chevron *Space" and "*Skidmore *Weight,"* 112 COLUM. L. REV. 1143, 1163–64 (2013).

[31] *See* Yehonatan Givati, *Strategic Statutory Interpretation by Administrative Agencies*, 12 AM. L. & ECON. REV. 95, 96 (2010) ("In the model, the agency, which maximizes some objective function, adopts a rule that interprets a statute . . . ."); Matthew C. Stephenson, *The Strategic Substitution Effect: Textual Plausibility, Procedural Formality, and Judicial Review of Agency Statutory Interpretations*, 120 HARV. L. REV. 528, 535, 536, 544 (2006) (assuming that agencies are "interpretive instrumentalists, attaching no intrinsic importance to textual fidelity or analogous concerns" but instead attempting to "secure whatever interpretation would best advance its substantive policy agenda"); John R. Wright, *Ambiguous Statutes and Judicial Deference to Federal Agencies*, 22 J. THEORETICAL POL. 217, 226 (2010) (also modelling agency action as a function of policy goals).

[32] Saiger, *supra* note 5, at 1233.

[33] *Id.* at 1234.

decisionmaking has been posited much more often than it has been demonstrated. The widespread belief in this normative shift has been supported primarily by anecdote,[34] which is troubling given that it is the main basis for the critique of *Chevron* leveled by current Justices of the Supreme Court.[35]

Of course, *Chevron* deference only applies to one aspect of agency activity: traditional regulatory rulemaking.[36] Subregulatory guidance (including, for the IRS, revenue rulings and revenue procedures) is instead subject to *Skidmore*[37] deference, under which "courts are obliged to take an

[34] *See* Brett M. Kavanaugh, Book Review, *Fixing Statutory Interpretation*, 129 HARV. L. REV. 2118, 2150 (2016) ("From my more than five years of experience at the White House, I can confidently say that *Chevron* encourages the Executive Branch (whichever party controls it) to be extremely aggressive in seeking to squeeze its policy goals into ill-fitting statutory authorizations and restraints."); David S. Tatel, *The Administrative Process and the Rule of Environmental Law*, 34 HARV. ENVTL. L. REV. 1, 2 (2010) ("[I]t looks for all the world like agencies choose their policy first and then later seek to defend its legality."); *supra* notes 28–29 and accompanying text.

[35] *See, e.g.*, Michigan v. EPA, 135 S. Ct. 2699, 2713 (2015) (Thomas, J., concurring) (complaining that *Chevron* empowers agencies "not to find the best meaning of the text, but to formulate legally binding rules to fill in gaps based on policy judgments made by the agency rather than Congress"); Kavanaugh, *supra* note 34, at 2151 ("*Chevron* invites an extremely aggressive executive branch philosophy of pushing the legal envelope . . . . After all, an executive branch decisionmaker might theorize, 'If we can just convince a court that the statutory provision is ambiguous, then our interpretation of the statute should pass muster as reasonable. And we can achieve an important policy goal if our interpretation of the statute is accepted.'").

[36] This is a relatively recent development with respect to the IRS—prior to the Supreme Court's 2011 ruling in *Mayo Foundation for Medical Education and Research v. United States*, it was unclear whether all IRS regulations were subject to *Chevron* deference or whether some might be subject to (weaker) *Skidmore* deference instead. *See* Mayo Found. for Med. Educ. & Research v. United States, 562 U.S. 44, 53 (2011) ("We see no reason why our review of tax regulations should not be guided by agency expertise pursuant to *Chevron* to the same extent as our review of other regulations."); MICHAEL SALTZMAN & LESLIE BOOK, IRS PRACTICE AND PROCEDURE ¶ 3.02[4] (2019) (describing the rise and fall of tax exceptionalism in judicial deference to IRS regulations); Michael Hall, *From* Muffler *to* Mayo*: The Supreme Court's Decision to Apply* Chevron *to Treasury Regulations and Its Impact on Taxpayers*, 65 TAX LAW. 695 (2012) (same). If the IRS expected weak *Skidmore* deference rather than stronger *Chevron* deference for some of its regulations prior to *Mayo*, then we might expect the shift toward normative decisionmaking discussed in Section III.A to be even more pronounced at other agencies, where *Chevron* always applied across the board.

[37] Skidmore v. Swift & Co., 323 U.S. 134, 140 (1944) (holding that subregulatory guidance, "while not controlling upon the courts by reason of their authority, do constitute a body of experience and informed judgment to which courts and litigants may properly resort for guidance"). Despite the terminology, the concept that agency statutory interpretation might be "entitled to very great respect" precedes *Skidmore*. Edwards' Lessee v. Darby, 25 U.S. (12 Wheat.) 206, 210 (1827) ("In the construction of a doubtful and ambiguous law, the

agency's view about statutory meaning into account when interpreting statutes the agency administers."[38] Scholars have been less opinionated on the implications of *Skidmore* deference for statutory interpretation. Peter Strauss has suggested that it should be rebranded "*Skidmore* weight," since it is not deference so much as a factor that courts are obliged to consider in their decisions.[39] Connor Raso and William Eskridge describe it as just "mildly deferential, . . . a judicial willingness to go along."[40] Saiger considers this question in the alternative: If *Skidmore* requires courts to give deference, then agencies have a duty (which they may or may not fulfill in practice) to produce subregulatory guidance that is interpretive rather than normative.[41] And, in Saiger's view, even if *Skidmore* does not demand deference, agencies would still be "wise" to emphasize interpretation in order to avoid reversal by courts.[42]

Here, then, is the overall picture. Trial courts always read statutes with an eye to interpretation, not least because they know that reviewing courts will do so. Agencies are generally thought to have greater flexibility to issue regulations and other guidance based on normative criteria than courts, although there is debate over the legitimacy of this approach and unclarity about deference to subregulatory guidance. And, if this theoretical account is descriptively correct, we would expect to see a shift toward normative decisionmaking after 1984 (*Chevron*), and perhaps after 1944 (*Skidmore*) as well.

cotemporaneous construction of those who were called upon to act under the law and were appointed to carry its provisions into effect is entitled to very great respect."); *see also, e.g.*, Fawcus Mach. Co. v. United States, 282 U.S. 375, 378 (1931); Swendig v. Wash. Water Power Co., 265 U.S. 322, 331 (1924);.

[38] Strauss, *supra* note 30, at 1153; *see also* SALTZMAN & BOOK, *supra* note 36, ¶ 3.03[1][b] ("Prior to the Supreme Court's decision in *Mead*, some courts applied *Chevron* deference to revenue rulings while others gave no deference whatsoever. After *Mead*, the general consensus is that *Skidmore* is the more appropriate standard by which to evaluate revenue rulings, not *Chevron*. The Supreme Court itself, however, has not expressly ruled on the question in a post-*Mead* world.").

[39] Strauss, *supra* note 30, at 1146.

[40] Connor N. Raso & William N. Eskridge, Chevron *as a Canon, Not a Precedent: An Empirical Study of What Motivates Justices in Agency Deference Cases*, 110 COLUM. L. REV. 1727, 1737, 1744 (2010).

[41] Saiger, *supra* note 5, at 1281.

[42] Saiger, *supra* note 5, at 1281–83 ("If courts defer under *Skidmore* to agency interpretations they think are interpretively suboptimal, then agencies are saying what the law is and must promulgate the interpretation they think is interpretively the best. If courts will not accept interpretations with which they do not agree, agencies are both entitled and usually wise to privilege the courts' anticipated interpretation over their own best interpretation of the statute.").

B. Textualism or Purposivism?

Once a particular authority has decided to engage in statutory interpretation, the next question will be what *kind* of interpretation it should conduct. Here, the key questions have been whether particular interpreters are more textualist or purposivist and how their practices have changed over time.[43]

To place the relationship between textualism and purposivism in context, consider the best-known trend in statutory interpretation: the rise and fall of purposivism at the Supreme Court. The standard story is that modern purposivism took root around 1940, tracking President Franklin Roosevelt's appointment of purposivist Justices and the development of new judicial methodologies to complement the expanded administrative state.[44] Purposivism continued its ascent into the 1970s, which have been described as the "heyday of purposive analysis."[45] But after peaking in the 1970s, the use of purposivism by the Supreme Court sharply declined, thanks to the appointment of textualist Justices by Republican Presidents (especially Justice Scalia in 1986).[46]

[43] See, e.g., Aaron-Andrew P. Bruhl, *Statutory Interpretation and the Rest of the Iceberg: Divergences Between the Lower Federal Courts and the Supreme Court*, 68 DUKE L.J. 1 (2018) (evaluating textualism and purposivism in the Supreme Court, Circuit Courts, and District Courts); Corey Ditslear & James J. Brudney, *The Warp and Woof of Statutory Interpretation: Comparing Supreme Court Approaches in Tax Law and Workplace Law*, 58 DUKE L.J. 1231 (2009) (evaluating textualism and purposivism at the Supreme Court); Anita S. Krishnakumar, *Statutory Interpretation in the Roberts Court's First Era: An Empirical and Doctrinal Analysis*, 62 HASTINGS L.J. 221 (2010) (evaluating textualism and purposivism in the Roberts Court).

[44] Nicholas R. Parrillo, *Leviathan and Interpretive Revolution: The Administrative State, the Judiciary, and the Rise of Legislative History, 1890-1950*, 123 YALE L.J. 266, 266 (2013) ("[T]his Article reveals that judicial use of legislative history became routine quite suddenly, in about 1940. The key player in pushing legislative history on the judiciary was the newly expanded New Deal administrative state."); *see also, e.g.*, JOHN W. JOHNSON, THE DIMENSIONS OF NON-LEGAL EVIDENCE IN THE AMERICAN JUDICIAL PROCESS: THE SUPREME COURT'S USE OF EXTRA-LEGAL MATERIALS IN THE TWENTIETH CENTURY (1990); Jorge L. Carro & Andrew R. Brann, *Use of Legislative Histories by the United States Supreme Court: A Statistical Analysis*, 9 J. LEGIS. 282 (1982); Nancy Staudt et al., *Judging Statutes: Interpretive Regimes*, 38 LOY. L.A. L. REV. 1909 (2005); Nicholas S. Zeppos, *The Use of Authority in Statutory Interpretation: An Empirical Analysis*, 70 TEX. L. REV. 1073 (1992).

[45] Anita S. Krishnakumar, *Backdoor Purposivism*, 69 DUKE L.J. (forthcoming 2020) (manuscript at 2).

[46] *See, e.g.*, John Calhoun, *Measuring the Fortress: Explaining Trends in Supreme Court and Circuit Court Dictionary Use*, 124 YALE L.J. 484, 498 (2014) ("[T]he sharpest increase in the use of dictionaries began in the mid-1980s, around the time Justice Scalia arrived at the Court."); Paul Clement, Editorial, *Arguing Before Justice Scalia*, N.Y. TIMES (Feb. 17, 2016), (describing 1987 as "when Justice Scalia started writing opinions for the

Figure 1 illustrates the conventional story, using the same methodology that this Article applies to the IRS and Tax Court further below.[47] Each point in the Figure represents the average term frequency of purposivist terms or textualist terms among all Supreme Court cases for the relevant year,[48] normalized to avoid inappropriately emphasizing the absolute magnitudes of term frequencies.[49] Because term frequency is inevitably based on the subjective choice of particular terms, as explained in greater detail below,[50] the absolute magnitudes of term frequencies are less important than relative magnitudes over time.

For ease of reading, the points are used to generate a trend line using locally estimated scatterplot smoothing (LOESS), a non-parametric form of local regression that fits a smooth curve to data points, with a 95% confidence interval represented by the shaded area.[51] These charts are presented as

court emphasizing the importance of statutory text and the unreliability of legislative history, and that made all the difference").

[47] *See infra* Section II.A (discussing empirical methods in greater detail).

[48] All the Figures in this Article were produced calculating the average of the term frequencies for all judicial opinions (or regulatory documents) for that year, weighted based on the word count of each document. For example, in calculating the textualist score for each year, a Tax Court opinion that is twice as long will count twice as much toward that score.

[49] *See infra* Section II.A (discussing the problems with comparisons of absolute term frequency magnitudes between interpreters).

[50] *See infra* Section II.A.

[51] *See* WILLIAM S. CLEVELAND, THE ELEMENTS OF GRAPHING DATA 168–73 (rev. ed. 1994) (describing LOESS); Aaron-Andrew P. Bruhl, *Statutory Interpretation and the Rest of the Iceberg: Divergences Between the Lower Federal Courts and the Supreme Court*, 68 DUKE L.J. 1, 189 (2018) (applying LOESS to a similar analysis of term usage, but using a smoothing factor of 0.33 rather than 0.5, resulting in a more tightly fitted curve). I use a smoothing factor of 0.5. *Smoothed Conditional Means*, GGPLOT2, https://ggplot2.tidyverse.org/reference/geom_smooth.html (last visited Sept. 15, 2019).

The confidence intervals in Figures 1 through 7, 11, and 20 through 26, are all calculated using bootstrapping. The bootstrapping process used is analogous to the ones described in Section B.IV and Section G of the Appendix. Given a sample of data points (in this case, with years and the term frequency of a particular methodology for that year), bootstrapping recreates a sample of the same size by randomly sampling (with replacement) from the original sample. This is repeated a number of times, here 1000 times, and LOESS curves are recalculated with respect to each bootstrapped sample. For each point on the graph's x-axis (here, each point in time), the values of each bootstrapped LOESS curve are stored and then used to calculate a confidence interval.

The confidence intervals follow the basic bootstrap (also known as the "reverse percentile," "pivotal," or "empirical" bootstrap) equation, such that at each point on the *x*-axis, where $\hat{\theta}$ is the LOESS value in the original sample, $\theta^*_{0.025}$ is the 2.5th-percentile bootstrapped value, and $\theta^*_{0.975}$ is the 97.5th-percentile bootstrapped value, the confidence interval equals:

$$(2\hat{\theta} - \theta^*_{0.975}\,,\ 2\hat{\theta} - \theta^*_{0.025})$$

exploratory data analysis rather than reflecting causal inferences, since the year an opinion was written is likely not the primary driver of interpretive methodology so much as it is correlated with deeper shifts in judicial philosophy.

Figure 1: Purposivist and Textualist Terms in Supreme Court Opinions

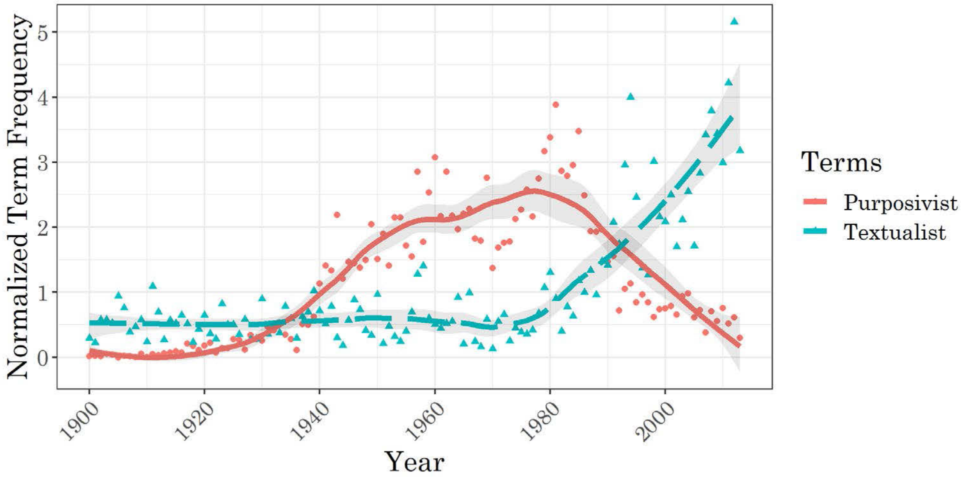
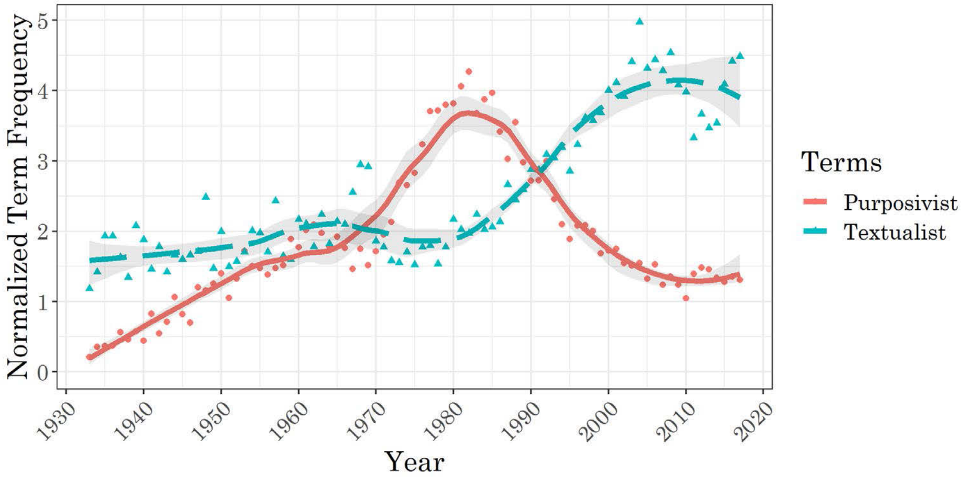


Figure 1 squares neatly with existing literature, showing the same rise in purposivism during the 1930s and 1940s, the peak in the 1970s, and the subsequent decline to the present, accompanied by a sharp uptick in textualism. The fact that Figure 1 is consistent with past scholarship is an early reassurance of the validity of the methods in this Article. Prior empirical research has also concluded that appellate and district courts have followed the same rough directional trend as the Supreme Court, albeit less dramatically and with a slight lag.[52] This methodology again generally confirms this result in Figure 2:

A.C. DAVISON & D.V. HINKLEY, BOOTSTRAP METHODS AND THEIR APPLICATION 194 (1997). Note that the confidence intervals are the confidence intervals of the *curve*, not confidence intervals of *observations*. That is, within each interval with respect to a given point on the *x*-axis, there is a 95-percent probability that the true regression line lies within that interval. But this does *not* imply that there is a 95-percent probability that any observation will lie within that interval. The latter probability would be captured by a prediction interval, which would take into account both uncertainty regarding the regression line as well as pointwise variance in the distribution of observations.

[52] *See* Bruhl, *supra* note 43, at 1 ("[A]ll federal courts have shifted toward more frequent use of textualist tools in recent decades. However, that shift has been less pronounced as one moves down the judicial hierarchy."); *see also* James J. Brudney & Lawrence Baum, *Two Roads Diverged: Statutory Interpretation by the Circuit Courts and*

Figure 2: Purposivist and Textualist Terms in District Court Opinions



The decline in purposivism and the ascent of textualism both begin slightly later at the District Courts. But the overall modern trend, away from purposivism and toward textualism, is clearly visible at both levels of court.

The crucial empirical question for this Article is whether agencies have followed the courts in their move toward textualism. Most scholars have argued that agencies should remain purposivist as a normative matter, although some have disagreed.[53] But whether they have actually done so is an open question and one that past studies have not attempted to answer.[54] This Article will address this question empirically, exploring more than a century of IRS guidance.

C. Cohesion Among Courts or Among Specialists?

The pattern of the purposivist/textualist shift at agencies and courts presents competing hypotheses with respect to the Tax Court. If the IRS and

the Supreme Court in the Same Cases, 87 FORDHAM L. REV. (forthcoming 2019) (generally confirming Bruhl's findings). *But see* Abbe R. Gluck & Richard A. Posner, *Statutory Interpretation on the Bench: A Survey of Forty-Two Judges on the Federal Courts of Appeals*, 131 HARV. L. REV. 1298, 1309–15 (2018) (arguing that judicial methodology in federal appellate courts is more complicated than the traditional textualist/purposivist divide, but acknowledging the general shift in recent decades toward textualist methods, even by those judges unwilling to self-identify as textualists).

[53] *See supra* note 3 and accompanying text.

[54] Semet, in particular, did not analyze trends over time, since her study was a snapshot of a fourteen-year period, too short to illustrate long-term methodological trends. Semet, *supra* note 6, at 2282.

generalist courts differ methodologically (and this Article concludes that they do), which will more strongly influence the Tax Court: cohesion with the IRS or cohesion with generalist courts?

The Tax Court handles almost all federal tax cases,[55] operating much like a centralized federal trial court. It takes cases after administrative adjudication by the IRS's internal Office of Appeals,[56] and, if cases are appealed from the Tax Court, they are reviewed de novo by the circuit court that had jurisdiction over the taxpayer.[57] Although the Tax Court is an Article I court, the Supreme Court ruled in *Freytag v. Commissioner*[58] that it "exercises judicial power to the exclusion of any other function . . . in much the same way as the federal district courts exercise theirs,"[59] concluding that the Tax Court's "exclusively judicial role distinguishes it from other non-Article III tribunals that perform multiple functions . . . ."[60]

Given the Tax Court's judicial role, there is reason to suspect that it would follow the general federal judicial trend toward textualism. More pragmatically, because Tax Court cases are reviewed de novo by circuit courts[61] and because the Tax Court "follows the law of the circuit in which a taxpayer's appeal would lie,"[62] the Tax Court has every incentive to conform its interpretive practice to that of the courts of appeals. If the Tax Court had remained purposivist, it might have found itself reversed with increasing frequency by textualist-leaning circuit courts.

On the other hand, scholars have long observed that tax law operates differently than other fields of law. In particular, "tax exceptionalists" have argued that federal tax statutes must be read in a more purposivist manner

[55] *See supra* note 9 and accompanying text.

[56] *See generally* 26 C.F.R. § 601.106 (2019) (describing the procedures for the Office of Appeals).

[57] I.R.C. § 7482 (2012); Smith, *supra* note 25, at 78 ("[D]e novo review represents the status quo . . . .").

[58] 501 U.S. 868 (1991).

[59] *Id.* at 891.

[60] *Id.* at 892 (ruling in the context of a dispute over the method for appointing special trial judges). This conclusion is, however, somewhat controversial. The D.C. Circuit's *Kuretski* ruling appears to cut the other way. Kuretski v. Comm'r, 755 F.3d 929, 932 (D.C. Cir. 2014); *see also* Hellwig, *supra* note 24, at 326 ("The exercise of attempting to definitively locate the United States Tax Court in a particular branch of government proves difficult at best, and at times feels like a hopeless exercise.").

[61] *See supra* note 57 and accompanying text.

