SIMPLIFICATION OR SEMANTICS? EVALUATING *VAVILOV*'S IMPACT ON STANDARD OF REVIEW

PAUL A. WARCHUK*

The Supreme Court of Canada's pivotal decision in Canada (Minister of Citizenship and Immigration) v. Vavilov *introduced a categorical approach to standard of review analysis, aiming to simplify the existing framework. This article traces the evolution of standard of review analysis and outlines previous empirical studies that examine* Vavilov's *effect on this analysis. The article describes a new empirical study that employs a current large language model to measure various variables pertaining to Federal Court and Federal Court of Appeal decisions, such as length of standard of review analysis and party agreement on standard of review. The findings confirm that* Vavilov *has simplified the standard of review analysis, but perhaps that this simplification may have resulted from an evolving approach that began in the years preceding* Vavilov.

TABLE OF CONTENTS

INTRODUCTION

In the 2019 case of *Canada (Minister of Citizenship and Immigration) v. Vavilov*, the Supreme Court of Canada promised to simplify one of Canadian law's most vexing problems: selecting the appropriate standard of review in judicial review proceedings.[1] Early accounts lauded the Supreme Court's new categorical approach for delivering on its promise of simplification, even if it came with conceptual trade-offs. Five years and thousands of decisions later, we can now evaluate whether the predicted simplification has materialized in practice.

* Assistant Professor, University of New Brunswick. I am grateful to Robert Green, Naomi Pinno, and Mary Pilcher for their excellent research assistance. Thank you also to the participants of the *Vavilov* at 5 Conference for their helpful feedback.

1 2019 SCC 65 [*Vavilov*].

Empirical research on the practical effects of *Vavilov* has already revealed surprising results. A recent study by Andrew Green found that landmark administrative law cases — including *Vavilov* — have had limited impact on the ground in the Federal Court. This limited impact was evident in the stability of both standard selection and case outcomes before and after *Vavilov*.[2]

To examine *Vavilov*'s impact on the process of selecting a standard of review, I used computational legal research methods to examine over 18,000 Federal Court and over 2,000 Federal Court of Appeal judicial review and statutory appeal decisions. The data reveals that *Vavilov* has simplified the standard of review analysis, with mean analysis length dropping by a staggering 48 percent in the Federal Court, while also suggesting that simplification was already underway before the Supreme Court intervened.

Beyond simplification, the analysis also sheds light on which standards courts actually select. Consistent with earlier studies, it reveals that the reasonableness standard had come to dominate judicial review before *Vavilov* instituted the presumption of reasonableness. However, following *Vavilov*'s instruction to apply standards to statutory appeals set forth in *Housen v. Nikolaison*,[3] the rate of reasonableness review has fallen in the Federal Court of Appeal.

Finally, the data addresses the relationship between standard selection and case outcomes. Although *Vavilov* itself does not appear to have affected the rate of granting relief, the data challenges previous findings that suggest the standard of review is unrelated to outcomes.

This article proceeds in three parts. Part I reviews the evolution of the law of standard of review, focusing on the state of the law before *Vavilov* and scholarly critiques of the pre-*Vavilov* framework. It then summarizes the changes introduced in *Vavilov* and the scholarly assessment of those reforms. Part II examines previous empirical studies of judicial review in Canada and then sets out the methodology for this study. Part III presents the results and analysis of this study.

I. THE EVOLUTION OF STANDARD OF REVIEW DOCTRINE

Standard of review is a relatively recent addition to administrative law. For most of the history of judicial review, there was no question as to the degree of scrutiny that a court would apply.[4] Courts would review issues going to the jurisdiction of the administrator without

---

2 Andrew Green, "How Important Are the Groundbreaking Cases in Administrative Law?" (2023) 73:4 UTLJ 426 [Green, "Groundbreaking Cases"]. He also clearly notes that this could be due to changes in behaviour based on the new legal framework, but sought to account for this possibility with his straddle methodology (*ibid* at 428–29).

3 2002 SCC 33 [*Housen*].