[62] Amandeep S. Grewal, *The Un-Precedented Tax Court*, 101 IOWA L. REV. 2065, 2078 (2016). This is known as the "*Golsen* rule." *See* Golsen v. Comm'r, 54 T.C. 742, 757 (1970) ("[W]here the Court of Appeals to which appeal lies has already passed upon the issue before us, efficient and harmonious judicial administration calls for us to follow the decision of that court"), *aff'd on other grounds*, 445 F.2d 985 (10th Cir. 1971).

than other federal statutes, due to idiosyncrasies of the tax code or the tax legislative process.[63] Corey Ditslear and James Brudney have found that, as a descriptive matter, the Supreme Court has been more purposivist in its tax opinions than in other opinions, although they largely attribute this to the influence of Justice Blackmun.[64] And Steve Johnson has directly speculated that the Tax Court's subject matter expertise might free it to apply purposivist techniques, much like the IRS.[65] Since the Tax Court and the IRS are both staffed by tax experts, known for their cultural insularity, one might expect them to converge in their interpretive techniques.[66]

Moreover, despite the Supreme Court's view that the Tax Court is "exclusively judicial,"[67] the Tax Court's status as an Article I court carries some distinctions from district courts. Tax Court judges are specialists,[68] they

[63] *See, e.g.*, Bradford L. Ferguson, Frederic W. Hickman & Donald C. Lubick, *Reexamining the Nature and Role of Tax Legislative History in Light of the Changing Realities of the Process*, 67 TAXES 804, 806–07 (1989) (citing the Code's complexity, age, extensive legislative history, specialized nature, and specialized drafting process); Mary L. Heen, *Plain Meaning, the Tax Code, and Doctrinal Incoherence*, 48 HASTINGS L.J. 771, 786 & nn.73, 818–19 (1997) (arguing against textualism in tax law); Michael Livingston, *Congress, the Courts, and the Code: Legislative History and the Interpretation of Tax Statutes*, 69 TEX. L. REV. 819, 882 (1991) ("The Article argues that the unique characteristics of tax law render generalized theories of interpretation inadequate for tax cases. These characteristics include the complex and constantly changing character of the tax code; the contextual style of tax interpretation, which emphasizes Treasury regulations, previous judicial decisions, and the broader statutory structure rather than the literal or plain meaning of the provision being construed; and the conceptual nature of the tax legislative process, in which members of Congress set only general guidelines for both the statute and legislative history."); Clint Wallace, *Congressional Control of Tax Rulemaking*, 71 TAX L. REV. 179, 183 (2017) (arguing for a "JCT Canon" under which tax statutes would be interpreted with a special eye toward legislative history generated by the Joint Committee on Taxation).

Some scholars have resisted the notion of tax exceptionalism. *See* Paul L. Caron, *Tax Myopia, or Mamas Don't Let Your Babies Grow up to Be Tax Lawyers*, 13 VA. TAX REV. 517, 518 (1994) (accusing the tax bar and tax scholars of "tax myopia"); Michael Livingston, *Practical Reason, "Purposivism," and the Interpretation of Tax Statutes*, 51 TAX L. REV. 677 (1996) (criticizing "the myth of tax essentialism").

[64] Ditslear & Brudney, *supra* note 43, at 1270-75. Ditslear and Brudney note that "after Blackmun departed . . . the Court's willingness to invoke legislative history in its tax majorities significantly declined." *Id.* at 1274.

[65] Steve R. Johnson, *The Canon that Tax Penalties Should Be Strictly Construed*, 3 NEV. L.J. 495, 518 (2003) ("It may well be that the Tax Court, as a result of its greater expertise, feels greater confidence in applying the copious interpretive materials that, I have argued, should be the proper bases for construing tax penalty statutes.").

[66] Caron, *supra* note 63, at 530.

[67] Freytag v. Comm'r, 501 U.S. 868, 891 (1991).

[68] Lederman, *supra* note 25, at 1880 ("[T]he Tax Court is specialized—its judges only decide tax cases—and accordingly has greater expertise in tax matters than do other courts.").

are appointed for limited fifteen-year terms (although they are frequently reappointed),[69] and they can be removed (with cause) by the President, whereas Article III judges must be impeached.[70] Practically speaking, Circuit Courts might be more reluctant to overturn the judgments (including the purposivist judgments) of specialists than generalists. Consequently, the Tax Court might also differ from other courts for procedural, rather than substantive, reasons.

If the tax exceptionalists are right, or if Article I courts tend to be distinct, then the Tax Court should resist the trend toward textualism and remain purposivist, like the IRS. But if cohesion among courts is the stronger force, then we should expect the Tax Court to trend toward textualism, like other federal courts.

II. Empirical Methods

To answer these questions, I created a new dataset of all IRS publications ever released, dating back to 1919. This includes all regulatory rulemaking and published subregulatory guidance, but excludes unpublished, non-precedential guidance provided directly to specific taxpayers. The publications were converted to plain text using optical character recognition ("OCR"), and then cleaned both manually[71] and using computer code—for example, by spell-checking, regularizing whitespace, and removing sections of the publications irrelevant to this Article's analysis. In addition, I downloaded court data from Harvard Law School's Caselaw Access Project, a high-quality dataset that includes almost every court case ever decided in the United States until 2015. Section A of the Appendix provides additional detail on the data used in this Article.

A. Natural Language Processing

The primary measure of interpretive methodology in this Article is the frequency with which agencies and courts cite particular tools, such as legislative history, dictionaries, or canons of construction. This is the dominant approach in existing literature, and maps closely onto conventional

[69] *See infra* note 148.

[70] Smith, *supra* note 25, at 95–96.

[71] In particular, I read through the plain text of each Cumulative Internal Revenue Bulletin to ensure that my code had correctly removed legislative history that did not represent original IRS writing. *See infra* Appendix Section A.1.

conceptions of textualism and purposivism.[72] For example, if a particular document had sixteen phrases relating to legislative history, and the document had 8000 words, the term frequency score for the document with respect to legislative history would be:

$$\frac{16}{8000} = 0.002$$

A single document might have a positive term frequency score for both textualism and purposivism, or both interpretive and normative decisionmaking. Judicial decisions sometimes weigh both textualist and purposivist considerations in the alternative, so this is not uncommon.[73]

[72] *See, e.g.*, Bruhl, *supra* note 43, at 29 ("To a significant degree, the observable difference between competing interpretive approaches lies in which tools they prioritize and emphasize. A judge that uses linguistic canons and dictionaries extensively but uses legislative history sparingly is more textualist than a judge who displays the opposite tendencies."); Lawrence Solan, *Private Language, Public Laws: The Central Role of Legislative Intent in Statutory Interpretation*, 93 GEO. L.J. 453, 453–55 (2005) (citing judicial references to legislative intent as primary evidence of judicial intentionalism). *See generally* James J. Brudney & Lawrence Baum, *Dictionaries 2.0: Exploring the Gap Between the Supreme Court and Courts of Appeals*, 2015 YALE L.J. FORUM 104 (studying the frequency of dictionary citations by the Supreme Court and courts of appeals, using word searches); Calhoun, *supra* note 46 (studying the frequency of dictionary citations by the Supreme Court and courts of appeals, using word searches). For other applications of term frequency analysis not limited to statutory interpretation methodology, see, for example, Keith Carlson, Michael A. Livermore & Daniel Rockmore, *A Quantitative Analysis of Writing Style on the U.S. Supreme Court*, 93 WASH. U. L. REV. 1461, 1478–80 (2016) (using term frequency to evaluate judicial "friendliness"). *See generally* Daniel Martin Katz et al., *Legal N-Grams? A Simple Approach to Track the Evolution of Legal Language*, 235 FRONTIERS ARTIFICIAL INTELLIGENCE & APPLICATIONS 167 (2011) (using n-gram analysis to track the evolution of legal language); David E. Pozen, Eric L. Talley & Julian Nyarko, *A Computational Analysis of Constitutional Polarization*, 105 CORNELL L. REV. (forthcoming 2019) (using textual analysis to analyze constitutional polarization).

More generally, term frequency underlies the "bag of words" model that is one of the most common classification schemes used in natural language processing and machine learning. *See* MICHAEL MCTEAR, ZORAIDA CALLEJAS & DAVID GRIOL BARRES, THE CONVERSATIONAL INTERFACE, TALKING TO SMART DEVICES 167 (2016). This Article uses the bag-of-words model to implement machine learning, which is the standard approach. *See, e.g.*, Pozen et al., *supra*, at *16 (analyzing the frequency with which terms are used without taking into account the context in which they are used). Term frequency also underlies many measures of "similarity" between different documents. *See, e.g.*, Carlson, Livermore & Rockmore, *supra*, at 1483–86 (measuring divergence in judicial writing styles); Elliott Ash & Omri Marian, The Making of International Tax Law: Empirical Evidence from Natural Language Processing, at *16 (unpublished manuscript) (on file with author).

[73] *E.g.*, Whistleblower 21276-13W v. Comm'r, 147 T.C. 121, 128 & n.8 (2016); Gardner v. Comm'r, 145 T.C. 161, 164, 176, 179 (2015).

Different scholars have different specific definitions of textualism and purposivism, and this Article does not argue that purposivism is merely the act of using legislative history to interpret statutes, which would be an oversimplification. Nevertheless, textualists' skepticism toward legislative history and the general view of purposivism as a philosophy in opposition to textualism makes the use of legislative history a useful proxy for purposivist methodology.[74] In contrast, textualist judges are typically distinguished by their emphasis on the "plain meaning" of statutes,[75] the use of dictionaries to determine plain meaning,[76] and canons of interpretation.[77]

The specific terms selected, and the rationales behind them, are described in Section B of the Appendix. The full source code, including all of the phrases used as proxies in this Article, is publicly available online.[78] I conduct several robustness checks in Part IV to ensure that the measures used in this Article are valid; in particular, I spot-check term frequency results in Section IV.A by randomly sampling opinions containing terms I designate textualist, purposivist, interpretive, or normative, in order to ensure that they match conventional conceptions of these methodologies.

All of the analysis in this Article was conducted by downloading bulk data and using Python code to analyze text. Past research has generally relied either on manual tabulation of the occurrences of certain terms, or on searches in Westlaw or Lexis.[79] Programming automates these tasks and makes the

[74] Bruhl, *supra* note 43, at 29.

[75] *See* William N. Eskridge, *The New Textualism*, 37 UCLA L. REV. 621, 623–25 (1990) ("[N]ew textualism posits that once the Court has ascertained a statute's plain meaning, consideration of legislative history becomes irrelevant.").

[76] Bruhl, *supra* note 43, at 29 ("A judge that uses linguistic canons and dictionaries extensively but uses legislative history sparingly is more textualist than a judge who displays the opposite tendencies.").

[77] *See, e.g.,* Gluck & Posner, *supra* note 52, at 1303–04 ("Textualists advanced the canons, in particular, as a more objective and coordinating set of tools for resolving statutory disputes than alternatives like legislative history . . . ."); John F. Manning, *Legal Realism & the Canons' Revival*, 5 GREEN BAG 2D 283, 290 (2002) ("Because textualists believe in a strong version of legislative supremacy, their skepticism about actual intent or purpose has predictably inspired renewed emphasis on the canons of interpretation, particularly the linguistic or syntactic canons of interpretation."). Borrowing from Aaron Bruhl, I divide canons of construction between "substantive canons," "language canons," and "holistic-textual canons." Bruhl, *supra* note 43, at 26, 64. The language canons and holistic-textual canons are most closely associated with textualists. *See* Bruhl, *supra* note 43, at 36; *infra* Appendix Section B.2.

[78] *Code*, JONATHAN H. CHOI, https://www.jonathanhchoi.com/code (last updated Aug. 1, 2019).

[79] Bruhl, *supra* note 43, at 30 ("[T]he analyses in this Article rely on electronic searches, primarily in Westlaw, to identify and count cases."); Solan, *supra* note 72, at 454 nn.118–19 (using Lexis searches to assess methodology).

analysis more flexible. This enables the application of machine learning techniques, as well as more granular detection and avoidance of false positives and negatives—for example, this Article counts appearances of the phrase "tax administration" but excludes the phrase "effective tax administration" (a term of art referring to a particular type of IRS settlement[80]), which would not be possible using a typical Boolean search without entirely excluding any documents that discuss "effective tax administration."[81] It also permits more detailed analysis by political affiliation and case outcome,[82] and the robustness checks in Part IV.

Most importantly, coding on raw data allows analysis of term frequency—the number of times a phrase appears in a document divided by the word count of the document—rather than a binary analysis of whether or not a phrase appears in the document at all, which is all that is feasible using a word search in Westlaw or Lexis.[83] Word searches only return the raw number of documents that contain any mention of a particular search term and cannot account for characteristics of the documents retrieved. This means that they cannot consider the number of times a search term appears in the document, or the length of the document.

Because the average length of judicial and administrative decisions has varied over time, certain terms might appear more or less purely as a function of greater or lesser detail, rather than due to trends in judicial methodology. For example, the average length of Tax Court opinions has significantly increased over time. If this phenomenon simply resulted from a trend toward more thorough descriptions of the rationales behind rulings,

[80] *See* DEP'T OF TREASURY, INTERNAL REVENUE MANUAL § 4.18.3 (defining "Effective Tax Administration Offers.")

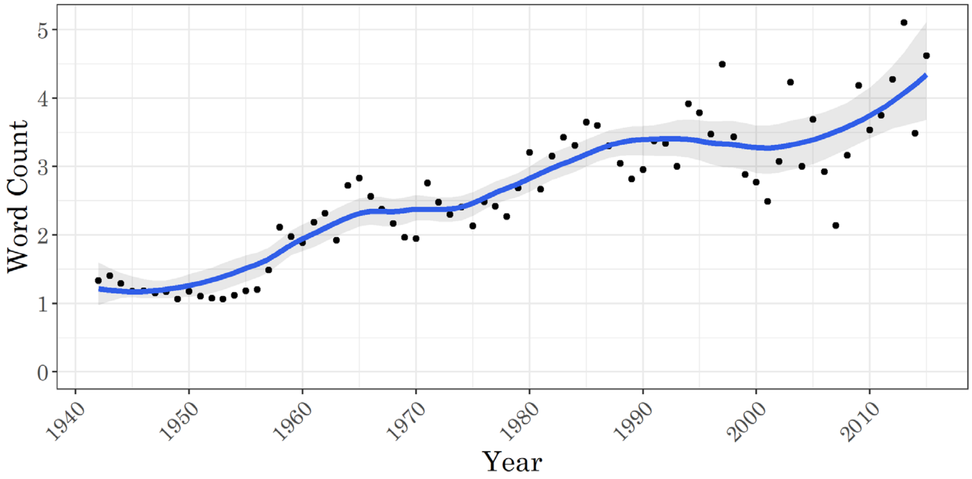
[81] For example, in a Westlaw search, one could search for "tax administration," and one could search for "tax administration % 'effective tax administration,'" ("%" is the symbol for "not" in Westlaw searches), but the latter search would not pick up a document that *both* included a legitimate occurrence of "tax administration" *and* an occurrence of "effective tax administration." *See* WESTLAW, SEARCHING WITH TERMS AND CONNECTORS 4 (2009).

[82] *See infra* Sections III.F, III.G.

[83] Lexis and Westlaw do allow searches for documents that contain a particular term at least a certain number of times. But this would be an impracticably unwieldy method to determine term frequency count, since it would have to be run many times to determine how many terms use a term at least once, at least twice, at least three times, and so on. For example, to determine normative, interpretive, textualist, and purposivist scores just for the IRS would take 339,000 separate manual searches, conservatively assuming thirty occurrences per term per year, and assuming that the "at least" search function could be used with proximity searches (which it cannot). (339,000 equals 113 terms, times thirty searches per term per year, times 100 years.)

then a study that counted the number of opinions containing certain tools would overestimate reliance on those tools in later periods.[84]

Figure 3: Average Word Count of Tax Court Opinions



In addition, a mere count of documents containing a particular phrase cannot measure the "intensity" of that phrase's usage. It might be cited once in passing, or many times as the central rationale to a ruling, but the numerical result would be the same. Term frequency addresses both the problems addressed above—it places a lower value on a phrase that only appears once in a long document, compared to a phrase that occurs many times in that same document.

B. Machine Learning

This Article uses term frequency analysis to illustrate broad trends, such as the Tax Court's movement toward textualism. For more granular analysis on specific interpretive tools, it turns to machine learning. Machine learning, broadly stated, uses algorithms based on a mathematical model to

[84] Note that while measuring term frequency tends to mitigate this problem, it is not a complete solution. Term frequency merely applies a linear adjustment for word count, but the relationship between interpretive depth and word count is not likely to be perfectly linear. For example, if court opinions less than 3000 words never engaged in any interpretation, but all of the words between the 3000th and the 4000th involved interpretation, then word count minus 3000 (minimum 1) would be the more appropriate denominator in calculating the degree of textualism or purposivism in an opinion.

make predictions or decisions without explicit human direction.[85] In doing so, machine learning can uncover trends and test hypotheses that would be onerous or potentially unreliable for humans to analyze manually.

This Article uses a binary classification model in order to test whether the court that wrote a given Tax Court opinion can be identified based on methodology alone. First, each opinion in the dataset is converted from plain text into a "vector" of numbers based on the occurrences of each interpretive tool in that opinion.[86] The classifier must be trained to predict which court wrote a particular opinion based on its vector.[87] To accomplish this, the opinions are randomly divided into a "training set," consisting of 80% of the opinions in the sample, and a "test set," consisting of the other 20%. The classifier repeatedly attempts to classify the opinions in the training set, hundreds of thousands of times, with small tweaks to the classifier between each iteration. The tweaks are retained if the classifier's performance improves and discarded otherwise. By iterative machine learning, the classifier is improved until its accuracy reaches a maximum.[88]

After the training is completed, the performance of the classifier is evaluated using the test set. This entire process is then repeated five times (this is known as "five-fold cross-validation") in order to ensure that the results are robust and not dependent on the specific training and test sets chosen.[89] By comparing the classifier's predictions for the test set with the actual classifications, we can produce various metrics of its predictive abilities. Additional technical detail on the machine learning methodology is provided in Section D of the Appendix.

[85] For a general explanation of machine learning methods, see TREVOR HASTIE, ROBERT TIBSHIRANI & JEROME FRIEDMAN, THE ELEMENTS OF STATISTICAL LEARNING: DATA MINING, INFERENCE, AND PREDICTION (2d ed. 2009).

[86] All of the machine learning in this Article is conducted using a "bag-of-words" approach (i.e., analyzing only the terms used, without regard to grammar or word order), using a Python utility provided by the Scikit-Learn project. I specifically use a count vectorizer, term frequency-inverse document frequency transformer with logarithmic term frequencies, and logistic regression (with five-fold cross-validation, 500 maximum iterations, and refitting). *See* SCIKIT-LEARN, https://scikit-learn.org/stable (last visited Aug. 1, 2019). Section D of the Appendix discusses machine learning methodology in more detail.

[87] This is only a simplified description—in practice, the vector is transformed before it is used to classify data. *See infra* Appendix Section D.

[88] Specific algorithms will vary in how they implement the general concept of iterative improvement, often using mathematical models. *See, e.g.*, FABRIZIO SEBASTIANI, MACHINE LEARNING IN AUTOMATED TEXT CATEGORIZATION 10 (2001) (describing the "inductive construction of the classifiers").

[89] This Article uses five-fold cross validation. *See supra* note 86; *cf.* George Self, *Why and How to Do Cross Validation for Machine Learning*, TOWARDS DATA SCI., https://towardsdatascience.com/why-and-how-to-do-cross-validation-for-machine-learning-d5bd7e60c189 (last visited June 27, 2019) (describing cross validation in machine learning).

The most widely endorsed measure of predictive performance is the Matthews correlation coefficient (MCC),[90] which produces a score between -1 and +1, where +1 represents perfect correlation (perfect prediction), -1 represents perfect inverse correlation (again, perfect prediction), and 0 represents no correlation (the worst possible score, no better than random). The interpretation of coefficients is highly subjective; however, as an extremely rough rule of thumb, a coefficient might be considered weak or negligible if its absolute value were less than 0.3; moderate between 0.3 and 0.7; and strong above 0.7.[91]

For completeness, I also list each classifier's "accuracy" (also known as the "correct classification rate"[92]) and "F_1 score."[93] Accuracy is the most intuitive measure of predictive power, representing the percentage of all predictions that were correct.[94] However, it is ill-suited to imbalanced datasets—in an extreme case with ninety-nine observations in category 1, but just one observation in category 2, a classifier that always guessed category 1 would still have an accuracy of 99%. The MCC, and to a lesser extent F_1 score, accounts for this problem.[95]

The classification method I use[96] assigns weights to each of the terms in the vocabulary, which facilitates more granular analysis of how strongly each term is associated with each category—for example, to what degree the

[90] *See, e.g.*, Davide Chicco, *Ten Quick Tips for Machine Learning in Computational Biology*, 10 BIODATA MINING 1, 11 (2017) ("[W]e strongly encourage to evaluate [sic] each test performance through the Matthews correlation coefficient (MCC), instead of the accuracy and the F1 score, for any binary classification problem.").

[91] E. GARCIA, A TUTORIAL ON CORRELATION COEFFICIENTS 8–9 (2011). Garcia notes that correlation coefficients may be weaker than initially supposed if degrees of freedom are low due to a small sample size. *Id.* at 10. This is generally not an issue for the tests in this Article, which use relatively large sample sizes.

[92] *See, e.g.*, Pozen et al., *supra* note 72, at *10.

[93] *See* KEVIN P. MURPHY, MACHINE LEARNING: A PROBABILISTIC PERSPECTIVE 182–83 (2012) ("Precision measures what fraction of our detections are actually positive, and recall measures what fraction of the positives we actually detected. . . . These are often combined into a single statistic called the F score, or F1 score, which is the harmonic mean of precision and recall." (emphasis omitted)).

[94] *See id.* at 182.

[95] GARCIA, *supra* note 91, at 8–9. In technical terms, MCC is the only one of the three measures that factors in every quadrant of the "confusion matrix": that is, true classification into Category 1, false classification into Category 1, true classification into Category 2, and false classification into Category 2. *See* Pierre Baldi et al., *Assessing the Accuracy of Prediction Algorithms for Classification: An Overview*, 16 BIOINFORMATICS REV. 412, 415 (2000) (noting that MCC "uses all four numbers" and therefore "may often provide a much more balanced evaluation of the prediction"). I also correct for imbalanced datasets by undersampling from the over-represented dataset until the sample is evenly balanced between the two categories. *See infra* note 131.

[96] Specifically, I use logistic regression with cross-validation. *See supra* note 86.

"rule of lenity" is associated with the Tax Court or with district courts. In Section III.E, I use these data to produce word clouds to illustrate the interpretive tools most characteristic of each court.

C. Regression Analysis

While natural language processing and machine learning are useful in mapping general interpretive trends and identifying which courts use which particular tools, they are less appropriate in identifying the causal relationship between interpretive methodology and case characteristics. For example, as Section III.F illustrates, casual examination of cases might suggest that Democratic Tax Court judges are less likely to use purposivist terms in their opinions. But this apparent correlation could be caused by other factors, like the year an opinion was written or the year that the judge who wrote it was appointed. When controlling for these factors, the ultimate result is the reverse. Regression analysis allows separate consideration of each of these contributors to methodology.

Section E of the Appendix contains additional technical detail on regression methodology. Because the term frequencies in Tax Court cases do not follow a normal distribution, I rely on two-part regression (logit and a log-transformed generalized linear model) rather than ordinary least squares regression. Sections C through G of the Appendix present additional robustness checks in light of the distributional issues in the dataset.

D. Limitations

1. Term Frequency as a Proxy for Methodology

Despite its advantages, term frequency analysis has some limitations. For one, it does not capture whether courts cite a certain interpretive tool approvingly or disapprovingly. A critic might speculate that the Tax Court began to cite textualist tools not in order to follow general judicial trends, but merely to observe and criticize those trends. While reviewing sources to select the terms analyzed in this Article, as well as during the ex post checks in Section IV.A, I did not find this to be the case—in fact, I found no disapproving citations of either textualist tools or legislative history in any IRS or Tax Court document.[97] Moreover, a disapproving mention of a

[97] On the other hand, interpreters sometimes cite evidence for one view even if they ultimately decide the other way—but this is to be expected in the ordinary course of statutory interpretation, where different sources may disagree.

specific tool would still suggest that the author considers the tool important to others, even though the author disputes its validity.[98]

A second limitation is that term frequency will not always reflect how important a particular interpretive tool was to the judge's ultimate decision. Legislative history, for example, might be a decisive factor in a court's ruling, even though it is only mentioned once. Or it could be mentioned several times, even though the court ultimately decides the case on other grounds.

To address this limitation, this Article focuses not on absolute results, but on *relative* results. It would be problematic to use term frequency in isolation to assess how textualist a particular court opinion was or, indeed, how textualist the Tax Court as a whole was in any particular year. Instead, this Article always asks how many textualist terms the entire Tax Court used this year compared to last year, or compared to some other court in the same year.

Imagine that dictionaries were infrequently cited by courts but that, when they are cited, they are only mentioned once and with decisive effect. This would imply that term frequency is not a reliable means to compare dictionary use with, say, legislative history—and this Article does not do so. Instead, this Article considers whether an authority cites dictionaries more *over time*. Consequently, a skeptic would need to argue that the way they are cited has changed over time. I have found no evidence of such changes while individually reading cases and agency guidance to validate the terms selected. Moreover, it is reassuring that the term frequency metrics in this Article frequently move in opposite directions. So any hypothesis as to why textualist terms have become more commonly cited at the Tax Court would need to explain why, during the same period of time, purposivist terms have declined at the Tax Court and textualist terms have declined at the IRS.

More broadly, by examining long-term methodological trends, averaged over many different documents and many consecutive years, this Article avoids the idiosyncrasies of single documents and single authors. This again reduces the likelihood of bias from particular administrators or judges. The challenge must not merely be that one judge varies her usage between periods, but that all judges vary their usage on average between periods for some reason other than methodological shifts.

2. *Doing Different Things, Doing Things Differently*

[98] *See* Anita S. Krishnakumar & Victoria F. Nourse, *The Canon Wars*, 97 TEX. L. REV. 163, 182 (2018) ("It is not necessarily the case, for example, that the most frequently invoked interpretive rule is also the most universally accepted. Nevertheless, frequency of judicial invocation does capture an important aspect of what it means to be well-established and entrenched in the legal community.").

Differences between the IRS and the Tax Court might arise not from differences in methodology, but differences in subject matter. For example, perhaps more complex issues inherently demand more purposivist methodology, and perhaps the IRS generally handles more complex matters than the Tax Court. Consequently, methodological divergence between the IRS and the Tax Court may not reflect a difference in their dispositions toward the same interpretive questions, but merely that the IRS and Tax Court serve fundamentally different roles.[99] To borrow Aaron Bruhl's terminology, the IRS and the Tax Court may be both "doing different things" and "doing things differently."[100]

It is undoubtedly true that the IRS and Tax Court do different things in a broad sense. Published Tax Court decisions focus on novel and substantive legal questions, whereas many IRS publications focus on procedural issues. The initial visual presentation of methodological trends in Sections III.A through III.D therefore do not conduct comparisons of the absolute frequencies of particular terms between authorities. Instead, these Sections focus on relative methodological changes *within* various authorities *over time*. In doing so, they avoid the difficulties attending comparisons between various authors.

This approach ameliorates but does not eliminate the problem. Especially over long periods, any interpreter might both change the statutes it interprets (as the statutes themselves are amended) and its interpretive preferences holding statutes constant. For example, Section III.A suggests that the IRS may have become less interpretive as the tax code itself expanded, leaving less statutory ambiguity for the IRS to resolve (doing different things). But Section III.A also suggests that the IRS may have been inspired by *Chevron* to take a more normative approach in reading the tax code (doing things differently).

Similarly, the machine learning analysis in Section III.E compares Tax Court methodology in tax cases with the methodology of District Courts and the Court of Federal Claims across those courts' complete dockets. Again, the distinction between doing different things and doing things differently is blurred; as Section III.E notes, it is likely that both differences play a role in the ability of the algorithm to distinguish opinions written by the various courts. While terms specific to tax law are not included in the analysis, it is hardly surprising that, for example, an area of law dominated by the practice of an agency (the IRS) tends to cite *Chevron* more often.[101] And this finding does not imply that the Tax Court would be more likely to

[99] *See infra* note 3; *infra* Section I.B.

[100] Bruhl, *supra* note 43, at 6 ("[C]ourts at different levels of the system are both *doing different things* and *doing things differently*.").

[101] *See infra* fig. 8.

cite *Chevron* than a District Court if they were both interpreting the same statute.

Ultimately, these examples illustrate the difficulty of drawing causal inferences from descriptive term frequency statistics alone. It would be risky to attempt to assess the extent to which different authorities are "doing different things" or "doing things differently" based solely on term frequencies. Instead, I try to tease out causal explanations using historical and primary sources.

The task of this Article is a more modest one, set against the virtual absence of any existing empirical evidence. This Article merely asks whether different courts differ in their methodological approach, for whatever reason. In doing so, it sets a baseline by suggesting that there are indeed substantial differences in interpretive style between different courts. Whether different courts would use different interpretive methodologies when confronted with the *same* statutes remains an important question for future research.

III. Results

Part I describes two major methodological dichotomies: normativity versus interpretation, and textualism versus purposivism. Agencies and courts make "normative" decisions when they justify their rulings on policy grounds, like "fairness" or "efficient administration."[102] They make "interpretive" decisions when they describe the interpretation of statutes, as when they "interpret the Code."[103]

Once an agency or court decides to engage in statutory interpretation, it may further decide to use textualist tools—like dictionaries—or purposivist tools—like legislative history.[104] Finally, an authority that leans either textualist or purposivist might still use different specific interpretive tools—one purposivist might emphasize committee hearings, for example, and another might emphasize committee reports. This Part examines variation among authorities and over time, along all these dimensions.

A. The IRS Has Become More Normative and Less Interpretive

As between normative and interpretive decisionmaking, the IRS has substantially moved over the past century away from interpretation and toward justifying its rulings on normative policy grounds.

[102] *See infra* Appendix Section B.4.

[103] *See infra* Appendix Section B.3.

[104] *See infra* Appendix Sections B.1, B.2.

Figure 4: Interpretive and Normative Terms in IRS Publications

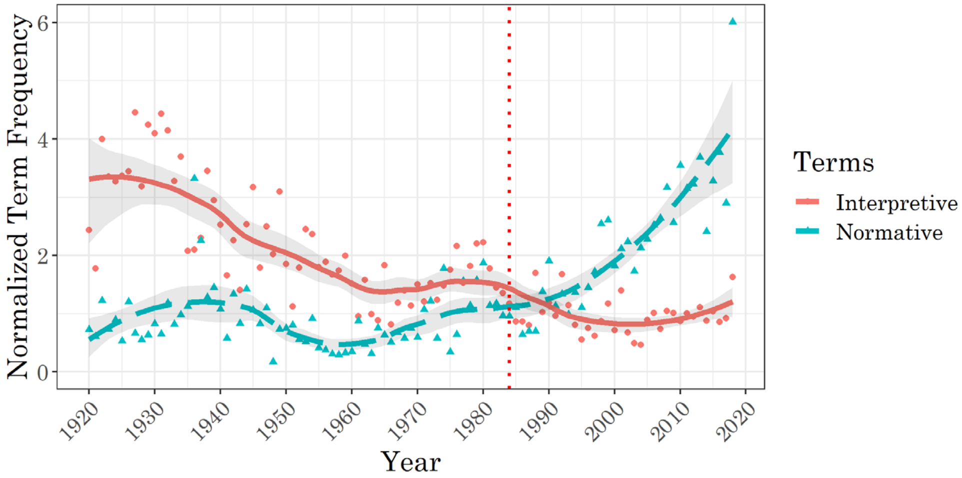


Figure 4 is consistent with the view that *Chevron* marked a normative shift. The IRS's use of normative terms markedly accelerated after 1984, the year that *Chevron* was decided. The year *Skidmore* was decided, 1944, seems to reflect a mere continuation of a downward trend in interpretation. At the same time, *Chevron* is not the sole plausible causal explanation for the rise in normative terms—*Chevron* coincided with a number of other Reagan-era events that may have affected agency practice, such as the institution of cost-benefit analysis in 1981,[105] the appointment of Justice Scalia in 1986 and the rise of textualism throughout the 1980s, and the continuing popularization of law and economics through the 1980s. However, while *Chevron* deference may not be the sole or even the proximate cause of the increase in normative terms, it is notable that the normative shift would not have been possible if agencies had continued to constrain themselves strictly to interpretive matters, as Elliott has argued.[106]

In contrast to the sudden uptick in normative terms, how can we explain the long decline in interpretive terms that predated both *Skidmore* and *Chevron*? One potential explanation is that as the tax code matured, there was less and less statutory ambiguity to be resolved by regulations and rulings. When federal income tax laws were first passed, a greater part of the IRS's work consisted of basic interpretive issues, deciding on the correct reading of this or that section of the Code.[107] As the interstices of the Code were filled,

[105] Exec. Order No. 12,291, 46 Fed. Reg. 13,193 (Feb. 17, 1981).

[106] *Supra* notes 28-29 and accompanying text.

[107] *Cf.* Jonathan H. Choi, *The Substantive Canons of Tax Law*, 72 STAN. L. REV. (forthcoming 2020) (manuscript at 41–43) (on file with author) (describing how, "[d]uring the infancy of the federal income tax, . . . statutes were relatively sparse and agency practice

the IRS shifted to more granular policymaking details, beginning to offer clarifications of its own regulations rather than original interpretations of statutes.

A related hypothesis is that the IRS has gained expertise over time. The IRS today employs a variety of technical experts, including statisticians, economists, and computer researchers.[108] These specialists might provide the IRS the means to make more sophisticated normative judgments, including more accurately estimating the real-world impact of particular tax policies.

Regardless of the precise explanation, historical documents reflect the overall narrative that the IRS has grown more normative and less interpretive over time. The IRS itself was concerned with a declining focus on faithful interpretation as early as the 1960s. In 1964, it issued a Revenue Procedure stating, in part:

> At the heart of administration is interpretation of the Code. It is the responsibility of each person in the Service, charged with the duty of interpreting the law, to try to find the true meaning of the statutory provision and not to adopt a strained construction in the belief that he is "protecting the revenue." The revenue is properly protected only when we ascertain and apply the true meaning of the statute.[109]

This statement was reproduced at the front of every Cumulative Internal Revenue Bulletin from 1970 to 1999, "to emphasize [its] importance to all employees of the Internal Revenue Service."[110] The IRS's stress on faithful interpretation may be responsible for the bump in interpretive terms during this period.

But the shift away from interpretive decisionmaking has resumed in the past few decades. A survey of recent trends in IRS policy reflects this.

was relatively uncertain"). *See generally* Lawrence A. Zelenak, *Leaving It Up to Treasury: Congressional Abdication on Major Policy Issues in the Early Years of the Income Tax*, 81 L. & CONTEMP. PROBS. 137 (2018) (describing how the early income tax code was silent or ambiguous on a number of essential issues, leaving them to be resolved at the discretion of the Treasury).

[108] *Research & Analysis | IRS Careers*, INTERNAL REVENUE SERV., https://www.jobs.irs.gov/resources/job-descriptions/research-analysis (last visited Nov. 3, 2019).

[109] Rev. Proc. 64-22, 1964-1 C.B. 689; *see also* Rev. Proc. 2000-43, 2000-2 C.B. 404 (citing with approval the portion of Revenue Procedure 64-22 discussing administration); Rev. Proc. 2012-18, 2012-10 I.R.B. 455 (same).

[110] *E.g.*, 1984-1 C.B. ii. The Cumulative Internal Revenue Bulletin is a compilation of all the Internal Revenue Bulletins issued in each year.

The IRS was reformed in the mid-1990s to have an increased emphasis on service to taxpayers and taxpayer rights.[111] In the 2000s, the IRS shifted its focus to the increased proliferation of abusive multibillion-dollar tax shelters.[112] And the most recent movement, following a series of directives by the Trump administration, has been to cultivate regulations that are "simple, fair, efficient, and pro-growth."[113]

The IRS continues to juggle each of these concerns in its modern policymaking: simplicity, clarity, fairness, efficiency, and most of all its central function of raising revenue. It has been aided by an extensive scholarly literature addressing each of these goals.[114] But these are all *normative* goals, not interpretive ones. Whether inspired by *Chevron*, by modern political trends, or by some combination thereof, the IRS has moved decisively toward normativity in its rulings.

[111] *See, e.g.*, Taxpayer Bill of Rights 2, Pub. L. No. 104-168, 110 Stat. 1452 (1996) (listing the rights of taxpayers in dealing with the IRS); Internal Revenue Service Restructuring and Reform Act of 1998, Pub. L. No. 105-206, 112 Stat. 685 (1998) (reforming the IRS, with an eye to improving taxpayer service); I.R.C. § 1503 (1998) (requiring, generally, that IRS employees be fired if they engage in one of ten kinds of anti-taxpayer conduct).

[112] *See, e.g.*, Joseph Bankman, *Tax Enforcement: Tax Shelters, the Cash Economy, and Compliance Costs*, 31 OHIO N.U. L. REV. 1, 2 (2005) (describing evidence of huge tax shelters in the early 2000s); Tanina Rostain, *Sheltering Lawyers: The Organized Tax Bar and the Tax Shelter Industry*, 23 YALE J. REG. 77, 79 (2006) (describing efforts since the late 1990s to fight tax shelters). For a history of the tax shelter movement of the 2000s, see generally TANINA ROSTAIN & MILTON C. REGAN JR., CONFIDENCE GAMES: LAWYERS, ACCOUNTANTS, AND THE TAX SHELTER INDUSTRY (2014).

[113] Exec. Order No. 13,789, 82 Fed. Reg. 19,317 (Apr. 21, 2017). The IRS consequently identified and removed 296 regulations that it deemed "no longer necessary because they do not have any current or future applicability." T.D. 7805 (Mar. 15, 2019). While these executive orders were stated in very general terms, they are nominally binding on the IRS and would have constituted explicit pressure to take normative considerations into account. *See* Mashaw, *supra* note 3, at 506 ("[B]oth as a practical political and as a normative constitutional matter, we should expect agencies to interpret statutes in the context of presidential direction."); *see also* Exec. Order No. 13,777, 82 Fed. Reg. 12,285 (Feb. 24, 2017) (requiring agencies to undertake reforms intended to "lower regulatory burdens on the American people"); Exec. Order No. 13,789, 82 Fed. Reg. 19,317 (Apr. 21, 2017) (same, specifically with respect to the IRS). *See generally* Elena Kagan, *Presidential Administration*, 114 HARV. L. REV. 2245 (2001) (discussing presidential direction of agencies).

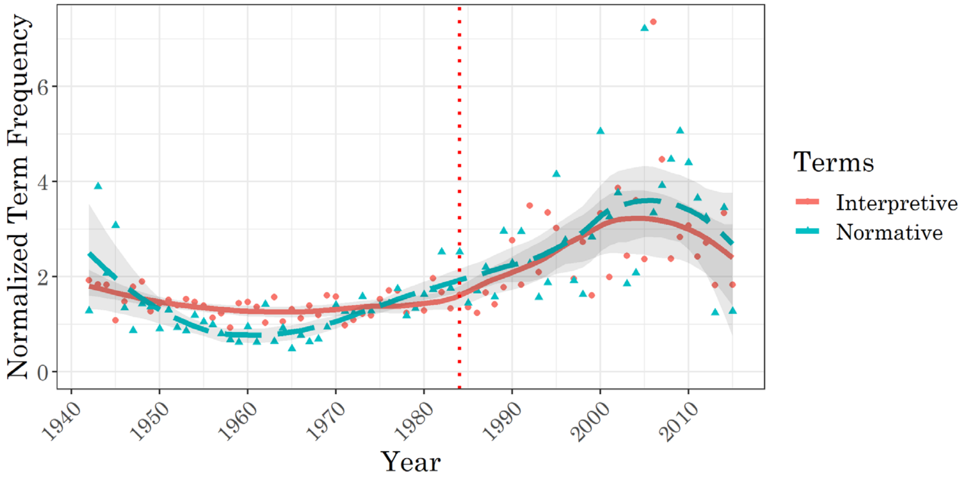
[114] *See, e.g.*, Lily L. Batchelder, Fred T. Goldberg & Peter R. Orszag, *Efficiency and Tax Incentives: The Case for Refundable Tax Credits*, 59 STAN. L. REV. 23 (2006) (discussing efficiency and revenue-raising); John A. Miller, *Indeterminacy, Complexity, and Fairness: Justifying Rule Simplification in the Law of Taxation*, 68 WASH. L. REV. (1993) (discussing simplicity, clarity, and fairness).

B. The Tax Court Has Maintained the Same Proportion of Interpretation and Normativity

But what about the Tax Court? Tax Court judges are likely aware of broader trends, like the controversies surrounding tax shelters from the 2000s. At the same time, Tax Court judges are (at least in theory) impartial arbiters not directly responsible to the executive branch,[115] such that their priorities might vary from the current priorities of the administration.

It turns out that the Tax Court has remained remarkably steady over the years in its mix between normative and interpretive terms. The two have fluctuated within a smaller range for most of the life of the Tax Court. And, importantly, they have generally moved in tandem rather than inversely, in contrast to the IRS.

Figure 5: Interpretive and Normative Terms in Tax Court Opinions



In addition to providing evidence of consistent priorities over time at the Tax Court, Figure 5 also contrasts well with Figure 4, suggesting that the variation demonstrated in Figure 4 is a true effect rather than just noise.

On the other hand, while the Tax Court has remained relatively consistent in the proportion of interpretive terms and normative terms it uses for any given year, the frequency of both types of term has changed over time. Most broadly, both types of term have become more common from a relatively low level during the 1940s through 1970s, to a higher level at present. This could reflect the trend noted in Figure 3, that later Tax Court opinions tend to be longer. If all Tax Court opinions reflect some fixed amount of factual and procedural recitation, longer opinions might cause

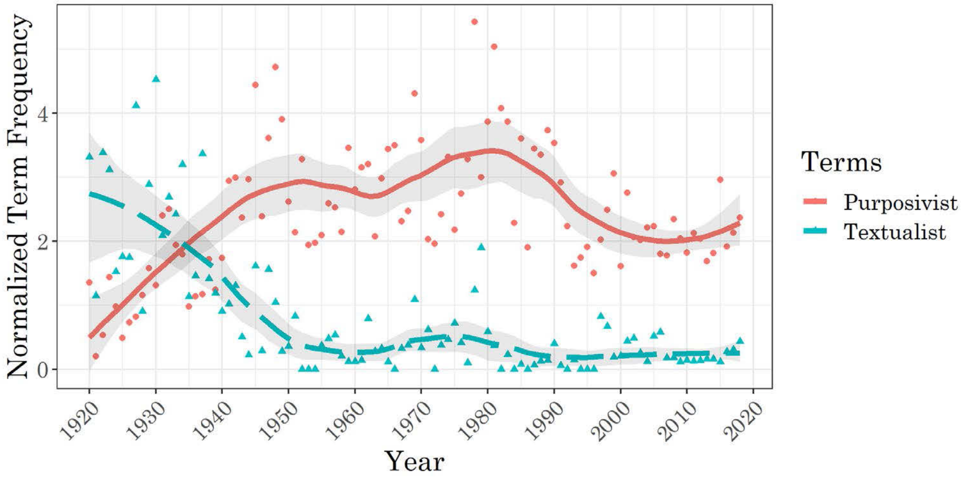
[115] Tax Court judges may only be removed with cause. *See supra* note 70.

more space to be allotted to legal analysis. However, this story cannot explain the small dips in the frequency of normative and interpretive terms during the 1940s and 2010s. While these dips cannot be attested with a high degree of confidence—they fall within the confidence intervals, suggesting they may be aberrations rather than true trends—they might be a fruitful subject for future research.

C. The IRS Has Become More Purposivist and Less Textualist

Given that the IRS has significantly reduced the amount of interpretation that it conducts, the next question is whether it has also changed the *type* of interpretation it conducts. As noted above, the Supreme Court and district courts became more purposivist during the 1930s and 1940s but then became less purposivist and more textualist during the 1980s and 1990s.[116] Interestingly, the IRS followed the first shift but not the second, remaining resolutely purposivist despite the rise of the judicial new textualism:

Figure 6: Purposivist and Textualist Terms in IRS Publications



In fact, in many recent years, the IRS has made almost no references to plain meaning, dictionaries, or the various language canons that I use as proxies for textualism and which the Supreme Court (like the Tax Court, as noted below) has readily adopted. While at first it may appear that the IRS has reduced its use of purposivist terms since 1980, this matches the overall decline in interpretation discussed in Section III.A. It is worth noting that the IRS's use of textualist terms has declined by at least as much over the same

[116] *Supra* Section I.B.

period, such that the relative mix between textualism and purposivism still strongly favors purposivism.

What has prevented the IRS from adopting textualism? My view is that the IRS's close involvement in the legislative process—its role in advising Congress during the drafting of bills[117] and its deep institutional knowledge of the intended meaning of bills[118]—provide it with the means and the motivation to pay special attention to legislative history.[119] This is reflected by the fact that the IRS has chosen to publish legislative history, including relevant reports and hearings, in its Internal Revenue Bulletins since 1941.[120] 1941 marks the original rise of the administrative state as well as purposivism, since specialist agencies like the IRS were able to effectively interpret legislative history in a way that laypeople were not.[121]

This explanation is not specific to tax law—most agencies are involved in the process of drafting statutes and accordingly might have special expertise in interpreting legislative history.[122] Future research could usefully explore whether other agencies have also resisted the modern move toward textualism, like the IRS.[123]

[117] *See, e.g.*, Parrillo, *supra* note 43, at 266 ("By reason of its unprecedented manpower and its intimacy with Congress (which often meant congressmen depended on agency personnel to help draft bills and write legislative history), the administrative state was the first institution in American history capable of systematically researching and briefing legislative discourse and rendering it tractable and legible to judges on a wholesale basis."); Jarrod Shobe, *Agencies as Legislators: An Empirical Study of the Role of Agencies in the Legislative Process*, 85 GEO. WASH. L. REV. 451, 451 (2017) (finding "that agencies are deeply involved in drafting and reviewing statutory text before enactment, and . . . that Congress often relies heavily on agencies' significant legislative resources and expertise"); Strauss, *supra* note 30, at 1146 ("The agency may have helped to draft the statutory language, and was likely present and attentive throughout its legislative consideration.").

[118] *See* Wallace, *supra* note 63.

[119] *See supra* Section I.B.

[120] *E.g.*,1941-1 C.B. 479-576; 1941-2 C.B. 331-525.

[121] *See supra* notes 44, 117.

[122] *See supra* notes 117-121.