4 For a general overview of the history of standard of review in Canada, see: Audrey Macklin, "Standard of Review: Back to the Future?" in Colleen M Flood & Lorne Sossin, eds, *Administrative Law in Context*, 2nd ed (Toronto: Emond Montgomery Publications, 2013) 279; Paul Daly, "The Struggle for Deference in Canada" in Hanna Wilberg & Mark Elliott, eds, *The Scope and Intensity of Substantive Review: Traversing Taggart's Rainbow* (Oxford: Hart Publishing, 2015) 297.

deference, while all other issues were unreviewable.[5] This all-or-nothing approach created perverse incentives for courts to expand the murky concept of jurisdiction to permit review.[6]

In 1979, Justice Dickson famously opined: "The question of what is and is not jurisdictional is often very difficult to determine. The courts, in my view, should not be alert to brand as jurisdictional, and therefore subject to broader curial review, that which may be doubtfully so."[7] It was likely to avoid the continuing expansion of the concept of jurisdiction that he then introduced the patent unreasonableness standard of review to Canadian administrative law.[8] This intermediate standard permitted courts to review and quash a legal interpretation made within an administrator's jurisdiction, where such interpretation was "so patently unreasonable that its construction cannot be rationally supported by the relevant legislation."[9] While Justice Dickson referenced administrative expertise and the presence of privative clauses as reasons for judicial deference to these legal interpretations, the application of patent unreasonableness followed a categorical rule: all legal interpretations within jurisdiction were subject to the patent unreasonableness standard.[10]

A decade later, in *U.E.S., Local 298 v. Bibeault*, Justice Beetz adopted a more nuanced framework for determining jurisdictional boundaries themselves. Rather than relying on the traditional preliminary or collateral questions doctrine — based on the idea that jurisdictional questions were those that had to be answered before the administrator could turn to the merits — Justice Beetz instructed judges to focus "directly on the intent of the legislator."[11] The key question was whether the legislator intended the question to be within the jurisdiction of the administrator.[12] And to determine this, judges were to perform a "pragmatic and functional" analysis, which required courts to examine the wording of the enabling statute, its purpose, and the expertise of the decision-makers to determine the scope of their jurisdiction.[13]

This pragmatic and functional analysis for determining jurisdictional boundaries would soon expand beyond its original purpose. Courts began to shift their focus from using the pragmatic and functional approach to determine the scope of an administrator's jurisdiction to applying the same contextual factors to determine the appropriate standard of review for all administrative decisions. Throughout the 1990s, the Supreme Court refined the pragmatic and functional approach, adding factors that would help courts ascertain legislative intent, such as the tribunal's role and presence or absence of a privative clause.[14] The Supreme Court

---

5 See PW Hogg, "The Supreme Court of Canada and Administrative Law, 1949-1971" (1973) 11:2 Osgoode Hall LJ 187 at 204.

6 In theory, questions of jurisdiction were preliminary or collateral questions that had to be answered before the administrator could exercise their powers. But in practice, the concept was extremely perplexing, with courts recognizing a few dozen types of errors that would result in the loss of jurisdiction: RA Macdonald, "Absence of Jurisdiction: A Perspective" (1983) 43:2 R Barreau 307.

7 *CUPE v NB Liquor Corporation*, 1979 CanLII 23 at 233 (SCC) [*CUPE*].

8 *Ibid* at 237. See also Paul Daly, "The Unfortunate Triumph of Form Over Substance in Canadian Administrative Law" (2012) 50:2 Osgoode Hall LJ 317 at 319–20 [Daly, "The Unfortunate Triumph"].

9 *CUPE*, *supra* note 7 at 237.

10 *Ibid* at 235–37.

11 [1988] 2 SCR 1048 at 1089.

12 *Ibid* at 1087.

13 *Ibid* at 1088.

14 *Pezim v British Columbia (Superintendent of Brokers)*, [1994] 2 SCR 557 at 589–90.

also introduced a third standard of review, reasonableness *simpliciter*, positioned between correctness and patent unreasonableness.[15]

In *Pushpanathan v. Canada (Minister of Citizenship and Immigration)*, the Supreme Court systematized this evolution by articulating a comprehensive four-factor test for the pragmatic and functional approach: the presence or absence of a privative clause, the relative expertise of the decision-maker, the purpose of the enabling legislation as a whole and the provision in particular, and the nature of the issue.[16] Crucially, *Pushpanathan* clarified that this contextual analysis should be applied to determine the appropriate standard of review for all administrative decisions, with "jurisdictional error" redefined as simply "an error on an issue with respect to which, according to the outcome of the pragmatic and functional analysis, the tribunal must make a correct interpretation and to which no deference will be shown."[17] This marked the complete transition from the formalistic preliminary question doctrine to a unified contextual framework for determining both jurisdictional boundaries and standards of review.

Although the *Pushpanathan* "pragmatic and functional" approach successfully rejected the formalistic preliminary question doctrine that scholars had long critiqued, it created its own practical difficulties.[18] Its multi-factor approach was complicated for courts to apply, with factors often pointing in opposite directions, conflicting Supreme Court of Canada jurisprudence, and a hodgepodge of categories of exceptions. Matthew Lewans referred to the period from 2002 to 2008 as the "dis-functional" period of judicial deference.[19] David Mullan acknowledged that standard of review "has been a distracting feature of Canadian judicial review law for over 50 years but more intensely so since the advent of the 'pragmatic and functional' analysis."[20] The frustration grew such that lower courts began to criticize it.

A. *DUNSMUIR*

In *Dunsmuir v. New Brunswick*, Justices Bastarache and LeBel acknowledged these difficulties in setting the stage for another reformulation:

> The Court has moved from a highly formalistic, artificial "jurisdiction" test that could easily be manipulated, to a highly contextual "functional" test that provides great flexibility but little real on-the-ground guidance, and offers too many standards of review. What is needed is a test that offers guidance, is not formalistic or artificial, and permits review where justice requires it, but not otherwise. A simpler test is needed.[21]

The new approach returned to two standards of review: reasonableness and correctness. It also created a new standard of review analysis to determine which standard of review to apply.

---

15 *Canada (Director of Investigation and Research) v Southam Inc*, 1997 CanLII 385 at paras 58–60 (SCC).
16 1998 CanLII 778 at paras 29–38 (SCC) [*Pushpanathan*].
17 *Ibid* at para 28.
18 Daly, "The Unfortunate Triumph", *supra* note 8.
19 Matthew Lewans, "Deference and Reasonableness Since *Dunsmuir*" (2012) 38:1 Queen's LJ 59 at 71.
20 David Mullan, "*Dunsmuir v. New Brunswick*, Standard of Review and Procedural Fairness for Public Servants: Let's Try Again!" (2008) 21:2 Can J Admin L & Prac 117 at 118.
21 2008 SCC 9 at para 43 [*Dunsmuir*].

The first step of the new standard of review analysis was to ascertain whether the jurisprudence had already determined in a satisfactory manner the degree of deference to be accorded with regard to a particular category of question. If it had, then the court need not go further. If it had not, the court must proceed to an analysis of the factors making it possible to identify the proper standard of review.[22] These factors were the same four factors as the pragmatic and functional approach.

The Supreme Court also listed factors that would heavily favour one standard or the other. It advised that privative clauses; questions of fact, discretion, or policy; interpretations of an administrator's home statute or statutes closely connected to its function; or the application of a general common law or civil law rule in relation to a specific statutory context would likely attract reasonableness review.[23] Meanwhile correctness was likely to apply to constitutional questions, true questions of jurisdiction, questions of general law "that [are] both of central importance to the legal system as a whole and outside the adjudicator's specialized area of expertise," and questions regarding the jurisdictional lines between two or more competing specialized tribunals.[24]

Initial reviews of *Dunsmuir* were positive. Scholars agreed that the majority's more categorical approach would simplify standard of review analysis.[25] Andrew Green, for instance, argued that the categorical approach would make it easier for lower courts to determine the standard of review and reduce the number of mistakes in getting there.[26] However, Paul Daly criticized *Dunsmuir* as a return to formalism.[27] While Daly acknowledged that the pragmatic and functional approach had significant pre-decision costs, he warned that the categorical approach was not as straightforward as it may appear. The categories were both under and over inclusive, overlapping and without clear instructions on how to go about resolving conflicts or boundary disputes.[28] Others warned that the simplicity of the new standard of review analysis was artificial as it pushed the disagreement from the selection of the standard to the application of reasonableness.[29]

In the years after *Dunsmuir*, Daly's early warnings would come to pass. The likely categories soon solidified into presumptions and the Supreme Court became bitterly divided on when it was appropriate to depart from those presumptions, and what factors should guide such a decision.[30] In early 2016, Justice Stratas circulated a blistering critique, which detailed the chaos plaguing administrative law, including standard of review analysis.[31] Later that

---

22 *Ibid* at para 62.

23 *Ibid* at paras 52–54.

24 *Ibid* at paras 58–61, quoting LeBel J in *Toronto (City) v CUPE, Local 79*, 2003 SCC 63 at para 62.

25 Macklin, *supra* note 4 at 320. See also Gerald P Heckman, "Substantive Review in Appellate Courts since *Dunsmuir*" (2009) 47:4 Osgoode Hall LJ 751 at 783. Note that while Heckman concluded that *Dunsmuir* simplified standard of review analysis, he also warned that the *Dunsmuir* framework risked reducing deference (*ibid*).