[123] An alternative explanation could be that the legislative history of tax statutes might be especially useful due to the work of the JCT. The JCT is a nonpartisan congressional committee that "assists with devising and drafting legislation, and, importantly, produces revenue estimates of every tax provision and prepares explanations of revenue-raising legislative proposals that Congress relies on throughout the legislative process." Wallace, *supra* note 63, at 183. However, the results in Section III.D immediately below weigh against this explanation. The Tax Court has the same access to JCT publications as the IRS, but it does not participate in the drafting of statutes as the IRS does and did not remain purposivist, unlike the IRS.

D. The Tax Court Has Become More Textualist and Less Purposivist

Section I.B observes the movement of federal courts over the past three decades away from purposivism and toward textualism. Yet the preceding Section indicates that purposivism remains dominant at the IRS. We might ask, as Section I.C does, which is the stronger driver of methodology: cohesion among specialists or cohesion among courts?

Figure 7: Purposivist and Textualist Terms in Tax Court Opinions

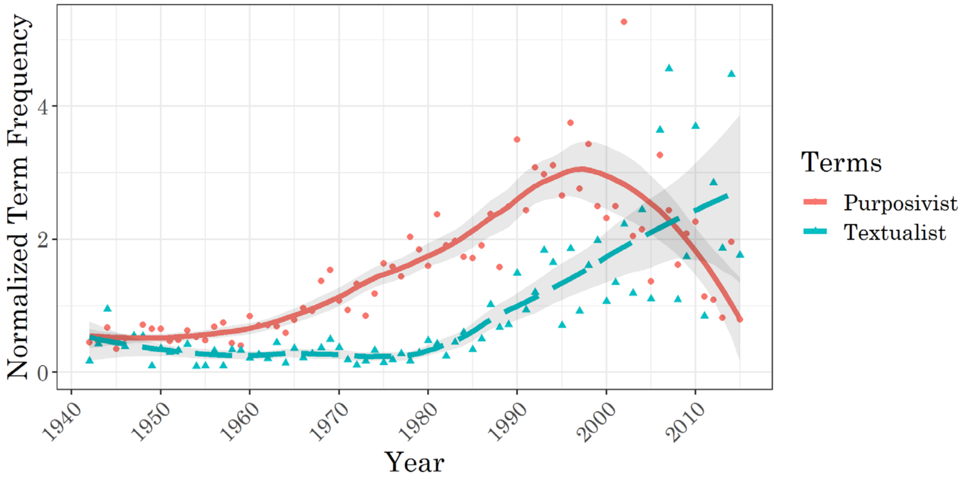


Figure 7 shows that methodological trends in the Tax Court most strongly resemble those in other federal courts. Like district courts, the Tax Court embraced textualist tools in the 1980s and 1990s.[124] The Tax Court also peaked and then declined in its use of legislative history, although it did so approximately a decade later then district courts, in the 1990s rather than the 1980s.[125]

The lag in the Tax Court's turn away from purposivism is somewhat puzzling. It may be a product of the Tax Court's continued reliance on certain interpretive tools, such as committee reports, in light of their continued use by the IRS and tax experts, or it may reflect reluctance to give up especially

[124] *Cf.* Bruhl, *supra* note 43, at 58–61 (using slightly different methodology but finding the same trend).

[125] *Id.* at 57–58. Tax Court data are only available from the court's founding in 1942, so it is difficult to gauge whether it would have participated in the move toward purposivism around that time. The Tax Court's predecessor, the Board of Tax Appeals, was an "independent agency in the executive branch" whose "decisions were not final and could be collaterally attacked in federal court," making it less appropriate for a study of judicial methodology. *See* Lederman, *supra* note 25, at 1841.

useful sorts of legislative history, like the "bluebooks" published by the JCT that summarize legislation in each session of Congress.[126] The following Section addresses this possibility in greater detail, but it remains a question that might be elucidated by future research.

A reader might also wonder how it is possible for the IRS and Tax court to diverge methodologically in the first place. Why would a textualist Tax Court not simply strike down guidance issued by a purposivist IRS? Judicial deference surely plays a role here—both *Chevron* and *Skidmore* provide agencies latitude to read statutes differently from courts.[127] The Tax Court is also constrained as a practical matter, since fifteen judges can only do so much to police the voluminous guidance that the IRS produces each week. Finally, professional respect may play a role. Deference aside, Tax Court judges may informally feel reluctant to repudiate IRS purposivism even if they would have applied more textualist tools when considering the same question ab initio. This Article does not draw any conclusion on the precise causal mechanism for the disconnect between the Tax Court and IRS. Likely each of these explanations plays some role, but future research could usefully investigate further.

As Section II.D.2 discusses, this Article attempts to distinguish "doing different things" from "doing things differently" by focusing on changes in relative term frequency over time. But there is a more pointed potential criticism that remains. *Chevron* tends to sort interpretive issues between those within the *Chevron* space and outside of the *Chevron* space. What if issues within the *Chevron* space require purposivist tools (like legislative history) more often, perhaps because they are more complex and ambiguous? Since agencies have exclusive jurisdiction over issues within the *Chevron* space, this would imply that the IRS appears more purposivist merely because *Chevron* has precluded courts from addressing the most purposivist questions. Likewise, it could be that courts became more textualist merely because the most purposivist issues were removed from their remit.

There are two main reasons to doubt this account. First, it contradicts most of the theoretical and anecdotal literature discussing the rise of the new textualism. This literature generally attributes the modern resurgence in

[126] *See Joint Committee Bluebooks*, JOINT COMMITTEE ON TAX'N, https://www.jct.gov/publications.html?func=select&id=9 (last visited July 15, 2019) (describing the bluebooks and providing access to copies since 1969). Note that the bluebooks might not technically be "legislative history," since they are produced after legislation has already been passed, but they are considered good evidence of contemporaneous understandings about legislation from the last session of Congress.

[127] *See supra* Section I.B.

textualism to the intellectual activity of textualists like Justice Scalia,[128] and the reports of judges that lean toward textualism generally reflect theoretical commitments to the primacy of statutory text.[129] I am not aware of any commentator or judge who has suggested that judges have become more textualist because they face a different set of issues than they did before *Chevron*.

Second, this sorting imperfectly fits the stories told by Figures 6 and 7. In Figure 6, IRS purposivism did not increase after *Chevron*—instead, it remained flat or perhaps even slightly declined. If more purposivist issues were sorted toward the IRS, we would expect the IRS to increase its use of purposivist tools.[130] In Figure 7, the decline of Tax Court purposivism occurred a decade later than the rise of Tax Court textualism, suggesting that the relationship between the two is more complex than a direct tradeoff due to sorting and that the decline in purposivism was not directly attributable to *Chevron*.

E. The Tax Court Has Developed a Unique Interpretive Methodology Relative to Other Courts

Although the Tax Court has generally become more textualist, the specific flavor of its textualism may differ from other trial courts. I use a machine learning classifier to test whether Tax Court opinions may be distinguished based on interpretive methodology alone.[131] I employ two binary classifications: the Tax Court versus generalist district courts, and the

[128] See, e.g., Abbe R. Gluck, *Justice Scalia's Unfinished Business in Statutory Interpretation: Where Textualism's Formalism Gave Up*, 92 NOTRE DAME L. REV. 2053, 2058 (2017).

[129] *See generally* Abbe R. Gluck & Richard A. Posner, *Statutory Interpretation on the Bench: A Survey of Forty-Two Judges on the Federal Courts of Appeals*, 131 HARV. L. REV. 1298 (2018) (reporting results from a survey of appellate judges, including judges that lean toward textualism).

[130] Note, however, that the IRS might not grow more purposivist if it were to solely apply normative criteria inside the *Chevron* space, so that interpretive issues within that space are not sorted to the IRS so much as removed from consideration by any authority.

[131] The results in this Section were produced using Tax Court and district court opinions from 2004 to 2018, the modern textualist era of these courts, in order to obtain current results. Because district courts, taken together, produce many more opinions each year than the Tax Court, the sample used for machine learning would tend to be highly imbalanced in favor of district courts. To correct for this, I randomly "undersample" district court opinions by excluding district court cases at random until the two samples are of the same size. *See generally* Nitesh V. Chawla, *Data Mining for Imbalanced Datasets: An Overview*, *in* DATA MINING AND KNOWLEDGE DISCOVERY HANDBOOK 853–67 (Oded Maimon & Lior Rokach eds., 2010) (describing undersampling).

Tax Court versus the Court of Federal Claims (CFC). The CFC is another Article I court that handles claims for monetary damages against the federal government.[132]

Both classifications perform moderately well based on the performance measures described above.[133]

Table 1: Tax v. District Court Classifier Performance

MCC:	0.546
Accuracy:	0.772
F_1 Score:	0.762

Table 2: Tax v. CFC Classifier Performance

MCC:	0.446
Accuracy:	0.717
F_1 Score:	0.707

This suggests that the Tax Court has indeed produced a style of statutory interpretation distinct from the district courts and the CFC, even though all of these courts have taken a broadly textualist turn.

Because the algorithm classifies opinions between the courts by assigning weights to each interpretive term, we can analyze these weights to see which terms are most strongly associated with the Tax Court. Figure 8 presents these weights, evaluating which terms are most predictive of Tax Court opinions (above the dotted line) and which are most predictive of District Court opinions (below the dotted line).

The listed values are coefficients generated through machine learning, from a logistic regression with log-transformed tf-idf as the independent variables.[134] Generally speaking, the coefficients should be interpreted as the products of a log-log regression, scaled (by virtue of the tf-idf transformation) so that rarer terms are not disproportionately significant. That is, before scaling, a coefficient of β implies that a k-fold increase in the frequency of a term is associated with an odds ratio of k^{β}. More concretely, if the coefficient in a log-log regression for a particular term (say, "*in pari materia*") were 2, then doubling the number of times "*in pari materia*" is used in a case would

[132] *See* 28 U.S.C. § 1491 (2012).

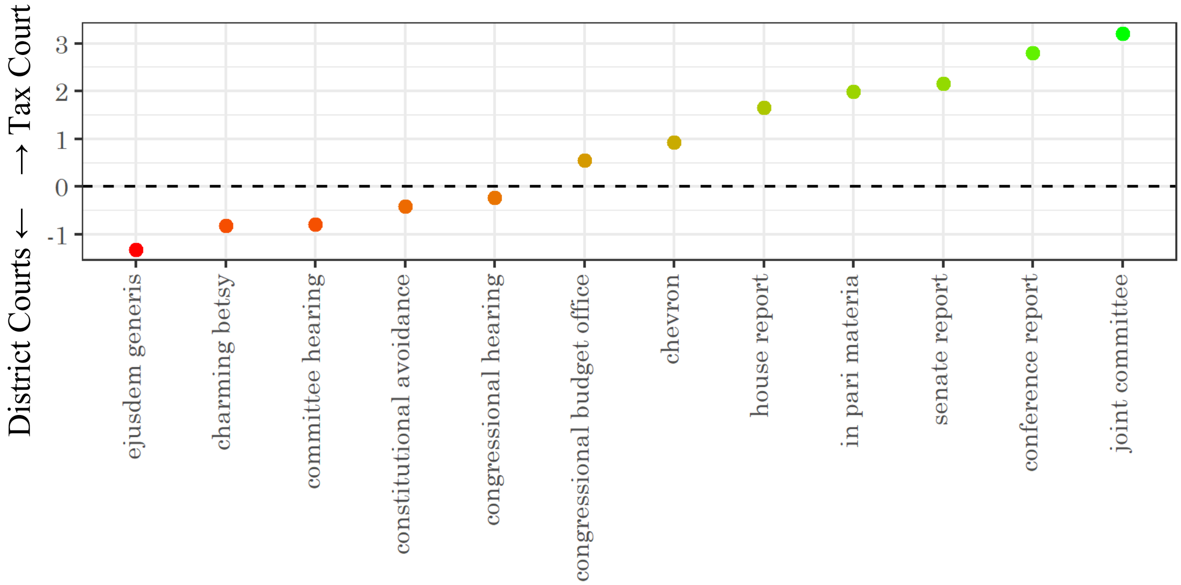
[133] Rather than relying on the results from a single iteration of the algorithm, these tables reflect the median values from repeated bootstrapping generated in Section IV.B.

[134] *See infra* Appendix Section D (discussing tf-idf transformation).

also increase the probability of a particular case being a Tax Court case to $2^2 = 4$ = 400% of what it would be otherwise. The coefficients in a log-log tf-idf regression can be interpreted in roughly the same manner, but they are scaled to reduce the outsize influence of rare terms. Appendix Section D contains a more specific mathematical description of how the coefficients are calculated.

These terms were selected because they were statistically significant at the 99% confidence level, using bootstrapped percentile confidence intervals, from bootstrapping with 100 iterations.[135] No other term was statistically significant above the 95% level, and consequently they were omitted.[136]

Figure 8: Interpretive Tools, Tax Court v. District Courts



The results in Figure 8 are intuitively sensible.[137] For legislative history, the Tax Court most heavily prioritizes congressional reports, Congressional Budget Office materials, and materials from the JCT, eschewing congressional hearings. Among textual canons, the Tax Court favors the *in pari materia* canon (requiring that sections of the tax code

[135] *See infra* Section IV.B, Appendix Section G (describing bootstrapping to derive confidence intervals in machine learning).

[136] As in Tables 1 and 2, the coefficients in Figure 8 are median bootstrapped values.

[137] *See supra* note 126 and accompanying text.

dealing with similar material "must be construed together"[138]). The prominence of the *in pari materia* canon is not too surprising—many scholars have observed (and approved of) tax authorities' determination to construe the tax code in a consistent manner,[139] and it is a familiar interpretive move to clarify an ambiguous section of the tax code by reference to other sections.[140] Likewise, the Tax Court's reluctance to deploy the *ejusdem generis* canon (requiring that when a general word follows specific words, the general word is assumed to include only words of a similar type[141]—for example, a statute allowing "dogs, cats, and other animals" in a park might not permit tarantulas) has been observed by other scholars. Most prominently, the definition of "income" has been expanded by courts far beyond the initial list of examples provided in the tax code.[142]

For substantive canons, the Tax Court favors *Chevron* deference.[143] This likely reflects the IRS's importance as the primary nexus for the administration of federal tax law, and deference to its regulations frequently

[138] *See, e.g.*, Merrill v. Fahs, 324 U.S. 308, 311, 313 (1945).

[139] *See, e.g.*, Ditslear & Brudney, *supra* note 43, at 1298–99 (describing the rule "that when Congress expresses or describes a tax law concept in one part of the Internal Revenue Code, that expression or description should be deemed probative regarding Congress's treatment of the concept in a separate part of the code").

[140] *See, e.g.*, Yates v. Hendon, 541 U.S. 1, 13–16 (2004); Drye v. United States, 528 U.S. 49, 56–57 (1999); United States v. Reorganized CF&I Fabricators of Utah, Inc., 518 U.S. 213, 222–23 (1996); United States v. Hill, 506 U.S. 546, 555–56, 556 n.7 (1993); United States v. Dalm, 494 U.S. 596, 601–02 (1990); United States v. Rodgers, 461 U.S. 677, 695–98 (1983); United States v. Consumer Life Ins. Co., 430 U.S. 725, 745–46 (1977); Laing v. United States, 423 U.S. 161, 176–77 (1976).

[141] *Cf.* Circuit City Stores, Inc. v. Adams, 532 U.S. 105, 114–15 (2001) ("Where general words follow specific words in a statutory enumeration, the general words are construed to embrace only objects similar in nature to those objects enumerated by the preceding specific words.").

[142] *See* I.R.C. § 61(a) (2012) (listing items that qualify as income); Alice G. Abreu & Richard K. Greenstein, *The Rule of Law as a Law of Standards: Interpreting the Internal Revenue Code*, 64 DUKE L.J. ONLINE 53, 71 (2015) ("[W]hen interpreting the meaning of income, courts often ignore the constraints of *ejusdem generis*.").

[143] The status of judicial deference regimes as substantive canons is not wholly uncontroversial, but I treat them as such for purposes of this Article without taking a position on that debate. *See* Connor N. Raso & William N. Eskridge, Chevron *as a Canon, Not a Precedent: An Empirical Study of What Motivates Justices in Agency Deference Cases*, 110 COLUM. L. REV. 1727, 1727 (2010) ("As a descriptive matter, we find that deference regimes are more like canons of statutory construction, applied episodically but reflecting deeper judicial commitments, than like binding precedents, faithfully applied, distinguished, or overruled."). They are, at least, important determinants of how statutes are read, as indicated above.

appears in Tax Court cases.[144] Conversely, the Tax Court avoids the *Charming Betsy* canon, stating "that ambiguous congressional statutes should be construed in harmony with international law,"[145] and the constitutional avoidance canon. This too makes sense, given that the Tax Court is rarely faced with questions of constitutionality or international law.

F. Democratic Judges Are More Purposivist and Republican Judges Are More Textualist at the Tax Court

Past empirical work has frequently asked whether Republican-appointed judges interpret statutes differently from Democrat-appointed judges.[146] Conventional wisdom holds that Republican judges lean textualist, and Democratic judges lean purposivist. This tendency has been observed at the Supreme Court, for example.[147]

I investigate this issue at the Tax Court by dividing opinions by authorship, between Democratic and Republican appointees.[148] A casual

[144] *See, e.g.*, Good Fortune Shipping SA v. Comm'r, 148 T.C. 262, 263, 275–84 (2017) (applying *Chevron* analysis); Lindsay Manor Nursing Home, Inc. v. Comm'r, 148 T.C. 235, 243–61 (2017) (same); N.J. Council of Teaching Hosps. v. Comm'r, 149 T.C. 466, 472 n.7 (2017) (applying *Skidmore* analysis).

[145] Note, *The Charming Betsy Canon, Separation of Powers, and Customary International Law*, 121 HARV. L. REV. 1215, 1215 (2008).

[146] *E.g.*, Brudney & Baum, *supra* note 52, at *33; Ditslear & Brudney, *supra* note 43, 1301; Krishnakumar, *supra* note 43, 274-78; David S. Law & David Zaring, *Law Versus Ideology: The Supreme Court and the Use of Legislative History*, 51 WM. & MARY L. REV. 1653, 1671 (2010); Semet, *supra* note 6, at 2289-99, 2314-27.

[147] *See, e.g.*, Law & Zaring, *supra* note 146, at 1654 ("[L]iberal Justices are generally more likely than conservative Justices to cite legislative history.").

[148] Because Tax Court judges serve fifteen-year terms, sometimes they will be reappointed upon the expiration of their terms. Usually, the reappointing President is of the same party as the President originally appointing the judge. For example, Judge Maurice B. Foley was appointed by President Clinton and reappointed by President Obama, *see Chief Judge Maurice B. Foley,* U.S. TAX CT. (Apr. 23, 2019), https://www.ustaxcourt.gov/judges/foley.htm, while Judge Thomas B. Wells was appointed by President Reagan and reappointed by President George W. Bush, *see Judge Thomas B. Wells*, U.S. TAX CT. (Feb. 13, 2013), https://www.ustaxcourt.gov/judges/wells.htm. A few judges were appointed and reappointed by Presidents from different parties—these judges are not clearly either Democratic or Republican and were therefore excluded for purposes of this analysis. For example, Judges Mary Ann Cohen and Joel Gerber were both appointed by President Reagan and reappointed by President Clinton. *See Judge Joel Gerber*, U.S. TAX CT. (Apr. 8, 2013), https://www.ustaxcourt.gov/judges/gerber.htm; *Judge Mary Ann Cohen*, U.S. TAX CT. (Oct. 2, 2012), https://www.ustaxcourt.gov/judges/cohen.htm. Some Tax Court opinions, particularly memorandum opinions, are written by "special trial judges," who are appointed by the Chief Judge of the Tax Court rather than by the President. *See* I.R.C. § 7443A (2012); Wright v. Comm'r, T.C.M. (RIA) 2013-68 (2013); Madison

survey of interpretive trends among these judges suggests, if anything, the opposite of the conventional story. The two Tax Court judges who have cited legislative history most often (as of 2015, when the court data were assembled)—Judges Morrison and Wright—were both appointed by Republican Presidents.[149] And the three Tax Court judges who have used textualist tools most often—Judges Lauber, Buch, and Nega—were all appointed by President Obama.

However, we should be skeptical of apparent partisan trends as merely collateral effects of larger time trends. Given that the three most textualist judges were appointed by President Obama, the question then becomes whether they are textualist because they were appointed by a Democratic President or because they were appointed recently.[150] That is, to what extent is party affiliation a misleading proxy for the year that an opinion was written or the year the author was appointed?

Because interpretive methodology could have multiple determinants, visual analysis of time trends and machine learning classifier analysis is potentially unreliable. The better approach is to conduct regression analysis that controls for variables other than party affiliation. The results of this regression analysis are excerpted in Table 3; Section E of the Appendix provides additional detail on methodology and provides full tables of results.

Recycling Assocs. v. Comm'r, T.C.M. (RIA) 2001-85 (2001). Since the ideology of a judge appointed by another judge rather than the President will be more attenuated, these opinions are excluded as well.

[149] *Judge Richard T. Morrison*, U.S. TAX CT. (July 19, 2019), https://www.ustaxcourt.gov/judges/morrison.htm; Press Release, U.S. Tax Court, Death Announcement - Senior Judge Lawrence A. Wright (Mar. 20, 2000), https://www.ustaxcourt.gov/press/032000.pdf.

[150] *See* Gluck & Posner, *supra* note 52, at 1300 ("[Y]ounger judges, who attended law school and practiced during the ascendance of textualism, are generally more formalist and accepting of the canons of construction, regardless of political affiliation.").

Table 3: Two-Part Regression Results for Party Affiliation in Tax Court Opinions, 1942 - 2015

	Dependent variable: purposivist terms (per million words)		Dependent variable: textualist terms (per million words)	
Democrat	**-44.6** **(65.8)**	**154.3***** **(54.0)**	**-12.4** **(8.2)**	**-17.7**** **(7.9)**
Year Judge Appointed		3.4 (2.9)		-0.09 (0.36)
Taxpayer Wins		0.1 (42.0)		-4.8 (8.0)
Opinion Year Fixed Effects	No	Yes	No	Yes
Log Pseudo-likelihood	-23,958.91	-9,064.93	-5,818.29	-1,972.50
N	7,308	2,760	7,308	2,479

Note: Each column reflects the combined marginal effects from a two-part regression, excerpted from Tables 10 and 11. The fixed effects rows indicate whether dummy variables are included for each opinion year, each judge authoring opinions, or both. *N* varies between regressions because some observations lacked determinate values for some variables (for example, some cases lack a clear winner, since the taxpayer won on some issues and lost on others). Standard errors are clustered by judge. * denotes statistical significance at p<0.1, ** at p<0.05, and *** at p<0.01.

Table 3 presents the results of regressions that test the effect of party affiliation on the use of purposivist and textualist terms, both alone and will full controls. Without controls, Democratic judges appear less likely to use *both* purposivist and textualist terms, albeit not statistically significantly so. However, as noted above, this could simply be the product of confounding omitted variables. With full controls, Democratic judges are statistically significantly more likely to use purposivist terms (at a 99% confidence level) and statistically significantly less likely to use textualist terms (at a 95% confidence level). Moreover, the magnitude of these effects are large: the mean number of purposivist terms used across the sample is 365.3 per million words, and the mean number of textualist terms is 30.0 per million words, suggesting that party affiliation is an important predictor of interpretive methodology.