26 Andrew Green, "Can There Be Too Much Context in Administrative Law? Setting the Standard of Review in Canadian Administrative Law" (2023) 47:2 UBC L Rev 443 at 490.

27 Daly, "The Unfortunate Triumph", *supra* note 8.

28 Paul Daly, "*Dunsmuir*'s Flaws Exposed: Recent Decisions on Standard of Review" (2012) 58:2 McGill LJ 483.

29 Mullan, *supra* note 20 at 125.

30 Paul Daly, "The Scope and Meaning of Reasonableness Review" (2015) 52:4 Alta L Rev 799.

31 David Stratas, "The Canadian Law of Judicial Review: A Plea for Doctrinal Coherence and Consistency" (2016) 42:1 Queen's LJ 27.

year, Justice Abella wrote of the need to "simplify the standard of review labyrinth we currently find ourselves in."[32] But her colleagues refused to take up her charge and polarization continued.[33]

On the tenth anniversary of *Dunsmuir*, Paul Daly and Léonid Sirota organized a digital symposium, with dozens of scholars and practitioners contributing comments.[34] It was clear from many of the comments that *Dunsmuir* had failed to simplify and clarify the law, regarding both the selection of a standard of review and the application of the reasonableness standard.[35] A few weeks later, with criticism mounting and internal divisions persisting, the Supreme Court took the unprecedented step of announcing that it would use the *Vavilov* case as an opportunity to reconsider "the nature and scope of judicial review of administrative action."[36]

B. *VAVILOV*

Eighteen months after granting leave, and one year after hearing from the parties, 32 interveners, and two amici curiae, the Supreme Court of Canada released its decision in *Vavilov*.[37] Once again, the Supreme Court was divided, but with a clearer seven to two majority. The majority began by acknowledging the widespread criticism and agreed that "*Dunsmuir*'s promise of simplicity and predictability … [had] not been fully realized."[38] To remedy the confusion, they focused on creating a new standard of review analysis and providing additional guidance on the methodology of conducting a reasonableness review.

On selection of the standard of review, the majority announced that the analysis will always begin with a presumption that the standard is reasonableness — grounded in the legislature's choice to delegate decision-making authority to the administrator rather than a court.[39] The presumption of reasonableness can be rebutted where the legislature has indicated that it intends a different standard to apply or where the rule of law requires that the standard of correctness be applied.[40] To further simplify the process, the Supreme Court then established two categories where legislative intent points away from deference: where the legislature prescribes a specific standard of review or creates a statutory right of appeal.[41] It also established three categories where the rule of law requires correctness: constitutional questions, general questions of law of central importance to the legal system as a whole, and

---

32 *Wilson v Atomic Energy of Canada Ltd*, 2016 SCC 29 at para 19.

33 Robert Danay, "A House Divided: The Supreme Court of Canada's Recent Jurisprudence on the Standard of Review" (2019) 69:1 UTLJ 3.

34 These posts were published as a special edition of the Canadian Journal of Administrative Law and Practice: Paul Daly & Léonid Sirota, eds, *Canadian Journal of Administrative Law & Practice Special Issue - A Decade of Dunsmuir* (Toronto: Thomson Reuters, 2018).

35 Léonid Sirota, "The Paradox of Simplicity in Canadian Administrative Law" (2018) Can J Admin L & Prac 59 (Special Issue); Mark Mancini, "The Dark Art of Deference: Dubious Assumptions of Expertise on Home Statute Interpretation" (2018) Can J Admin L & Prac 83 (Special Issue).

36 *Minister of Citizenship and Immigration v Alexander Vavilov*, 2018 CanLII 40807 (SCC).

37 *Vavilov*, *supra* note 1. *Vavilov* was heard together with *Bell Canada v Canada (Attorney General)*, 2019 SCC 66, and the number of interveners represents the total unique interveners between both cases.

38 *Vavilov*, *supra* note 1 at para 7.

39 *Ibid* at paras 23–32.

40 *Ibid* at para 17.