G. Case Outcomes Do Not Statistically Significantly Predict Interpretive Methodology at the Tax Court

Scholars have previously studied the determinants of taxpayer wins and losses at the Tax Court, with mixed success.[151] None so far have tested the relationship between interpretive methodology and prevailing party in Tax Court cases. To test this question, I coded the winner in each of the Tax Court cases[152] and included the prevailing party in the regressions analyzing determinants of purposivist and textualist term frequencies.

[151] *See, e.g.*, James Edward Maule, *Instant Replay, Weak Teams, and Disputed Calls: An Empirical Study of Alleged Tax Court Judge Bias*, 66 TENN. L. REV. 351 (1999); Robert M. Howard, *Comparing the Decision Making of Specialized Courts and General Courts: An Exploration of Tax Decisions*, 26 JUST. SYS. J. 135 (2005); Daniel M. Schneider, *Assessing and Predicting Who Wins Federal Tax Trial Decisions*, 37 WAKE FOREST L. REV. 473 (2002). These studies have generally used case outcomes as the dependent variable in regression analysis, attempting to predict case outcomes based on case characteristics. This Article focuses instead on interpretive methodology, using case outcome as one of several independent variables used to attempt to predict methodology.

[152] The coding was conducting algorithmically, exploiting the statement at the end of every Tax Court decision identifying the prevailing party. When a Tax Court case had no clear winner—for example, if the taxpayer prevailed on some issues and the IRS prevailed on others—the case was excluded from the sample. This analysis considers all Tax Court cases from 1942-2015.

Table 4: Two-Part Regression Results for Party Affiliation in Tax Court Opinions, 1942 - 2015

	Dependent variable: purposivist terms (per million words)		Dependent variable: textualist terms (per million words)	
Democrat	154.3*** (54.0)		-17.7** (7.9)	
Year Judge Appointed	3.4 (2.9)		-0.09 (0.36)	
Taxpayer Wins	**0.1** **(42.0)**	**15.0** **(36.2)**	**-4.8** **(8.0)**	**-4.0** **(5.6)**
Opinion Year Fixed Effects	Yes	Yes	Yes	Yes
Judge Fixed Effects	No	Yes	No	Yes
Log Pseudo-likelihood	-9,064.93	-12,983.50	-1,972.50	-2,781.20
N	2,760	4,241	2,479	4,041

Note: Each column reflects the combined marginal effects from a two-part regression, excerpted from Tables 10 and 11. The fixed effects rows indicate whether dummy variables are included for each opinion year, each judge authoring opinions, or both. When judge fixed effects are included, judge characteristics (party and year of appointment) are omitted as multicollinear. *N* varies between regressions because some observations lacked determinate values for some variables (for example, some cases lack a clear winner, since the taxpayer won on some issues and lost on others). Standard errors are clustered by judge. * denotes statistical significance at p<0.1, ** at p<0.05, and *** at p<0.01.

Table 4, again excerpted from the full results in Section E of the Appendix, shows that the relationship between case outcomes and methodology is not statistically significant when controls are included, even at the 90% level.

A reader might wonder whether analysis of case outcomes is meaningful given case selection effects, especially in light of George Priest and Benjamin Klein's famous claim that "the proportion of observed plaintiff victories will tend to remain constant over time regardless of changes in the underlying standards applied."[153] Because the IRS and the taxpayer have the opportunity to settle prior to judgment,[154] the sample of decided cases may be unrepresentative and biased if litigants tend to settle in clear-cut cases. For example, it could be that more textualist judges are more likely to rule against taxpayers, but that, anticipating this, taxpayers and the IRS tend to settle cases before textualist judges (on terms favorable to the IRS), so that the only cases that go to trial have countervailing unobserved characteristics that make them close cases (for example, facts that favor the taxpayer). If so, the model in this Article might fail to capture the true relationship between textualism and taxpayer victories.

But there is some reason to doubt that Tax Court cases follow the Priest-Klein model of rational settlement. For one, Tax Court cases are unique in that the taxpayer need not pay any litigated taxes until the case is resolved[155]—so there are benefits to the taxpayer (liquidity and deferral) in litigating even a losing case to the end. These benefits may not have offsetting costs to the government, which is not subject to liquidity constraints and whose litigators may not receive sufficient credit for settling quickly.[156] In addition, most (more than 80% of[157]) Tax Court cases involve pro se litigants, whose cost of litigation may be lower than those retaining expensive counsel. And because the factual record generally must be assembled in order to respond to the initial IRS audit, Tax Court cases require less additional factfinding than more traditional court cases, again reducing the marginal costs of going to trial. These unusual features may explain why the IRS wins more than 75% of Tax Court cases,[158] contrary to the Priest-Klein hypothesis, which predicts that trial win rates will follow "a strong bias toward . . . fifty percent."[159]

Regardless of whether the Priest-Klein model applies, the failure to find a statistically significant relationship is *not* strong evidence that such a

[153] George Priest & Benjamin Klein, *The Selection of Disputes for Litigation*, J. LEGAL STUD. 13, 31 (1984).

[154] *Taxpayer Information: During Trial*, UNITED STATES TAX COURT, https://www.ustaxcourt.gov/taxpayer_info_during.htm#DURING9 (last visited Oct. 11, 2019) (noting that Tax Court trials may be settled even after the trial is complete).

[155] 1 TAXPAYER ADVOCATE SERVICE, 2018 ANNUAL REPORT TO CONGRESS 295 (2018) ("The U.S. Tax Court is the only prepayment judicial forum for taxpayers to resolve their disputes with the IRS.").

[156] This is essentially a principal-agent problem: even if government litigators receive credit for avoiding litigation costs of trials, they may not receive credit for bringing in tax revenue earlier than if the trial had not occurred.

[157] *Id.* at 295 ("More than 80 percent of cases in Tax Court are brought by unrepresented taxpayers . . . .").

[158] *See infra* tbl. 7.

[159] Priest & Klein, *supra* note 153, at 5, 23.

relationship does not exist, and this Article does not affirmatively claim that interpretive methodology has no relationship with case outcomes. Moreover, even if there is no consistent predictive relationship between case outcomes and interpretive methodology, methodology could still have an important effect on substantive case outcomes. It could be that every time a dictionary is cited, it decisively determines the prevailing party, but the prevailing party is equally likely to be the IRS or the taxpayer. In a well-functioning judicial system, this is in fact desirable—the absence of systemic bias is reassuring rather than a sign that interpretive methodology is superfluous.

IV. Robustness Checks

A. Reading Cases to Confirm Term Frequency Results

To confirm that term frequency results correspond with conventional notions of textualism, purposivism, interpretation, and normativity, I pulled forty Tax Court opinions and manually evaluated how the terms were used in those opinions. Although I spot-checked each search term more informally while producing the list of proxies for each methodology, this Section describes an additional ex post check to ensure the robustness of this Article's methods.

The dataset contained seventy-four years of opinions (1942–2015), which I separated into ten similarly sized time periods. For each methodology, I pulled one opinion at random from each period and reviewed it to confirm that the methodology was used as expected. The full list of these opinions is available online, along with specific details and citations for the methodologies used in each opinion.[160]

B. Bootstrapped Confidence Intervals for Machine Learning

MCC, Accuracy, and F_1 Score generally tell us about the *magnitude* of the differences between courts that can be captured by a machine learning classifier. But an important measure to determine the robustness of these results is whether they are *statistically significant*, that is, whether the classifier performs better than chance.

To this end, I employ a "bootstrapping" design that repeatedly tests the machine learning algorithm on a subsample of the data. By testing how much the estimates of classifier performance vary between tests, we can calculate the standard error of the test and derive confidence intervals. The algorithm is statistically significantly different from zero—that is, its

[160] *Online Appendix: Spot-Checking Terms*, Jonathan H. Choi (last updated Oct. 12, 2019), https://www.jonathanhchoi.com/s/Spot-Checking-Terms-10172019.pdf.

performance is better than random chance—if the confidence interval for MCC excludes zero, and if the intervals for accuracy and F_1 Score exclude 0.5 (50%).

Figures 9 and 10 present confidence intervals for each classifier performance metric after bootstrapping with 1000 test iterations. For each metric, the median value is represented by the white circle, the 95% confidence interval is represented by the blue inner bars, the 99% confidence interval is represented by the green outer bars, and the null hypothesis (the value that would be generated by a classifier performing no better than chance) is represented by the red line.

Figure 9: Bootstrapped Confidence Intervals, Tax Court v. District Courts

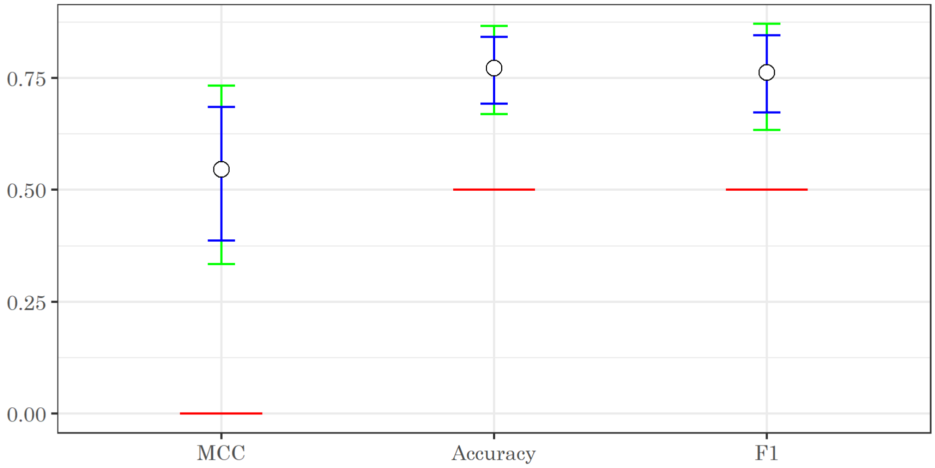
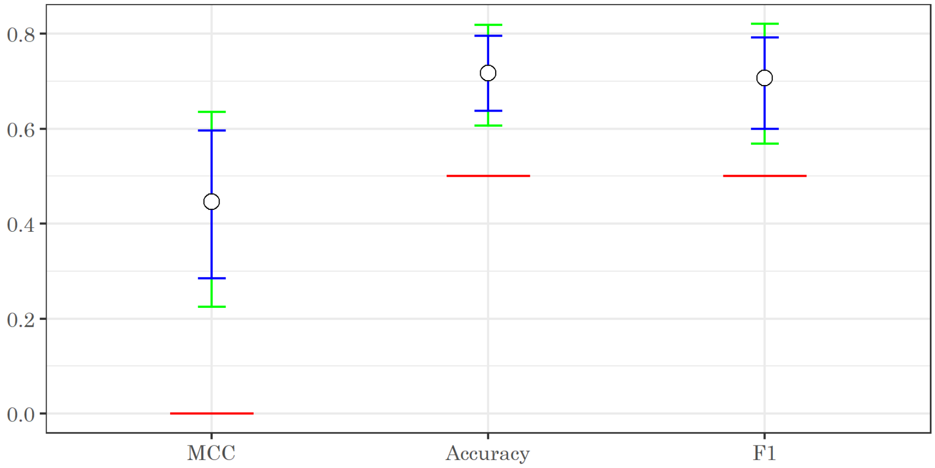


Figure 10: Bootstrapped Confidence Intervals, Tax Court v. CFC



Figures 9 and 10 demonstrate that under each of the performance metrics—MCC, accuracy, and F_1 Score—the classifier performs statistically significantly better than chance at a 99% confidence level, providing additional assurance of the results in Section III.E. Section G of the Appendix contains additional detail on the bootstrapping calculations.

C. Validating OCR Quality over Time

Another potential concern is that apparent trends could be produced merely by variation in the quality of computer OCR over time. This could introduce systematic bias if, for example, older documents were written in text that is more difficult to scan, or the quality of the records degraded over time (due to stains, tears, etc.). If (hypothetically) there were a ten percent chance that any particular word in the 1925 Cumulative Internal Revenue Bulletin were misspelled and therefore not identified, but a zero percent chance in 2018, the matching rate in 2018 would be overstated relative to 1925 by ten percent. A skeptical reader might particularly doubt the dataset of Internal Revenue Bulletins produced specifically for this Article.

I technologically mitigate this issue by using spell-checking to correct obvious errors.[161] But this is not a complete solution—for example, again purely hypothetically, if the OCR rendered the word "the" as "tbo," the spell-checker would not correct that misspelling.[162]

One way to judge the variation in spelling errors over time, which may be a proxy for OCR quality, is to examine the ratio of terms—purposivist, textualist, normative, and interpretive—before and after OCR is conducted. Figure 11 depicts this ratio, obtained for any year by taking the count of all terms examined in this Article in the Cumulative Internal Revenue Bulletin before conducting spellchecking, divided by the count of such terms after conducting spellchecking.[163]

[161] *See supra* Part II.

[162] This is because the spellchecking algorithm used only fixes words that are incorrect by one character (that is, whose Levenshtein distance is one). "Tbo" is different from "the" by two characters—in fact, it would likely be corrected as "to" rather than as "the."

[163] On some occasions, the ratio will exceed 100%. This could happen if, for example, the misspelled word "mcode" were corrected to "mode." In this case, the misspelling could be registered as an instance of "code," which (if used in conjunction with a word like "interpret") would be registered as an interpretive term, while the corrected spelling would not be. For Figure 11, the ratio is capped at 1.00.

Figure 11: Ratio of All Terms Before and After Spellcheck

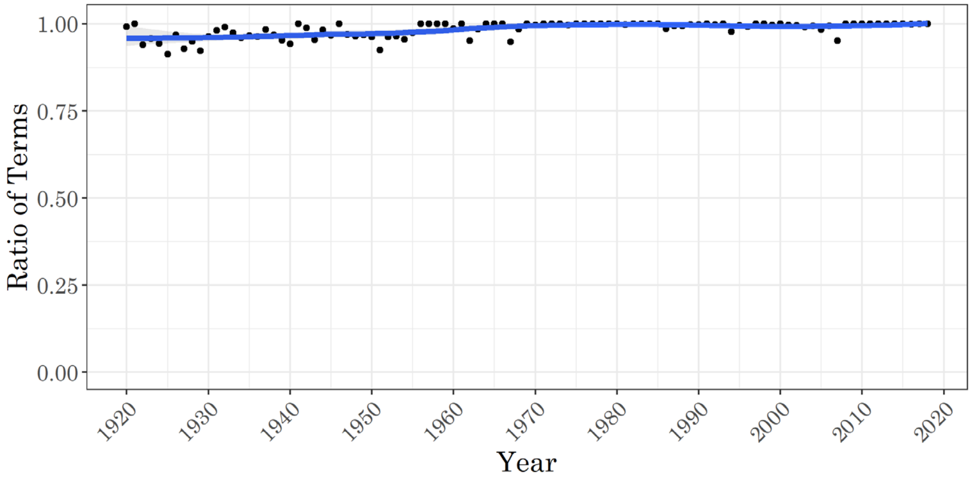


Figure 11 shows that the earliest years contain more misspellings, as might be expected. However, the trendline remains above 95 percent for every year analyzed, and in no year does the ratio of terms before and after spellchecking fall below 93 percent, suggesting that the quantity of misspellings is small (on average less than five percent, for any year). By comparison, the increase in the term frequency of normative terms at the IRS between 1984 and the present is approximately 300 percent.[164] In addition, the misspellings are mostly concentrated in the earliest years. The ratio stabilizes after the late 1960s, with most years afterward at 1.00 (implying that no spelling errors were found for any of the hundreds of occurrences of terms).

Figure 11 therefore suggests that the most significant recent trends analyzed by this Article—especially the increase in normative terms after 1984—are unlikely to be the product of variation in OCR quality. This is also borne out by the fact that a number of the most prominent trends in this Article are declines in term frequency, such as the decline in interpretive terms at the IRS that began in the 1920s.[165] The increase in OCR quality over time suggests that, if anything, these declines might be understated.

[164] *Supra* fig. 4.

[165] *Supra* fig. 4.

Conclusion

Most statutory interpretation occurs at agencies, rather than courts.[166] Yet little empirical scholarship exists on how agencies interpret statutes,[167] and none has contrasted the methodologies of agencies and courts. This has left jurists and scholars in the dark while grappling with some of the most important questions in modern jurisprudence, including the effect of judicial deference doctrines like *Chevron*.

This Article uses the IRS and the Tax Court as case studies in administrative and judicial statutory interpretation. It concludes that the two differ substantially. First, the data show that the IRS less often engages in statutory interpretation at all (especially after *Chevron*), instead shifting toward normative policy judgments in its decisionmaking. Second, the data show that the IRS has become more purposivist over time when interpreting statutes, unlike the now-textualist Tax Court.

This Article also has implications for the study of tax law. It helps taxpayers better to tailor their arguments before the IRS and the Tax Court. Moreover, it provides evidence confirming the "exclusively judicial role" that the Supreme Court has controversially held the Tax Court to play,[168] in that the Tax Court reads statutes more like other courts than like the IRS.

Finally, this Article complicates the standard story of tax exceptionalism. On one hand, the two primary interpreters of federal tax law significantly differ in their methodology, such that tax law is not uniformly more purposivist than other fields, as many scholars have proposed.[169] On the other hand, although the Tax Court has broadly become more textualist, it favors different specific interpretive tools than other courts,[170] suggesting that while certain authorities may be purposivist or textualist in broad terms, each may adopt its own specific flavor of purposivism or textualism.

[166] Mashaw, *supra* note 3, at 502–03 (describing agencies as "the primary official interpreters of federal statutes").

[167] See supra notes 6-7.

[168] Freytag v. Comm'r, 501 U.S. 868, 892 (1991); *see supra* note 60. The judicial status of the Tax Court is important because it implies that Tax Court opinions are subject to de novo review, rather than deferential review, as an agency determination would be. In addition, it has implications for the appointments process at the Tax Court. *See supra* note 60.

[169] *See supra* notes 63–65 and accompanying text.

[170] *Supra* Section III.E.

APPENDIX

A. Data Sources

All of the Python code used in this Article is available for reference online.[171] All of the data used in this Article are available upon request, except for court opinions that I am prohibited from sharing under the terms of my researcher license, as described below.[172]

3. IRS Publications

The IRS publications used in this Article were extracted from two sources. First, I downloaded all of the Cumulative Internal Revenue Bulletins, published annually by the IRS from 1919 until 2008, from the website of the U.S. Government Printing Office.[173] Second, I downloaded all of the Internal Revenue Bulletins posted on the IRS's website, which include the years from 2003 until the present.[174] Both sources provide files in .pdf format, which I converted to plain text using Adobe's OCR software. I found alternative OCR software to produce the same or slightly worse results. The OCR was of reasonably high quality, but to ensure accurate term frequency counts, I also wrote a program to conduct pre-processing (removing whitespace, regularizing capitalization, fixing hyphenation across pages, and conducting spell checking[175]). Where the documents could not be feasibly processed using algorithms, I edited them manually (for example, to remove irrelevant material such as legislation and legislative history). The beginning and ending years, 1919 and 2019, were omitted as partial years that might be biased if IRS guidance follows an annual cycle.

Internal Revenue Bulletins include all official IRS publications for each year—regulations, revenue rulings, revenue procedures, and other miscellaneous statements. They do not include unpublished guidance on which taxpayers (other than the petitioner) are not generally entitled to rely: for example, private letter rulings issued to particular taxpayers that services such as *Tax Notes* may obtain through FOIA requests. The Internal Revenue

[171] *Code*, *supra* note 78.

[172] *See infra* Appendix Section A.2.

[173] U.S. GOV'T PUB. OFF., https://www.govinfo.gov (last visited July 2, 2019).

[174] *IRS Online Bulletins*, INTERNAL REVENUE SERV., https://www.irs.gov/irb (last visited July 2, 2019).

[175] I used the pyspellchecker library in Python, version 0.4.0, with a Levenshtein Distance of 1, after excluding any terms analyzed in this Article. *See Pyspellchecker 0.5.0*, PYTHON SOFTWARE FOUND., https://pypi.org/project/pyspellchecker (last updated July 11, 2019).

Bulletins do contain copies of all tax legislation enacted for the year, along with relevant committee reports.[176] Since the tax legislation and legislative history were not original material produced by the IRS, I removed these from the documents for purposes of this Article.

This Article analyzes regulations and subregulatory guidance together. Historically, the line between different types of guidance has sometimes been fuzzy, and the significance of each type of guidance has changed over time. There was little formal distinction between regulations and subregulatory guidance before the Administrative Procedure Act (APA) was passed in 1946[177] and especially before the Federal Register Act was passed in 1935.[178] Even after the APA, most tax regulations are designated "interpretive" by the Treasury and not promulgated through normal notice-and-comment proceedings, again making them hard to distinguish from subregulatory guidance.[179] Moreover, many of the changes in IRS regulatory practice were endogenous with broader political movements that I am trying to capture in this Article—for instance, the passage of the Administrative Procedure Act was the culmination of years of New Deal politics,[180] the same politics that produced the shift toward purposivism that is a primary finding of this Article.

Given that it would be difficult and perhaps undesirable to disaggregate different types of guidance, I have analyzed all published tax guidance together. The fact that so many of the results discussed in this Article move in opposite directions suggests that this has not biased the

[176] The IRS began to publish committee reports in 1939. Its decision to publish committee reports may contribute to, or may reflect, the IRS's general emphasis on committee reports as indicia of legislative history.

[177] Pub. L. No. 79-404, 60 Stat. 237 (1946).

[178] Pub. L. No. 74-220, 49 Stat. 500 (1935).

[179] See, e.g., Kristin E. Hickman, *Coloring Outside the Lines: Examining Treasury's (Lack of) Compliance with Administrative Procedure Act Rulemaking Requirements*, 82 NOTRE DAME L. REV. 1727, 1729 (2007) ("Treasury also contends, however, that most Treasury regulations are interpretative in character and thus exempt from the public notice and comment requirements by the APA's own terms."). Many critics have alleged that because interpretive regulations did not pass through notice and comment, they lack force of law. See, e.g., Steve R. Johnson, *Intermountain and the Importance of Administrative Law in Tax Law*, TAX NOTES, Aug. 23, 2010, at 837, 843 ("Interpretive regulations do not have force of law; they merely inform the public of what the agency believes the statute means."); Stanley S. Surrey, *The Scope and Effect of Treasury Regulations Under the Income, Estate, and Gift Taxes*, 88 U. PA. L. REV. 556, 557 (1940) (arguing that the APA "does not invest interpretative regulations with the force of law").

[180] George B. Shepherd, *Fierce Compromise: The Administrative Procedure Act Emerges from New Deal Politics*, 90 NW. U. L. REV. 1557, 1560–61 (1996) ("The APA was a cease-fire armistice agreement that ended the New Deal war on terms that favored New Deal proponents.").

results by, for example, inflating the proportion of strictly procedural matters over time. However, more granular analysis of more specific slices of published guidance—such as the "legislative" regulations that do go through conventional notice and comment[181]—would be an interesting project for future research.

4. Court Opinions

The court opinions analyzed in this Article were downloaded from the Caselaw Access Project, a joint project of the Harvard Law School Library and Ravel Law.[182] The Project is an extensive and high-quality database that contains "nearly all cases from an American court" between 1658 and 2015.[183] In order to write this Article, I obtained a researcher license from Caselaw Access Project to download bulk data for the Tax Court and other courts. The terms of the license prohibit sharing bulk data with other researchers, so this is the only dataset used for this Article that I cannot make available upon request.

5. Excluding Non-Substantive Opinions

Past work has generally measured the percentage of judicial opinions containing a particular interpretive tool (say, dictionaries or legislative history) out of the opinions in which some statutory interpretation occurs. The goal is to exclude opinions that are largely procedural, in order to smooth variations in docket composition year over year. To achieve this, these papers have identified a "denominator" of interpretive opinions that divide the number of opinions containing hits for a particular tool.[184]

The Internal Revenue Bulletin contains a few texts in which novel statutory interpretation does not occur—particularly the IRB's reproduction of the past year's legislation and legislative history. I removed these from the

[181] *C.f.*, Hickman, *supra* note 178 (discussing how most, or all, Treasury regulations ought to be considered "legislative" and go through notice and comment).

[182] CASELAW ACCESS PROJECT, https://case.law (last visited Aug. 1, 2019). Ravel Law was subsequently acquired by LexisNexis. Thanks to Mike Lissner, Executive Director of the Free Law Project, for advice on obtaining these data and for providing the court data for early analyses of Tax Court and Supreme Court decisions. *See* COURTLISTENER, https://www.courtlistener.com/api/bulk-info (last visited Aug. 1, 2019).

[183] Jason Tashea, *Caselaw Access Project Gives Free Access to 360 Years of American Court Cases*, A.B.A. J. (Oct. 30, 2018), http://www.abajournal.com/news/article/caselaw_access_project_gives_free_access_to_360_years_of_american_court_cas.

[184] Bruhl, *supra* note 43, at 32-33; Calhoun, *supra* note 46, 495-96.

analysis, which is mathematically equivalent to the denominator approach used in other articles.[185] The issue of procedural opinions does not arise for the Tax Court, since the dataset for this Article includes only Tax Court "division opinions," which address novel legal issues.[186] (Tax Court opinions intended only to speak to settled law are called "memorandum opinions" or "oral opinions," which are unpublished and theoretically lack precedential weight.[187])

B. Terms Analyzed

The terms used in this Article were drawn from prior empirical work,[188] as well as my own reading of relevant sources. All terms are listed in lower case, since the searches I conducted were not case sensitive. All terms were treated as stems for purposes of the counts, meaning that terms with different suffices would also be included. For example, "senate report" below includes "senate reports" as well.

Synonyms for the same concept (for example, "implied repeal" and "implicit repeal") are all listed, for completeness. In order to prevent the machine learning algorithm from overestimating predictive performance based on mere stylistic variation, I group together different terms within a particular category for purposes of the machine learning analysis. For example, the number of citations to Senate reports are aggregated, regardless of whether they are written as "S. Rep.", "S. Rpt.", or "Senate report." Without these groupings, the algorithm might demonstrate a perfect ability to

[185] To illustrate, say that a sample of documents has 150 documents overall, 50 that cite dictionaries and 100 that engage in any statutory interpretation. The denominator approach divides the 50 citing dictionaries by the 100 (the denominator) citing statutory interpretation, yielding 50%. My approach, which is computationally simpler, divides 50 by the 150 minus 50, also yielding 50%.

[186] 2 HAROLD DUBROFF & BRANT J. HELLWIG, THE UNITED STATES TAX COURT: AN HISTORICAL ANALYSIS 750 (2014); Grewal, *supra* note 62, at 2073–79; *see* I.R.C. §§ 7459–60 (2012) (describing the process to issue division opinions).

[187] In practice, memorandum opinions are often cited and relied upon, and they do sometimes contain original legal decisionmaking. *See* DUBROFF & HELLWIG, *supra* note 185, at 750; Grewal, *supra* note 62, at 2073–81. However, the point remains that Tax Court division opinions are all intended to contain novel legal interpretation, and it is reassuring that the category is if anything underinclusive.

[188] *See* Bruhl, *supra* note 43, at 30-31, 38-39, 41, 53 (listing and describing the use of search terms to assess judicial purposivism, textualism, and canon use); Calhoun, *supra* note 46, at 524–25 (listing dictionaries); Staudt et al., *supra* note 44, at 1933-35, 1940-42, 1950-51, 1956-59 (listing terms associated with textualism, purposivism, judicial deference, and canons of construction). I thank Aaron Bruhl for sharing the search terms that he used in his comparative study of judicial statutory interpretation.

distinguish one court from another merely based on differences in citation practices.

One hazard attending machine learning is that conducting classification on an entire corpus of text—considering every word in a series of documents and testing whether each word has any predictive value—can produce seemingly strong predictive relationships purely by chance. This practice, known as "data dredging," is a perennial risk when machine learning is used for social science research.[189] To avoid it, I constrain the vocabulary of words that the classifier may consider in the learning process to the interpretive terms set out in this Section of the Appendix. Importantly, the interpretive vocabulary was selected based on my ex ante views on interpretive methodology and draws heavily on the existing vocabularies selected by other authors,[190] rather than being selected ex post based on which terms had predictive value after running a machine learning algorithm. In doing so, I reduce the risk that the classifier may appear to successfully predict a result merely by chance.

1. Purposivist Terms

Congressional Reports

conference report
conf. rep.
conf. rpt.
conf. rept.
conf.rep.
conf.rpt.
conf.rept.
house report
h. rep.
h. rpt.
h. rept.
h.rep.
h.rpt.
h.rept.
h.r. rep.
h.r. rept.
h. r. rept.
h.r.rep.
h.r.rpt.
h.r.rept.
senate report
s. rep.
s. rpt.
s. rept.
s.rep.
s.rpt.
s.rept.
committee report
comm. rep.
comm. rpt.

[189] Gregg R. Murray & Anthony Scime, *Data Mining*, *in* Emerging Trends in the Social and Behavioral Sciences 1, 3–4 (Robert Scott & Stephen Kosslyn eds., 2015)).

[190] In contrast, the normative terms were selected specifically for this article.

h. r. rep.
h.r. rpt.
h. r. rpt.
comm. rept.
comm.rep.
comm.rpt.
comm.rept.

Congressional Hearings

congressional hearing
congressional record
cong. rec.
cong.rec.
rec. doc.
committee hearing
senate hearing
house hearing
conference hearing

Miscellaneous Legislative History

legislative history
history of the legislation
conference committee
joint committee
jct
congressional budget office
cbo
senate committee
s. comm.
s. subcomm.
house committee
h.r. comm.
h. subcomm.
h. r. subcomm.

2. *Textualist Terms*

Some potential synonyms for "plain meaning" were excluded, on the basis that courts have not always used them in a textualist manner. For example, the "literal meaning" of a statute[191] is often cited as a criticism of textualism rather than an endorsement of it.[192] Accordingly, I excluded that term in order to avoid false positives.

[191] *Cf.* Stephen C. Mouritsen, *The Dictionary Is Not a Fortress: Definitional Fallacies and a Corpus-Based Approach to Plain Meaning*, 10 BYU L. REV. 1915, 1973 (2010) (analyzing the terms "plain meaning," "ordinary meaning," "natural meaning," "literal meaning," and "common meaning").

[192] *See, e.g.*, U.S. Padding Corp. v. Comm'r, 88 T.C. 177, 184 (1987) ("We may then look to the reason of the enactment and inquire into its antecedent history and give it effect in accordance with its decision and purpose, sacrificing, if necessary, the literal meaning in order that the purpose may not fail.").

Dictionaries[193]

dictionary[194]
dictionarium
linguae britannicae
world book
funk & wagnalls

Linguistic Canons[195]

expressio[196]
expresio
inclusio[199]
noscitur a sociis[200]
ejusdem generis[197]
last antecedent[198]
plain meaning

Textual-Holistic Canons

whole act
whole-act
meaningful variation
consistent usage

[193] These terms were borrowed in part from John Calhoun's listing. *See* Calhoun, *supra* note 46, app. I.

[194] Occurrences of the word "dictionary" in "dictionary act" are excluded.

[195] *See* Bruhl, *supra* note 43, at 56 ("The category of linguistic canons is composed of four familiar rules of word association and grammar: ejusdem generis, noscitur a sociis, expressio unius, and the rule of the last antecedent. All of these linguistic canons can be captured with good accuracy through electronic searches.").

[196] This phrase and its variants refer to the Latin maxim that *expressio unius est exclusio alterius*, meaning that express listing of certain items in a statute is presumed to exclude any unmentioned comparable items. Chevron USA Inc. v. Echazabal, 536 U.S. 73, 80 (2002) (quoting United States v. Vonn, 535 U.S. 55, 65 (2002)) ("[E]xpressing one item of [an] associated group or series excludes another left unmentioned.").

[197] *See supra* notes 141–142 and accompanying text.

[198] Jacob Scott, *Codified Canons and the Common Law of Interpretation*, 98 Geo. L.J. 341, 358 (2010) ("The last antecedent rule is somewhat confusing and hypergrammarian; it limits the operation of qualifying phrases to the last phrase in a sentence (rather than applying that limitation to the entire sentence).").

[199] "*Inclusio unius*" is a relatively rare variant whose effect is identical to the *expressio unius* canon. *See LawProse Lesson #227: Part 2: "Including but Not Limited to,"* LawProse, www.lawprose.org/lawprose-lesson-227-part-2-including-but-not-limited-to (last visited Aug. 1, 2019) ("In legal literature, *expressio unius* is more than 15 times as common as *inclusio unius*.").

[200] *See* Staudt et al., *supra* note 187, at 1933 ("[T]he meaning of one term is 'known by its associates' (i.e., understood in the context of other words in the list).").

whole code
whole-code
in pari materia[202]
surplusage[201]
superfluity
superfluities

3. *Interpretive Terms*

Unlike the other terms in this Article, a document's interpretive score was determined based on the number of interpretive *sentences*. A sentence was designated as interpretive if it included at least one word from the column on the left below, and one word from the column on the right below.

Includes:	AND	Includes:
construe		*statute*
construing		*statutory*
construction		*legislation*
interpret		*congress*
reading		*code*
		section

In addition, the following terms were included in the vocabulary used for machine learning analysis.

plain language
legislative intent
statutory purpose
vagueness
vague
ambiguity
ambiguities
ambiguous
unambiguous

4. *Normative Terms*

As noted above,[203] the phrase "effective tax administration" is excluded from counts using the following terms, as are the phrases "treasury inspector general for tax administration" and "small business regulatory

[201] *See, e.g.*, Corley v. United States, 556 U.S. 303, 314 (2009) (quoting Hibbs v. Winn, 542 U.S. 88, 101 (2004)) ("[O]ne of the most basic interpretive canons is that a statute should be construed so that effect is given to all its provisions, so that no part will be inoperative or superfluous, void or insignificant.").

[202] *See supra* notes 138–140 and accompanying text.

[203] *See supra* note 80 and accompanying text.

enforcement fairness act." In addition, any occurrences of normative terms in sentences that also contained purposivist terms, textualist terms, or substantive canons were excluded, in order to avoid policy judgments that occur in the interpretive process (for example, legislative history that discusses fairness).

good public policy
public policy goal
public policy grounds
tax administration
efficient administration
efficient tax collection
efficient enforcement
compliance burden
financial burden
administrative burden
regulatory burden
burdensome
compliance cost
complexity
intrusive
fairness
unfair
injustice
unjust
clarity

5. *Substantive Canons*

Deference regimes, such as *Chevron* and *Skidmore*, have sometimes been considered precedents and sometimes considered canons of construction.[204] I classify them as substantive canons for purposes of this Article but do not otherwise take a position on which categorization is more accurate.

General Substantive Canons

charming betsy
rule of lenity
absurd result
avoidance canon
canon of avoidance
constitutional avoidance
repeal by implication
implied repeal
implicit repeal
implicitly repeal
presumption against preemption
presumption against pre-emption

[204] *See* Raso & Eskridge, *supra* note 40, at 1727 ("As a descriptive matter, we find that deference regimes are more like canons of statutory construction, applied episodically but reflecting deeper judicial commitments, than like binding precedents, faithfully applied, distinguished, or overruled.").

Deference Canons

chevron
skidmore
seminole rock
auer

C. Non-Normal Distribution of Term Frequencies in Tax Court Opinions

Term frequencies in Tax Court opinions have several important distributional features that demand special attention in statistical analysis (including machine learning analysis). First, they are *semicontinuous*[205]: they vary continuously (they are not limited to whole numbers) but cannot be less than zero (since no opinion can use any term less than zero times). Second, they are *zero-inflated*[206]: many courts use *no* terms of any particular type, such that the median number of purposivist, textualist, interpretive, and normative terms used in Tax Court opinions is zero in each case. Third, they are *log-normal*: even excluding zero values, the distributions exponentially decrease, with long right tails (i.e., most cases use few terms, but some cases use a large number of terms), requiring log-transformation to turn them into normal distributions.

Each of these features violates the conventional assumption of normal distribution that underlies conventional statistical analysis, including standard ordinary least squares (OLS) regression and machine learning on raw term frequencies. Log-normality also casts doubt on visual analysis of the term frequency charts in this Article. There is a risk that any analysis of data following a log-normal distribution will be driven by outliers and therefore be less robust. Consequently, this dataset requires additional transformation to confirm the robustness of the results in this Article and should not be interpreted using OLS regression or raw term frequencies alone.

Table 5 illustrates the problem of zero-inflation in the data:

[205] *See id.* at ^#–^#.

[206] *See* Yongyi Min & Alan Agresti, *Modeling Nonnegative Data with Clumping at Zero: A Survey*, 1 JIRSS 7, 7 (2002) (“Applications in which data take nonnegative values but have a substantial proportion of values at zero occur in many disciplines. The modeling of such ‘clumped-at-zero’ or ‘zero-inflated’ data is challenging.”).

Table 5: Percentage of Tax Court Opinions with Zero Terms, 1942 - 2015

Type of Term	
Purposivist	69.89%
Textualist	93.46%
Interpretive	70.31%
Normative	88.27%

Figures 12 through 15 illustrate all three issues: semicontinuity, zero-inflation, and the log-normal distribution:

Figure 12: Histogram of Purposivist Terms in Tax Court Opinions, 1942-2015

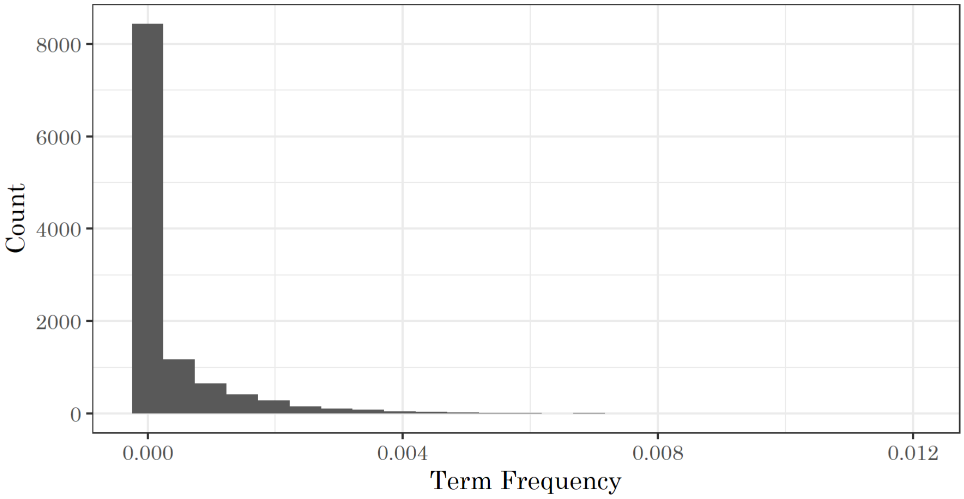


Figure 13: Histogram of Textualist Terms in Tax Court Opinions, 1942-2015

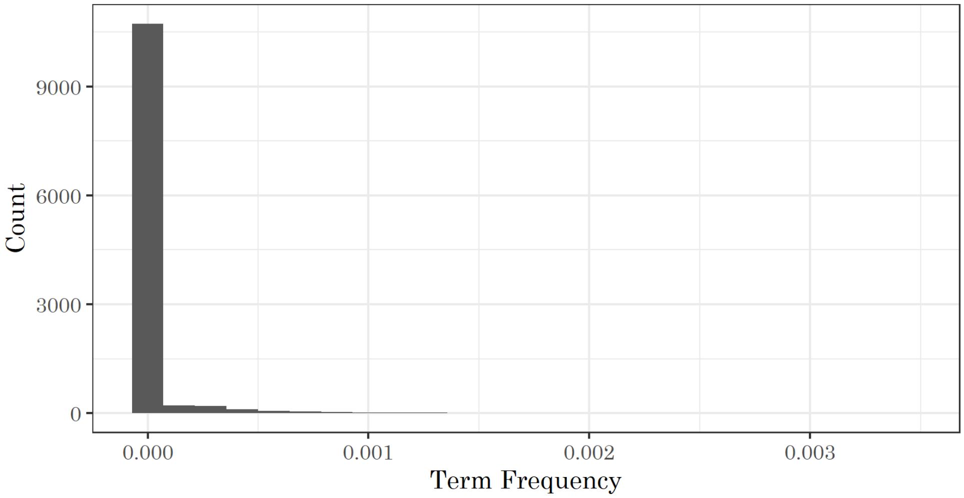


Figure 14: Histogram of Interpretive Terms in Tax Court Opinions, 1942-2015

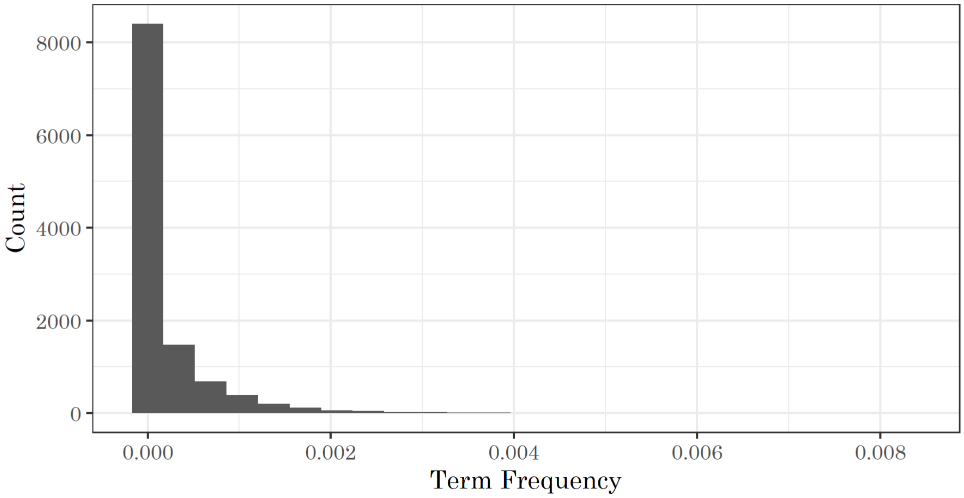
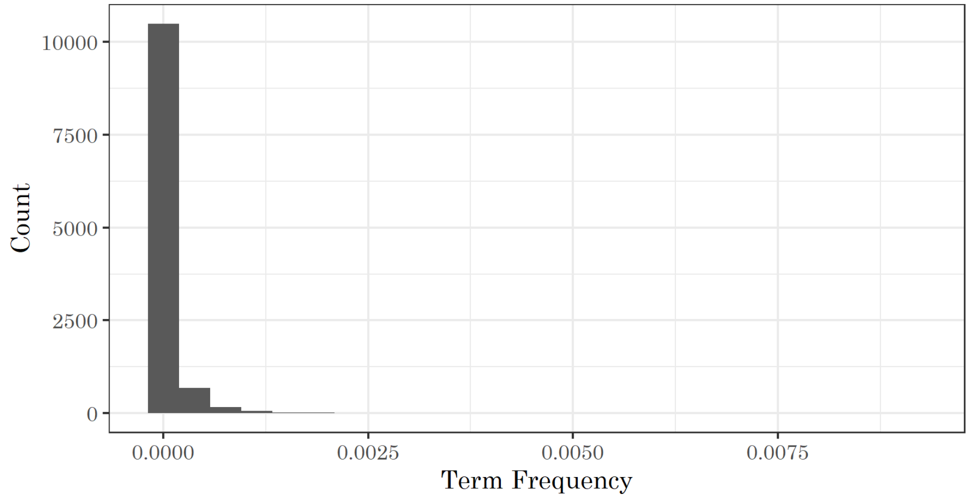


Figure 15: Histogram of Normative Terms in Tax Court Opinions, 1942-2015



Fortunately, a log-normal distribution can be easily addressed by logarithmically transforming the data. One method is to log-transform the data as follows:

$$\tilde{y} = \log(1 + y) \tag{1}$$

When the data are log-transformed in this way, they take the shape of the normal distribution (when excluding zeros—zero-inflation is a separate problem that I address in Section E.2 of the Appendix). Section E reproduces each term frequency chart in this Article subject to the log-transformation. Figures 16 through 19 illustrate that the log-transformation produces approximately normal distributions.