41 *Ibid* at paras 33–52.

questions related to the jurisdictional boundaries between two or more administrative bodies.[42]

While the Supreme Court left open the possibility of new categories, the majority were of the view that "at this time … these reasons address all of the situations in which a reviewing court should derogate from the presumption of reasonableness review."[43] Thus, any additional derogations would have to be exceptional and based on legislative intent or the rule of law (or both).[44] The majority was clear that "this decision conclusively closes the door on the application of a contextual analysis to determine the applicable standard, and in doing so streamlines and simplifies the standard of review framework."[45]

Vavilov was initially well-received by both academics and practitioners.[46] The new framework was clear and definitive, "respond[ing] effectively to many of the difficulties that plagued this area of law a decade ago."[47] Most comments expressed optimism that standard of review analysis would be durable and consistent. Criticism of the standard of review framework was directed toward the lack of consistent theory, rather than the simplicity of the new process.[48]

Although nearly six years have passed since the initial decision, the promise of simplicity appears to have held. No reports suggest that standard of review analysis has returned to the confusion of the pre-*Vavilov* era. To be sure, when the Supreme Court established a new category of correctness in *Society of Composers, Authors and Music Publishers of Canada v. Entertainment Software Association* — that of concurrent jurisdiction — some feared that it could inspire lower court judges to return to the pre-*Vavilov* battles over correctness and context.[49] However, that does not appear to have come to pass. In *Mason v. Canada*

---

42 *Ibid* at paras 53–64.

43 *Ibid* at para 69.

44 *Ibid* at para 70.

45 *Ibid* at para 47.

46 See e.g. Mark Mancini, "Vavilov: A Step Forward" (19 December 2019), online (blog): [perma.cc/LG28-SW8R]; Mark Mancini, "Canada Post: Vavilov's First Day in the Sun" (20 December 2019), online (blog): [perma.cc/9H7D-VMSY]; Paul Daly, "A Consensus, if You Can Keep it: Canada (Minister of Citizenship and Immigration) v. Vavilov, 2019 SCC 65" (20 December 2019), online (blog): [perma.cc/THW8-WKNJ]; Gerard Kennedy, "20 Things to Be Grateful For as Administrative Law Enters the 2020s" (23 December 2019), online: [perma.cc/XZK3-GYE8]; Michael Swanberg, "The Supreme Court's Decision in Vavilov: A New Framework for Reasonableness Review" (13 January 2020), online: [perma.cc/78H9-V2B7]; Dale Smith, "Presumption of Reasonableness: The Supreme Court Clarifies the Standard of Review", *CBA National Magazine* (19 December 2019), online: [perma.cc/PPJ2-RK9W].

47 Paul Daly, "The Scope and Meaning of Reasonableness Review After *Vavilov*" (5 June 2025) at 11, online (pdf): [perma.cc/ASU9-5UYG].

48 Paul Daly, "Vavilov on the Road" (2022) 35:1 Can J Admin L & Prac 1; Léonid Sirota, "Rebuilt on Sand: Canadian Administrative Law After Vavilov" (2020) 31 Pub L Rev 117; Kate Glover Berger, "The Missing Constitutionalism of Canada v Vavilov" (2021) 34:1 JL & Soc Pol'y 68. See also Mark Mancini & Léonid Sirota, "The End of Administrative Supremacy in Canada" (2024) 57:1 UBC L Rev 31; Léonid Sirota, "Not Good Enough" (20 December 2019), online (blog): [perma.cc/E343-FJSP]. See also Cristie Ford, "Vavilov, Rule of Law Pluralism, and What Really Matters" (27 April 2020), online (blog): [perma.cc/2ZNW-RTZ2]; Mary Liston, "Bell is the Tell I'm Thinking of" (29 April 2020), online (blog): [perma.cc/H3EL-YHY9].

49 2022 SCC 30 at paras 26–28. See Paul Daly, "The Return of Context? Society of Composers, Authors and Music Publishers of Canada v. Entertainment Software Association, 2022 SCC 30" (9 September 2022), online (blog): [perma.cc/Y6Y2-EA2P]; Paul Daly, "Concurrent Jurisdiction: How Broad is the

(Citizenship and Immigration), the Supreme Court held firm to the *Vavilov* categories — despite strong arguments to the contrary.[50]

While anecdotal evidence suggests *Vavilov* has maintained its promise of simplicity, systematic empirical analysis is needed to determine whether this simplification has materialized in judicial practice. The next part examines previous empirical research and sets out the methodology for measuring *Vavilov*'s actual effects.

II. PREVIOUS STUDIES AND METHODOLOGY

Judicial review of administrative action in Canada has been the subject of several empirical studies.[51] Two studies have directly investigated the process of selecting a standard of review, albeit with respect to *Dunsmuir* rather than *Vavilov*.

The first of these studies was a three-part series investigating the effects of *Dunsmuir* on judicial review in British Columbia, Nova Scotia, Quebec, Ontario, Alberta, and the federal courts.[52] Diana Ginn, William Lahey, Lauren Soubolsky, and Madison Veinotte examined 477 cases from between 2008 and 2015, using a coding framework of 104 questions.[53] They also compared their results with an earlier study, which looked at the first four years after *Pushpanathan*.[54]

To assess the simplicity of analysis, the authors looked at the length of analysis in paragraphs. In the federal courts, judges determined the standard of review in a single paragraph about one-third of the time, and in five paragraphs or less 80 percent of the time.[55] In the provincial superior courts of Nova Scotia, Quebec, Ontario, and Alberta, analysis was longer, with only slightly more than half the issues resolved in five paragraphs or less.[56]

The study also revealed that the standard of review was determined on the basis of precedent in 64 percent of the federal court cases and 67 percent of the provincial superior court cases.[57] For only 5 percent of the issues under review did the federal courts perform a

---

Entertainment Software Association Exception?" (4 November 2022), online (blog): [perma.cc/TA6H-WRF4].

50 2023 SCC 21; Léonid Sirota, "It's Nonsense, but it Works" (28 September 2023), online (blog): [perma.cc/F7U4-VQZA]; Paul Daly, "Context, Reasonableness Review and Statutory Interpretation: Mason v. Canada (Citizenship and Immigration), 2023 SCC 21" (28 September 2023), online (blog): [perma.cc/H4DB-Q89T].

51 In addition to the articles cited below, see also: Leonard Marvy & Voy Stelmaszynski, "Judicial Review of Ontario Labour Relations Board Decisions: From *CUPE* to *Dunsmuir*, and Beyond" (2009) 15:3 CLELJ 555; Erika L Ringseis & Allen Ponak, "Judicial Review of Arbitration Awards in Alberta: Frequency, Outcomes and Standard of Review" (2006–2007) 13 CLELJ 415.

52 Diana Ginn et al, "How Has *Dunsmuir* Worked? A Legal-Empirical Analysis of Substantive Review of Administrative Decisions After *Dunsmuir v. New Brunswick*: Findings from the Federal Courts" (2017) 30:1 Can J Admin L & Prac 51 [Ginn et al, "Federal Courts"]; Diana Ginn et al, "A Legal-Empirical Analysis of Substantive Review: Findings from the British Columbia Courts" (2017) 30:2 Can J Admin L & Prac 173; William Lahey et al, "How Has *Dunsmuir* Worked? A Legal-Empirical Analysis of Substantive Review of Administrative Decisions After *Dunsmuir v. New Brunswick*: Findings from the Courts of Nova Scotia, Quebec, Ontario and Alberta" (2017) 30:3 Can J Admin L & Prac 317.

53 Ginn et al, "Federal Courts", *supra* note 52 at 56.

54 Diana Ginn & William Lahey, "After the Revolution: Being Pragmatic and Functional in Canada's Trial Courts and Courts of Appeal" (2002) 25:2 Dal LJ 259.

55 Ginn et al, "Federal Courts", *supra* note 52 at 59.

56 Lahey et al, *supra* note 52 at 323.

57 Ginn et al, "Federal Courts", *supra* note 52 at 60; Lahey et al, *supra* note 52 at 327.

full four-factor analysis.[58] By contrast, provincial superior courts conducted a full four-factor standard of review analysis 21 percent of the time.[59] There was, however, significant variation among the provinces, with Ontario courts conducting a full standard of review analysis most frequently, 35 percent of the time, and Quebec courts, at the other end of the spectrum, doing so just 9 percent of the time.[60]

In terms of the outcome of the standard of review analysis, federal courts selected the deferential reasonableness standard approximately 75 percent of the time,[61] while the provincial superior courts selected reasonableness 82 percent of the time overall, ranging from 79 percent in Ontario to 88 percent in Quebec.[62]

Finally, the study found that federal courts granted judicial review in 66 percent of cases — almost exactly the same percentage as in their earlier pre-*Dunsmuir* study. Provincial superior courts showed greater deference, upholding 75 percent of administrative decisions, with individual provinces ranging from 68 percent of decisions upheld in Alberta to 79 percent in Ontario.[63]

Turning to the second study investigating the selection of standard of review, Robert Danay sampled 120 factums filed by litigants in the Ontario Divisional Court and the federal courts before and after *Dunsmuir* was decided.[64] To assess complexity, he calculated the total number of paragraphs and the proportion of each factum devoted to standard of review analysis. In the Ontario Divisional Court, the average number of paragraphs dealing with the standard of review increased from 8.5 paragraphs per factum before *Dunsmuir* to 10.4 paragraphs. Similarly, in the federal courts, the average number of paragraphs dealing with the standard of review increased from 2.6 paragraphs per factum before *Dunsmuir* to 3.4 paragraphs after.