Figure 16: Log-Transformed Histogram of Purposivist Terms in Tax Court Opinions, 1942-2015

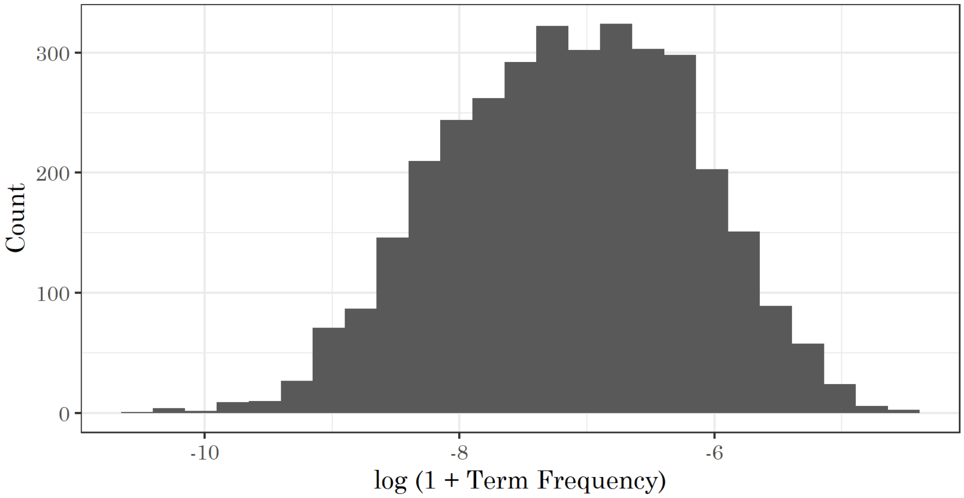


Figure 17: Log-Transformed Histogram of Textualist Terms in Tax Court Opinions, 1942-2015

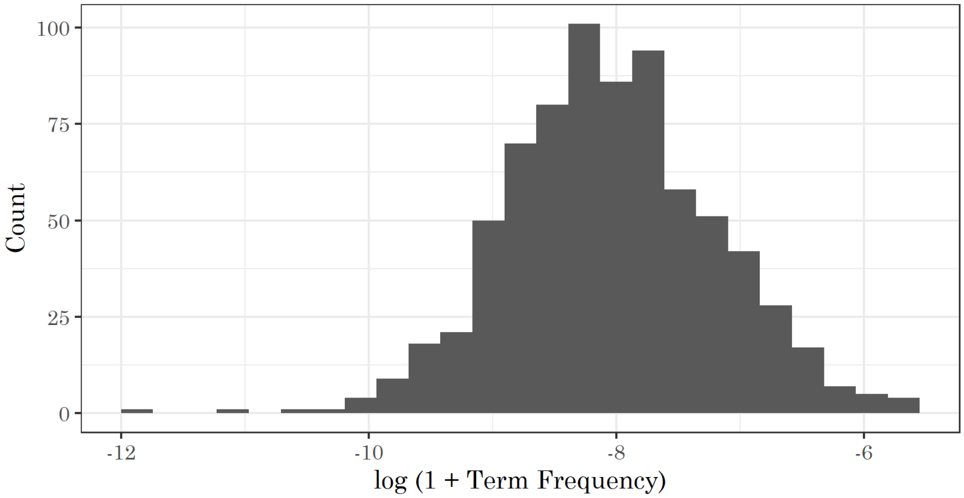


Figure 18: Log-Transformed Histogram of Interpretive Terms in Tax Court Opinions, 1942-2015

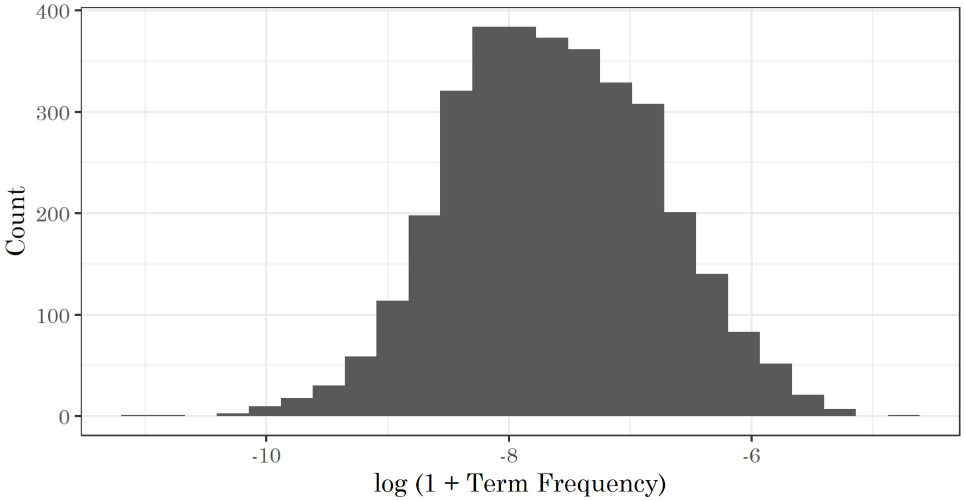
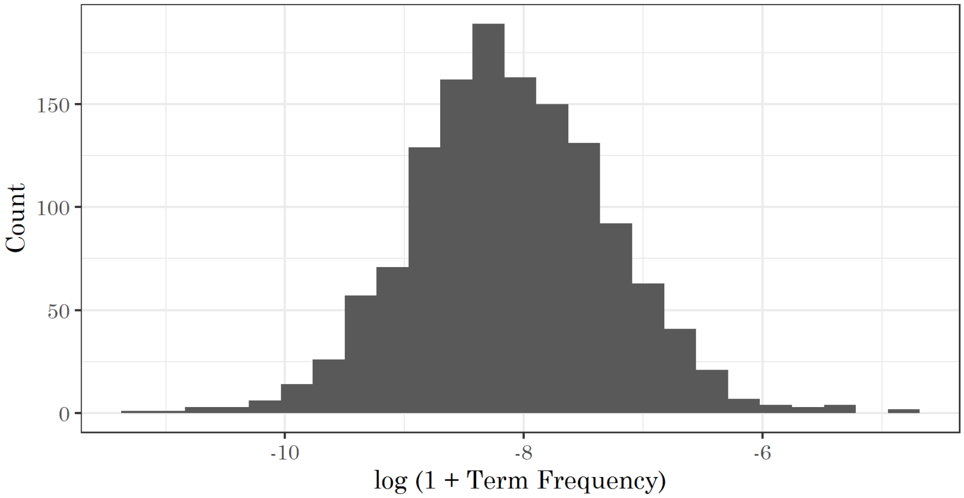


Figure 19: Log-Transformed Histogram of Normative Terms in Tax Court Opinions, 1942-2015



From the log-transformed histograms, it is evident that the distribution of data points is approximately log-normal when considering opinions with more than zero terms. This confirms that the data can be described as semicontinuous, with zero-inflation and a log-normal distribution.[207]

[207] I was not able to separate the dataset of IRS publications cleanly into discrete individual publications (which in any case are much more heterogeneous than court opinions;

This Article employs three methodologies, each of which must appropriately account for these distributional features. To ensure that the charts presented above are valid, Section F of the Appendix presents log-transformed versions of each of them. To ensure that the machine learning methodology is valid, Section D of the Appendix describes how the transformer used in the machine learning analysis normalizes the data prior to the operation of the classifier. Finally, to ensure that regression analysis is valid, Section E.2 of the Appendix employs a two-part regression model specifically designed to address semicontinuity, zero-inflation, and log-normality, which are common issues in natural datasets.

D. Tf-idf Transformation and Classification in Machine Learning

This Section provides additional detail on the methodology used for the machine-learning analysis in this Article, especially in light of the log-normal distribution of term frequencies discussed in the previous Section. Section II.B discussed how Tax Court opinions are first vectorized by obtaining term frequencies for each term of interest, and ultimately classified by an algorithm (in this case, a logistic regression) that improves by repeatedly iterating over a training set.

Between vectorizing and classification, however, the term frequencies are also *transformed* in order to make the classification statistically valid. The transformation converts raw term frequency to term frequency-inverse document frequency (tf-idf) and normalizes the data in the process. Mathematically, given term frequency $tf_{t,d}$ with respect to term t and document d, term frequency is log-transformed so that:

$$\widetilde{tf}_{t,d} = log(1 + tf_{t,d}) \tag{2}$$

Notice that this log-transformation is the same one used in Sections D and G of the Appendix. Next, inverse document frequency is calculated as a function of N, the number of documents in the corpus, and df_t, the number of documents in the corpus for which $tf_{t,d} > 0$:

$$\widetilde{idf}_t = log(N/\, df_t) \tag{3}$$

many IRS publications are merely administrative and only a few lines long). Consequently, I could not conduct the histogram analysis above for IRS publications. However, since it is plausible that IRS publications would also follow the same problematic distribution—anecdotally, I noticed outliers while cleaning the dataset, where the IRS heavily utilized certain interpretive tools in explaining particularly knotty guidance—out of caution, I apply the same logarithmic corrections in Section F of the Appendix for IRS publications as I do for Tax Court opinions.

Finally, tf-idf is calculated as a function of log-transformed term frequency and inverse document frequency:

$$tfidf_{t,d} = \widetilde{tf}_{t,d} \cdot \widetilde{idf}_t \qquad (4)$$

Conceptually, the use of tf-idf rather than raw tf reflects that terms before more indicative of classification when they are rarer. The inclusion of log-transformation in the tf-idf transformation also addresses the log-normality of the term frequency distribution.

Because the tf-idf statistic is then used in a classifier modeled as a logistic regression, the inclusion of idf rather than raw tf merely multiplies each coefficient in the regression by a scalar and therefore does not affect statistics such as MCC, accuracy, or F_1 score, nor does it affect the statistical significance of each term. This can be seen by considering the regression that the classifier conducts, where p / (1 - p) is the odds ratio with respect to the classification category (e.g., a Tax Court opinion), and *n* is the number of terms.

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \cdot tfidf_{1,d} + \beta_2 \cdot tfidf_{2,d} + \cdots + \beta_n \cdot tfidf_{n,d} + \epsilon_d \qquad (5)$$

Through Equations 4 and 5, we find that:

$$\beta_t \cdot tfidf_{t,d} = \beta_t \cdot \widetilde{tf}_{t,d} \cdot \widetilde{idf}_t \qquad (6)$$

$\widetilde{idf}_t$ varies with respect to each term and not with respect to each document. This means that it is a scalar multiplier against coefficient β_t. In other words, the relationship between β_t in this regression and $\widehat{\beta_t}$ from a different regression conducted only on log-transformed term frequency is that:

$$\widehat{\beta_t} = \beta_t \cdot \widetilde{idf}_t \qquad (7)$$

E. Regression Analysis of Tax Court Opinions

This Section employs regressions to more closely analyze the relationship between interpretive methodology in Tax Court opinions, on the one hand, and either case outcomes or party affiliation of judges, on the other hand.[208]

All of the regressions in this Section use clustered standard errors with clustering by judge, a variant of robust standard errors that accounts for heteroskedasticity across the "clusters" of opinions written by different judges. The regressions take each Tax Court opinion as a single observation, using term frequency (either purposivist or textualist) as the dependent variable, and taking as the independent variables: (1) party affiliation of the judge authoring the opinion, (2) case outcome, (3) the year that the judge writing the opinion was appointed, (4) fixed effects for the year the opinion was written, and/or (5) fixed effects for the judge that wrote the opinion. In regressions where judge fixed effects are used, party affiliation and the judge's year of appointment are dropped as multicollinear with the fixed effects. Fixed effects introduce dummy variables for each year or judge, which controls for variation in methodology over time and between judges,[209] isolating differences within a particular year and within a particular judge's docket.

Table 6 presents summary statistics for Tax Court opinions, to facilitate interpretation of the regression results in this Section.

[208] *See supra* Section III.F, III.G.

[209] *See generally* PAUL D. ALLISON, FIXED EFFECTS REGRESSION MODELS (2009) (describing fixed effects regression models).

Table 6: Summary Statistics for Tax Court Opinions, 1942 – 2015

	N	Minimum	Mean	Median	Maximum	Standard Deviation
Democrat[210]	7,308	0	0.561	1	1	0.496
Taxpayer Wins[211]	4,261	0	0.224	0	1	0.417
Textualist Term Frequency[212]	11,451	0	30.0	0	3,427.6	162.5
Purposivist Term Frequency[213]	11,451	0	365.3	0	11,869.4	967.3

It should be noted that regressions measure fundamentally different things from classifier accuracy results. Here, classifier accuracy measures how well opinions can be categorized into one of two categories based on interpretive methodology alone. This is roughly analogous to a regression with a binary category dummy (say, Democratic or Republican) as the dependent variable and each specific interpretive term (say, "dictionary") as the independent variables. (This description glosses over some additional nuance, of course, such as the transformation discussed in the previous Section and the fact that only some classifier techniques are analogous to regression.[214]) Classifier accuracy therefore measures to what extent methodology alone can explain the variation between the two categories.

In contrast, the regression analysis in this Section more narrowly asks whether specific variables have a statistically significant relationship with methodology. The regressions do not analyze a vector consisting of many different interpretive tools, but rather a single summary statistic reflecting the term frequency of all textualist or purposivist tools, respectively, in each opinion. Most importantly, the experimental hypothesis is completely different. Tests of statistical significance in regression analysis ask only

[210] This variable equals 1 if the judge authoring an opinion is a Democrat and 0 otherwise. Thus, this row indicates that 56.1% of opinions were authored by Democrats.

[211] This variable equals 1 if the taxpayer won and 0 otherwise. Thus, this row indicates that the taxpayer won in 22.4% of cases.

[212] Terms per million words.

[213] Terms per million words.

[214] *See supra* notes 85–89 and accompanying text.

whether a particular variable has *any* effect (i.e., whether we can reject the null hypothesis that the variable has no effect); classifier accuracy gauges the *magnitude* of the effect, by asking how much that variable drives outcomes. Classifier accuracy is therefore a cousin of R^2, the measure of what portion of variation in the dependent variable can be explained by all of the independent variables. A result might both be very statistically significant but still have a low R^2.

1. *Ordinary Least Squares (OLS) Regression Model*

Because term frequencies are not normally distributed, as described in Section C of the Appendix, OLS is not an appropriate regression model for these data. Nevertheless, I present OLS results for comparison with the results of the two-part regression model.[215] For document *d*, year *y*, and judge *j*, the models for each regression in Tables 8 and 9 are (in order, left to right):

$$tf_d = \beta_0 + \beta_1 \cdot Democrat_d + \epsilon_d \tag{8}$$

$$tf_d = \beta_0 + \beta_1 \cdot Year\ Judge\ Appointed_d + \epsilon_d \tag{9}$$

$$tf_d = \beta_0 + \beta_1 \cdot Taxpayer\ Wins_d + \epsilon_d \tag{10}$$

$$tf_d = \beta_0 + \beta_1 \cdot Democrat_d + \beta_2 \cdot Year\ Judge\ Appointed_d + \beta_3 \cdot Taxpayer\ Wins_d + \beta_{4,y} \cdot Year_{d,y} + \epsilon_d \tag{11}$$

$$tf_d = \beta_0 + \beta_1 \cdot Taxpayer\ Wins_d + \beta_{2,y} \cdot Year_{d,y} + \beta_{3,j} \cdot Judge_{d,j} + \epsilon_d \tag{12}$$

[215] I used ordinary least squares regression in Stata, version 16, using robust variance estimates. STATA, ROBUST VARIANCE ESTIMATES, https://www.stata.com/manuals13/p_robust.pdf (last visited Aug. 1, 2019).

Table 7: OLS Regression Results for Tax Court Purposivism, 1942 - 2015

Dependent variable: purposivist terms (per million words)

Democrat	-44.6			159.7**	
	(80.1)			(64.6)	
Year Judge Appointed		12.42***		2.8	
		(1.06)		(3.6)	
Taxpayer Wins			81.9*	-13.3	6.6
			(46.26)	(50.0)	(39.8)
Opinion Year Fixed Effects	No	No	No	Yes	Yes
Judge Fixed Effects	No	No	No	No	Yes
R^2	0.0006	0.0535	0.0012	0.1348	0.1540
N	7,308	11,451	4,261	2,763	4,255

Note: Each column of this table represents the results of a separate regression. The dependent variable in each regression is the frequency of textualist terms, in words per million. The fixed effects rows indicate whether dummy variables are included for each opinion year, each judge authoring opinions, or both. When judge fixed effects are included, judge characteristics (party and year of appointment) are omitted as multicollinear. *N* varies between regressions because some observations lacked determinate values for some variables (for example, some cases lack a clear winner, since the taxpayer won on some issues and lost on others). Standard errors are clustered by judge. * denotes statistical significance at p<0.1, ** at p<0.05, and *** at p<0.01.

Table 8: OLS Regression Results for Tax Court Textualism, 1942 - 2015

Dependent variable: textualist terms (per million words)

Democrat	-12.6			-7.3	
	(8.4)			(6.8)	
Year Judge Appointed		1.15***		-0.20	
		(0.19)		(0.53)	
Taxpayer Wins			1.2	-3.4	-3.2
			(5.9)	(6.1)	(4.7)
Opinion Year Fixed Effects	No	No	No	Yes	Yes
Judge Fixed Effects	No	No	No	No	Yes
R^2	0.0013	0.0130	0.0000	0.0623	0.0994
N	7,308	11,451	4,261	2,763	4,255

Note: Each column of this table represents the results of a separate regression. The dependent variable in each regression is the frequency of textualist terms, in words per million. The fixed effects rows indicate whether dummy variables are included for each opinion year, each judge authoring opinions, or both. When judge fixed effects are included, judge characteristics (party and year of appointment) are omitted as multicollinear. *N* varies between regressions because some observations lacked determinate values for some variables (for example, some cases lack a clear winner, since the taxpayer won on some issues and lost on others). Standard errors are clustered by judge. * denotes statistical significance at p<0.1, ** at p<0.05, and *** at p<0.01.

2. *Two-Part Regression Model*

Although OLS regression may be useful to set a baseline, it is a poor fit for the Tax Court data analyzed in this Article. As described in Section C of the Appendix, any regression method must specially adjust for the fact that term frequencies in this dataset are semicontinuous,[216] zero-inflated,[217] and log-normal. Each of these features violate the assumption of normal distribution that underlies OLS regression.

However, these features frequently appear in natural datasets, and econometricians have developed alternative regression methods to address them.[218] In this Section, I will use the two-part regression model first developed by Naihua Duan et al.[219] and implemented by Federico Belotti et al.[220] Conceptually, the model is divided between a first part to determine whether the dependent variable has a zero or positive value, and a second part to determine the positive value, conditional on the value being positive. This models a situation where a judge makes an initial decision on whether to use any textualist terms, and if she does so, a second decision on how many textualist terms to use.

Mathematically (and assuming a single independent variable for simplicity), and for our purposes applying a logistic regression, the first step may be represented as[221]:

$$logit[P(Y_i = 0)] = x'_{1i}\beta_1 + \epsilon_i \tag{13}$$

The second step is a regression on the value of y_i conditional on y_i being positive, for our purposes assuming a log-normal distribution[222]:

$$log[y_i | y_i > 0] = x'_{2i}\beta_2 + \epsilon_i \tag{14}$$

[216] *Id.* at 7-9.

[217] Min & Agresti, *supra* note 206, at 7-9.

[218] *See generally* J.A. Cole & J.D.F. Sherriff, *Some Single- and Multi-Site Models of Rainfall Within Discrete Time Increments*, 17 J. HYDROLOGY 97 (1972) (applying an early version of a two-part regression model to estimate rainfall); Naihua Duan et al., *Choosing between the Sample-Selection Model and the Multi-Part Model*, 2 J. BUS. & ECON STATS. 283 (1984) (applying a two-part model to estimate healthcare expenditures).

[219] Duan et al., *supra* note 218.

[220] *See* Federico Belotti et al., *Twopm: Two-Part Models*, 15 STATA J. 3 (2015).

[221] Min & Agresti, *supra* note 206, at 11. This particular example assumes that a logit model is used for the first part, which is the model I use in this Article. A probit model may also be used but would not have been appropriate for these data.

[222] *Id.* at 11.

The model separately estimates the marginal effect of each independent variable with respect both to the first part and the second part. But the two parts can also be combined to estimate the overall marginal effect of each independent variable with respect to the dependent variable. That is, the combined marginal effect of x_i both in changing the likelihood that y_i will be positive, as well as the marginal predictive effect of x_i on y_i assuming that y_i is positive. Mathematically, this is represented as[223]:

$$y_i = \hat{y}_i |\, x_i = (\hat{p}_i | x_i) \times (\hat{y}_i | y_i > 0, x_i) \tag{15}$$

Equations 8, 11, and 12 are modified in order to reflect Equations 13 through 15. Results from the two-part regression are presented in Tables 11 and 12. Each table contains three regressions, separated between the first part, second part, and combined marginal effect. Note that the coefficients in each column represent the results of very different regressions and are not directly comparable except in sign.

N changes between the tables, even for regressions with the same dependent and independent variables, because the first part of the regression drops any observations if the zero-positive dichotomy can be perfectly predicted based on any independent variable, including a dummy variable—for example, if any judge never uses a textualist term, or if no textualist terms were used in any opinion for a given year.

The first-step regression, as noted above, is a logit model. The second-step regression is a generalized linear model (GLM), which is a generalization of the OLS model with some assumptions relaxed. Specifically, I use a log-linked gamma GLM, in order to account for the distribution of term frequencies.[224] The coefficients from the first and second parts are retransformed in order to calculate combined marginal effects on a raw scale, because they are both calculated on non-linear scales.

One interesting supplemental finding to those in Section III.G is that the first-part coefficients for case outcomes are positive, but the second-part coefficients are negative. This suggests that cases where the taxpayer wins are more likely to include at least one purposivist or textualist term, but cases where the taxpayer loses are more likely to include more than one (conditional on including at least one). This result is not statistically significant but possibly warrants additional research.

[223] Belotti et al., *supra* note 220, at 7.

[224] For an example of this model in a two-part regression, see Belotti et al., *supra* note 220, at 10-13.

Table 9: Two-Part Regression Results for Tax Court Purposivism, 1942 - 2015

Dependent variable: purposivist terms (per million words)

	Logit (1st)	GLM (2nd)	Combined	Logit (1st)	GLM (2nd)	Combined	Logit (1st)	GLM (2nd)	Combined
Democrat	-0.407* (0.212)	0.161* (0.091)	-44.6 (65.8)	0.127 (0.108)	0.272*** (0.091)	154.3*** (54.0)			
Year Judge Appointed				0.0125 (0.0088)	0.0008 (0.0043)	3.4 (2.9)			
Taxpayer Wins				0.096 (0.113)	-0.05 (0.070)	0.1 (42.0)	0.153 (0.095)	-0.049 (0.067)	15.0 (36.2)
Opinion Year Fixed Effects	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Judge Fixed Effects	No	No	No	No	No	No	Yes	Yes	Yes
Log Pseudo-likelihood	-4,586.78	-19,372.13	-23,958.91	-1,526.21	-7,538.72	-9,064.93	-2,238.31	-10,745.19	-12,983.50
N	7,308	7,308	7,308	2,760	2,760	2,760	4,241	4,241	4,241

Note: Each “Logit” and “GLM” column reflects the first and second part of a two-part regression, with the “Combined” column reflecting the marginal effect calculated by combining both columns. The fixed effects rows indicate whether dummy variables are included for each opinion year, each judge authoring opinions, or both. When judge fixed effects are included, judge characteristics (party and year of appointment) are omitted as multicollinear. *N* varies between regressions because some observations lacked determinate values for some variables (for example, some cases lack a clear winner, since the taxpayer won on some issues and lost on others). Standard errors are clustered by judge. * denotes statistical significance at p<0.1, ** at p<0.05, and *** at p<0.01.

Table 10: Two-Part Regression Results for Tax Court Textualism, 1942 - 2015

Dependent variable: textualist terms (per million words)

	Logit (1st)	GLM (2nd)	Combined	Logit (1st)	GLM (2nd)	Combined	Logit (1st)	GLM (2nd)	Combined
Democrat	-0.555** (0.224)	0.146 (0.115)	-12.4 (8.2)	-0.14 (0.16)	-0.35** (0.15)	-17.7** (7.9)			
Year Judge Appointed				0.0014 (0.0074)	-0.0037 (0.0071)	-0.09 (0.36)			
Taxpayer Wins				0.20 (0.18)	-0.30** (0.14)	-4.8 (8.0)	0.17 (0.15)	-0.26* (0.135)	-4.0 (5.6)
Opinion Year Fixed Effects	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Judge Fixed Effects	No	No	No	No	No	No	Yes	Yes	Yes
Log Pseudo-Likelihood	-1,924.6	-3,893.69	-5,818.29	-613.26	-1,359.24	-1,972.50	-851.51	-1,929.69	-2,781.20
N	7,308	7,308	7,308	2,479	2,479	2,479	4,041	4,041	4,041

Note: Each "Logit" and "GLM" column reflects the first and second part of a two-part regression, with the "Combined" column reflecting the marginal effect calculated by combining both columns. The fixed effects rows indicate whether dummy variables are included for each opinion year, each judge authoring opinions, or both. When judge fixed effects are included, judge characteristics (party and year of appointment) are omitted as multicollinear. *N* varies between regressions because some observations lacked determinate values for some variables (for example, some cases lack a clear winner, since the taxpayer won on some issues and lost on others). Standard errors are clustered by judge. * denotes statistical significance at p<0.1, ** at p<0.05, and *** at p<0.01.