In another study, Danay reviewed 177 Supreme Court of Canada cases from the 1998 *Pushpanthan* decision to early 2016 to better understand *Dunsmuir*'s effect on deference.[65] He tracked individual votes of members of the Supreme Court, finding that 43 percent of votes were for correctness review pre-*Dunsmuir*, and only 17 percent after. Additionally, the reasonableness standard appeared to become more deferential after *Dunsmuir*. When reasonableness was applied pre-*Dunsmuir* it resulted in 31 percent of cases being overturned, compared to just 19 percent post-*Dunsmuir*.[66]

More recently, Andrew Green conducted an empirical study of 1,076 Federal Court decisions from 2007 to 2019.[67] The study tracked both the standard of review selected and

---

58 Ginn et al, "Federal Courts", *supra* note 52 at 61.
59 Lahey et al, *supra* note 52 at 336.
60 *Ibid* at 336–37.
61 Ginn et al, "Federal Courts", *supra* note 52 at 58.
62 Lahey et al, *supra* note 52 at 320.
63 *Ibid* at 321.
64 Robert Danay, "Did *Dunsmuir* Simplify the Standard of Review? An Empirical Assessment" (2018) Can J Admin L & Prac 201 (Special Issue).
65 Robert Danay, "Quantifying *Dunsmuir*: An Empirical Analysis of the Supreme Court of Canada's Jurisprudence on Standard of Review" (2016) 66:4 UTLJ 555.
66 *Ibid* at 577.
67 Green, "Groundbreaking Cases", *supra* note 2.

the result. Although the use of the reasonableness standard increased from under 60 percent in months following the *Dunsmuir* decision to nearly 90 percent after *Vavilov*, the rate of granting judicial review remained nearly unchanged.[68] Green performed a number of logistic regressions to try to isolate variables that may impact both the selection of standard of review and the rate of granting judicial review, including the subject-matter, administrative decision-maker, and issue under review.[69] While some of these variables were significant predictors, neither *Dunsmuir* nor *Vavilov* increased the odds of an administrative decision being upheld.

A. METHODOLOGY

This study examines applications for judicial review and statutory appeals heard by the Federal Court and Federal Court of Appeal and decided between 2008 and 2024. I chose those dates to capture the *Dunsmuir* period, as *Dunsmsuir* was released in March 2008, and the first five years of *Vavilov*, which was released December 2019. It is thus possible to roughly divide this period into the pre-*Vavilov* era (2008 to 2019) and post-*Vavilov* era (2020 to 2024).

I focused on the Federal Court and Federal Court of Appeal for two practical reasons. First, these courts are saturated with administrative law cases, hearing both a large volume of such cases and having administrative law represent a significant portion of their overall workload. This concentration reduces the resources needed to filter out irrelevant cases from the dataset. Second, accessing full-length judgments in bulk presents significant challenges. Major legal databases such as Quicklaw, Westlaw and CanLII prohibit users from bulk downloading decisions. The federal courts' decisions are available through the Refugee Law Lab's bulk datasets, making comprehensive analysis feasible.

To prepare the Federal Court and Federal Court of Appeal datasets for this study, I began by downloading the bulk decision datasets prepared and hosted by the Refugee Law Lab.[70] These bulk datasets contain full, unofficial reasons for judgment (hereafter decisions), scraped from the respective court websites. The versions I used included all cases posted as of 31 December 2024.

As the bulk datasets contain Federal Court or Federal Court of Appeal decisions across many areas of law, the first step was to filter out irrelevant cases.[71] I limited the final datasets to English language decisions on the merits of a judicial review or statutory appeal from an administrative decision-maker.[72] Motions and cases where there was no substantive review issue were eliminated.

Within the corpus of the two courts, there are a significant number of decisions that do not discuss standard of review at all — approximately 30 percent of the qualifying decisions from the Federal Court of Appeal and 15 percent from the Federal Court. Rather than

---

68 *Ibid* at 441. Note that Green coded cases as reasonableness, correctness, multiple, or other.

69 *Ibid* at 446–47.

70 Refugee Law Lab, "Federal Court of Appeal Bulk Decisions Dataset" (2025), online: [perma.cc/L62X-4U75]; Refugee Law Lab, "Federal Court Bulk Decisions Dataset" (2025), online: [perma.cc/F4JJ-ZYUY].