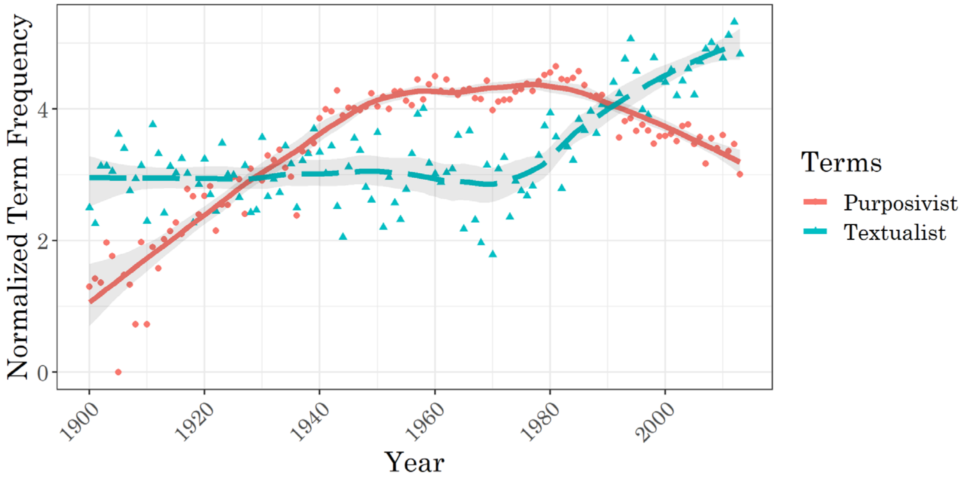
F. Log-Transformed Charts

As noted in Section C of the Appendix, LOESS regression analysis of a long-right-tailed non-normal distribution may inadvertently place outsize importance on outliers. In order to visually ensure that the figures used in this Article are robust and not merely driven by outliers, this Section recreates each term frequency chart using the log-transformation specified in Equation 1[225]:

$$\tilde{y} = \log(1 + y) \tag{16}$$

Visual examination of the log-transformed charts suggests approximately the same results as presented earlier in this Article.

Figure 20: Purposivist and Textualist Terms in Supreme Court Opinions



[225] Note that in each case, the term frequency subjected to the log-transform is expressed in terms per million words. This makes the left scale of the graph more readable but does not affect the shape of the curve.

Figure 21: Purposivist and Textualist Terms in District Court Opinions

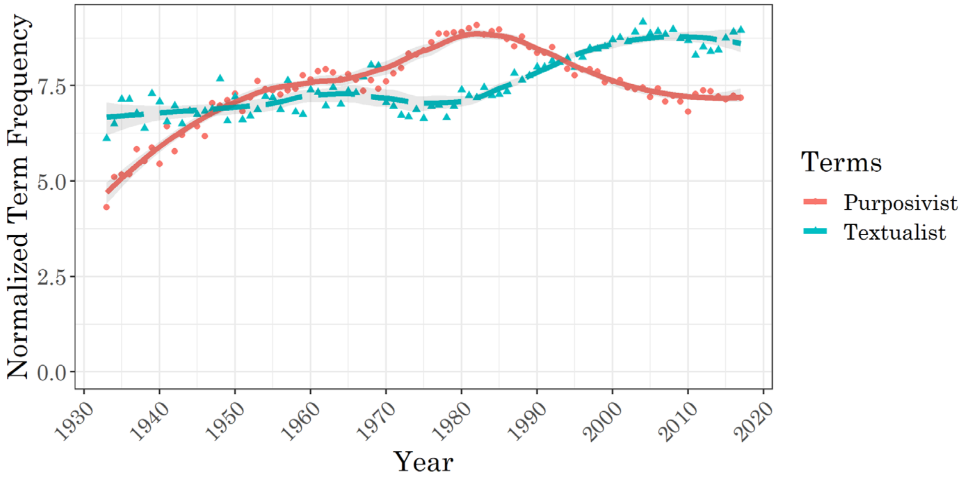


Figure 22: Average Word Count of Tax Court Opinions

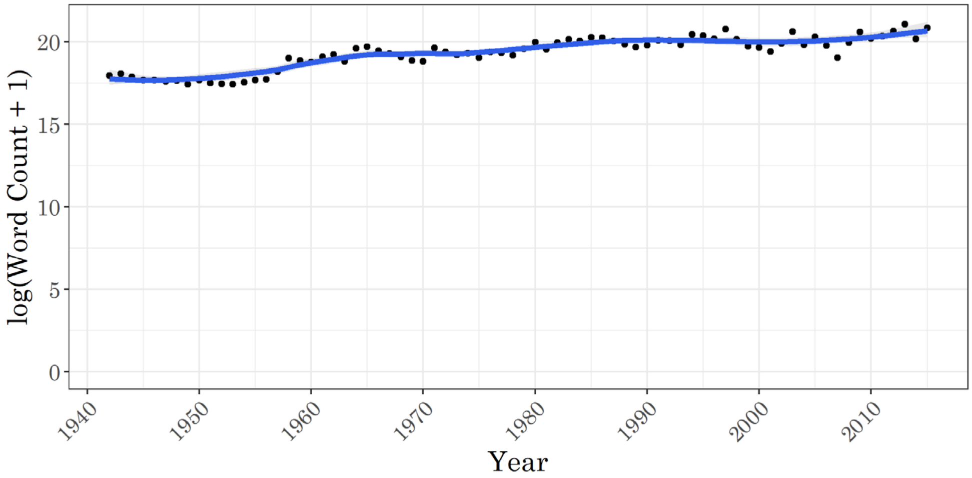


Figure 23: Interpretive and Normative Terms in IRS Publications

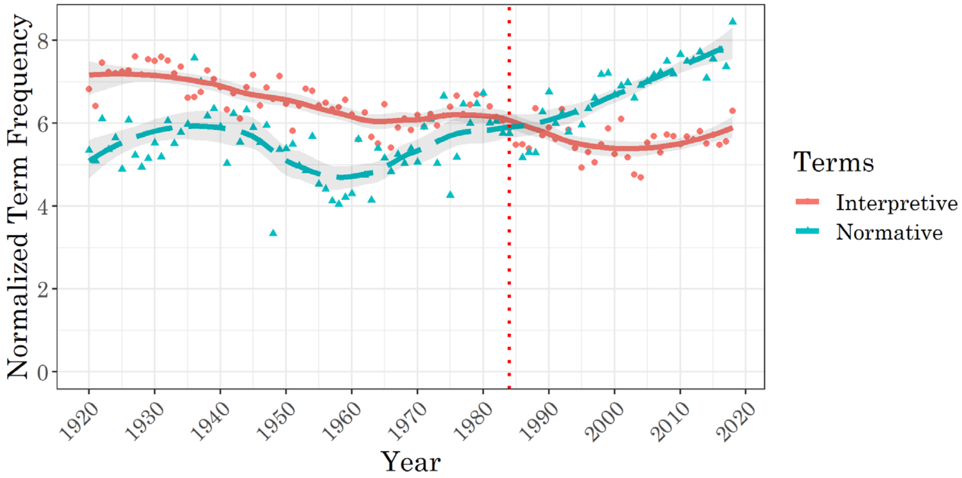


Figure 24: Interpretive and Normative Terms in Tax Court Opinions

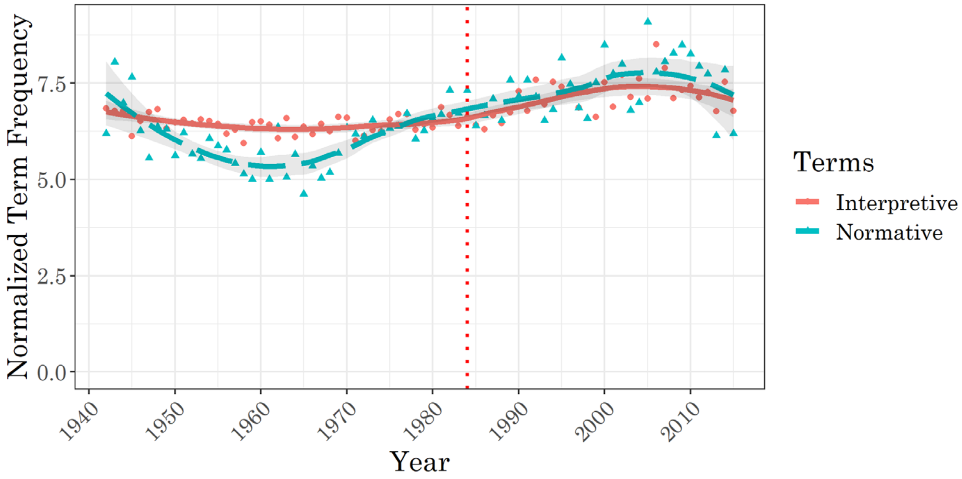


Figure 25: Purposivist and Textualist Terms in IRS Publications

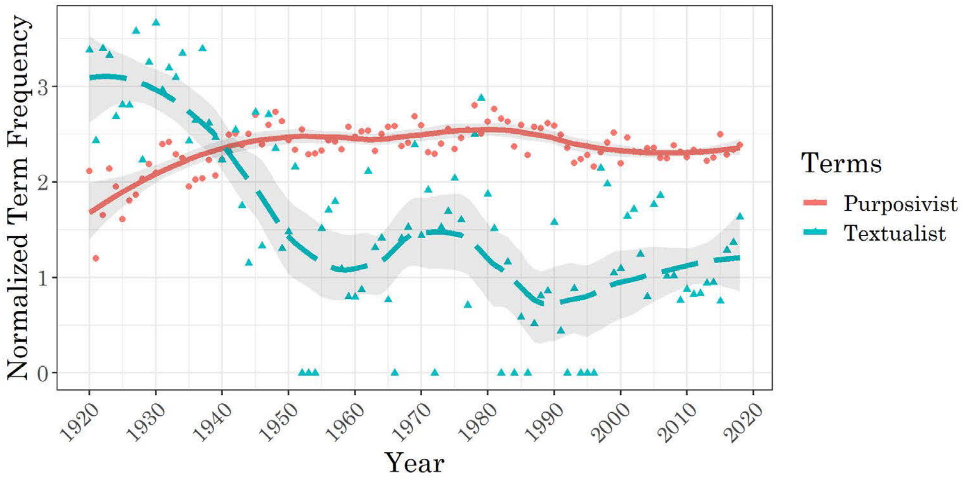
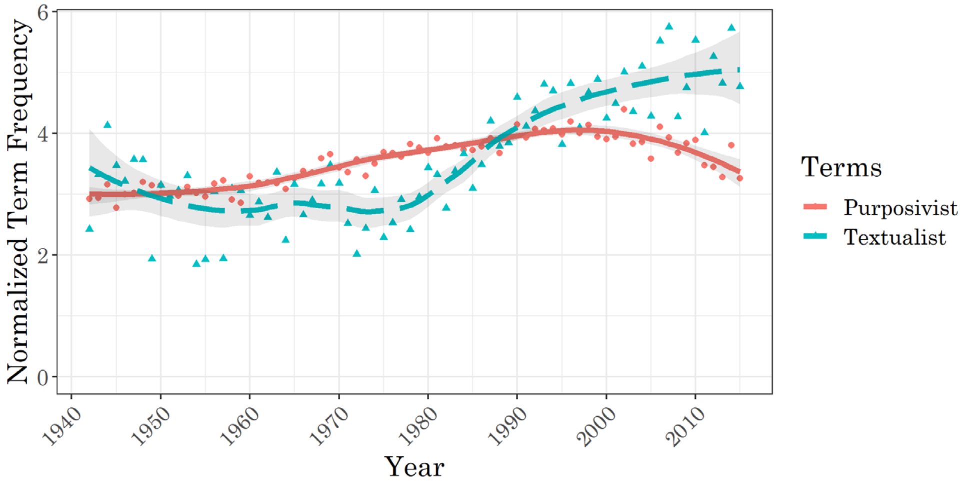


Figure 26: Purposivist and Textualist Terms in Tax Court Opinions



G. Bootstrapped Confidence Intervals for Machine Learning

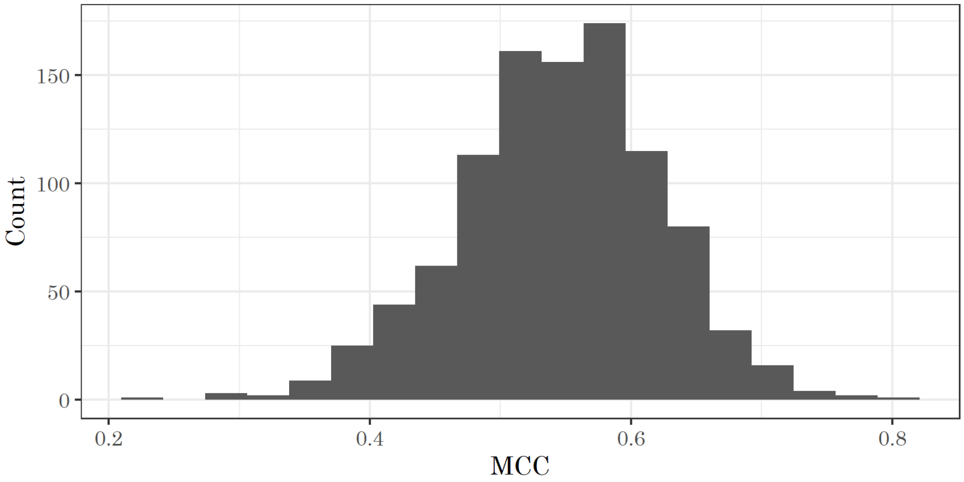
The bootstrapped confidence intervals in Section IV.B were calculated as percentile bootstrap confidence intervals, a form of non-parametric confidence interval initially developed by Bradley Efron and Robert Tibshirani.[226] These are sometimes known as "empirical confidence intervals" and avoid making certain assumptions about the functional form of standard errors. Consequently, they are better suited to bootstrapping than

[226] *See generally* BRADLEY EFRON & ROBERT J. TIBSHIRANI, AN INTRODUCTION TO THE BOOTSTRAP 171 (1993).

conventional confidence intervals. The Python code used to conduct the bootstrapping and to calculate the confidence intervals is available online.[227] I conducted bootstrapping with 1000 tests.

The histograms generated from the bootstrapping, Figures 27a through 29b below, suggest that each of the performance statistics used (MCC, accuracy, and F_1 score) was approximately normally distributed over the bootstrapping tests.

Figure 27a: Histogram of MCC Results from Bootstrapping, Tax Court v. District Courts



[227] *Code*, JONATHAN H. CHOI (last updated Aug. 1, 2019), https://www.jonathanhchoi.com/code.

Figure 27b: Histogram of MCC Results from Bootstrapping, Tax Court v. CFC

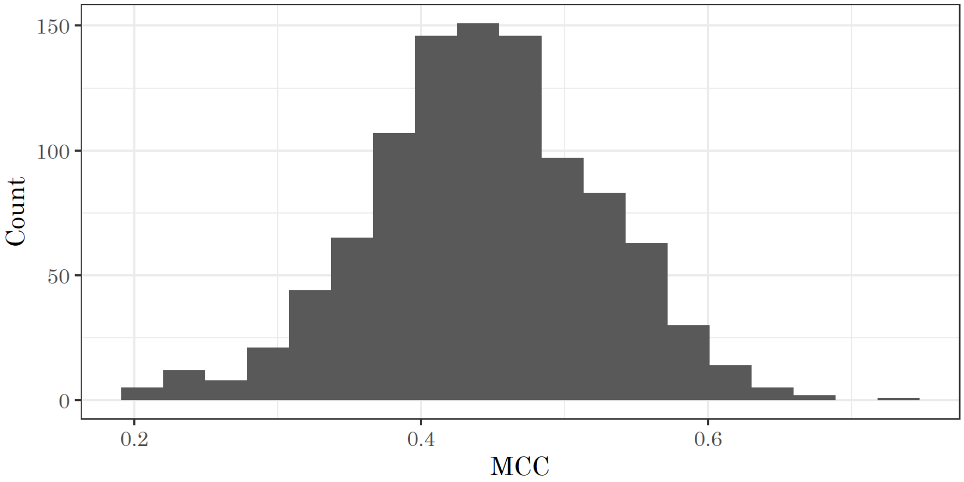


Figure 28a: Histogram of Accuracy Results from Bootstrapping, Tax Court v. District Courts

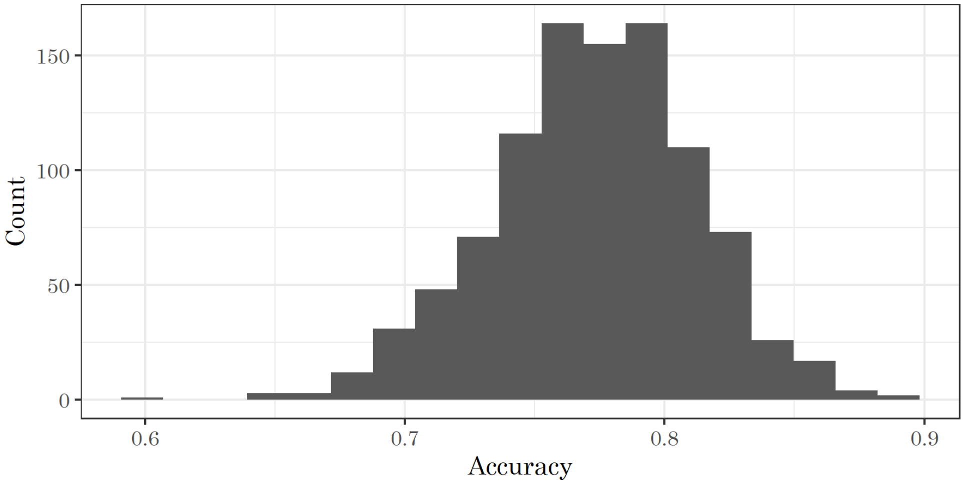


Figure 28b: Histogram of Accuracy Results from Bootstrapping, Tax Court v. CFC

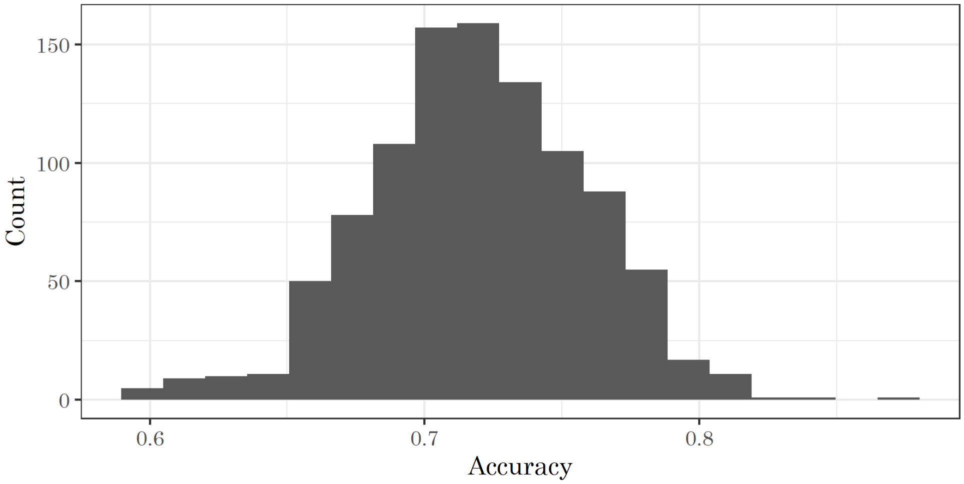


Figure 29a: Histogram of F1 Results from Bootstrapping, Tax Court v. District Courts

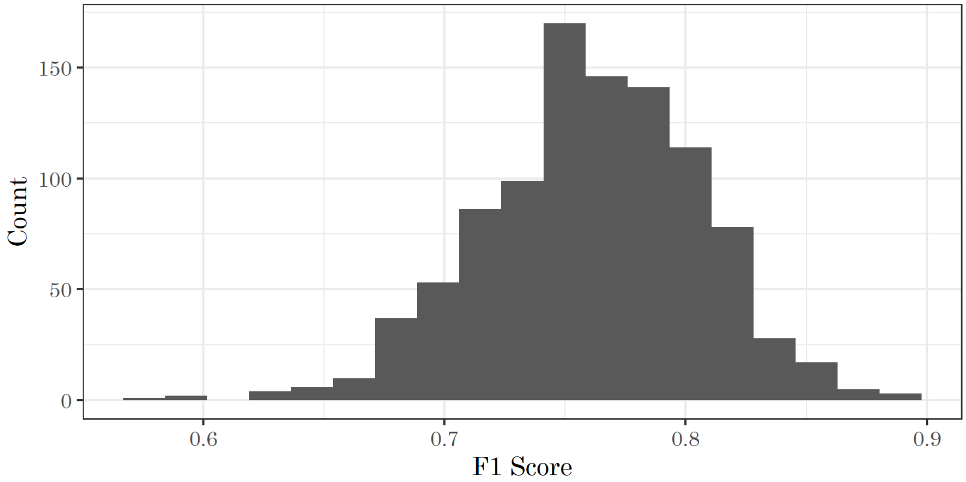
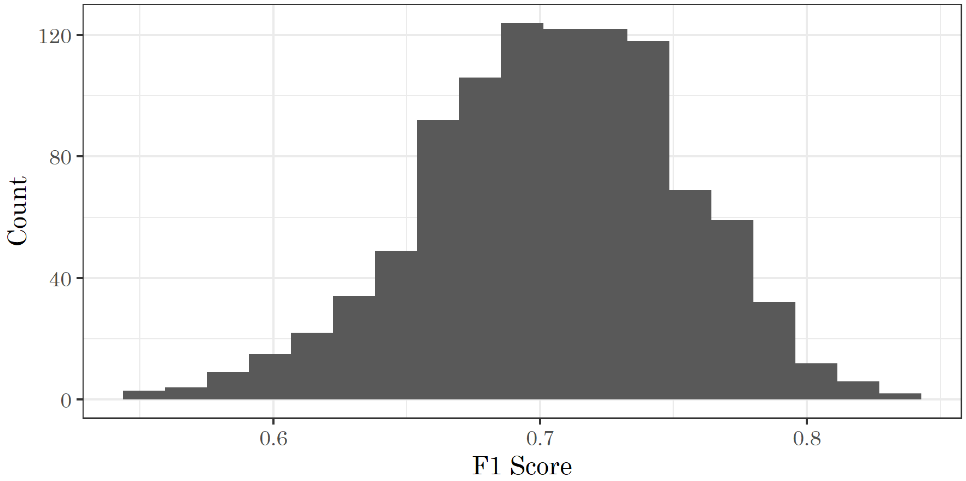


Figure 29b: Histogram of F1 Results from Bootstrapping, Tax Court v. CFC


Count
F1 Score