71 The filtering process was accomplished using the same method as described below for coding the cases.

72 For the Federal Court of Appeal dataset, a case could also be an appeal from the Federal Court (FC) where the FC decision was a judicial review or statutory appeal from an administrative decision-maker.

attempting to estimate the standard applied, I excluded these decisions from the final analysis because they were largely irrelevant to the primary question of whether standard of review analysis has become simpler.[73] Additionally, they lacked sufficient analytical content for reliable coding. Estimating the standard of review by interpreting the court's analysis alone is highly subjective and ultimately impossible in many instances.[74] Nonetheless, the elimination of these cases also produces limitations. For instance, courts that do not address standard of review may be disproportionately applying the correctness standard. However, this potential bias does not undermine the core findings about simplification trends among cases where courts do engage with standard of review analysis.

After the final datasets were established, I machine-coded each dataset separately. To code each case, I used Gemini Flash 2.5, a large language model (LLM).[75] An LLM is a type of artificial intelligence trained on enormous datasets of text to learn patterns in language, allowing it to predict the most probable sequence of words in response to a given input. The most well-known LLM is ChatGPT, but many alternatives are now commercially available.[76] Gemini Flash 2.5 is one such commercially available LLM, created by Google.

To get Gemini Flash 2.5 to filter and code the cases, I prepared a Python script that pulled cases from the bulk datasets one at a time and fed them to the LLM's application programming interface, or API, along with a pre-set prompt. The prompt contained a list of questions for the LLM to answer, such as "Does the Court explicitly select the standard of reasonableness for at least one of the substantive review issues?" To facilitate data analysis, most questions were phrased as yes or no questions. The LLM was also instructed to begin each answer with the "#number" of the question it was responding to. The script recorded the responses in an Excel workbook in the appropriate column.

I selected an LLM, rather than a human coder, for several reasons. The primary reason was the volume of decisions (approximately 28,000), which made manual review of the full dataset unrealistic. While research assistants took weeks to code a few hundred cases for auditing purposes, the LLM was able to code all 28,000 in a single day. Previous studies have all used sampling to get around the issue of volume. While good random sampling can approximate population characteristics, it can never achieve the same degree of statistical power as analyzing the complete dataset.

Additionally, human coders are not perfect. Repetitive coding tasks across thousands of cases are inherently error-prone, particularly as coder fatigue and attention lapses accumulate over time. When a large team of different coders is required, subjective differences between coders can introduce further inconsistencies. In such circumstances, machine coding presents

---

73 I included these decisions in the dataset so that I could consider the rate of ignoring standard of review. As I will suggest below, the failure to address standard of review at all may be connected to the complexity of the standard of review analysis.

74 I tested asking the coding agent to provide its best estimate; however, I was unable to get reliable results.

75 Version: Preview 04–17.

76 For a discussion of the technology and its application to legal research, see Sean Rehaag, "Luck of the Draw III: Using AI to Extract Data About Decision-Making in Federal Court Stays of Removal" (2024) 49:2 Queen's LJ 73.

an opportunity for increased accuracy, particularly because LLM settings can be tweaked to provide consistent and reproduceable results.[77]

This study therefore provides an opportunity to assess whether current LLM technology can reliably code complex legal doctrine. Some work has already been done on this, with positive results.[78] In Canada, Sean Rehaag successfully employed OpenAI's GPT-3.0 LLM to extract data from Federal Court stay decisions, using the LLM to identify stay-related docket entries, extract judge names, and classify case outcomes from natural language text. Rehaag's study demonstrated that LLMs can achieve 98 to 99 percent accuracy in legal document coding while processing tens of thousands of cases — a scale that would have required countless hours of human coding.[79]

This study provides an opportunity to expand our understanding of LLMs in legal research in several ways. First, it employs Gemini Flash 2.5, a more recent and powerful model than the GPT-3.0 used by Rehaag. As Rehaag noted in his paper, LLM technology has seen rapid advancements since his analysis was performed in 2022. Second, while Rehaag's coding tasks focused on extracting discrete datapoints like judge names and binary outcomes, this study tackles more complex analytical questions — determining whether courts engaged meaningfully with standard of review frameworks and assessing the depth of legal analysis. Third, whereas Rehaag fine-tuned GPT-3.0 by providing sample inputs and expected responses, fine-tuning is not available for Gemini Flash 2.5. This study therefore tests whether newer LLMs can handle complex legal doctrine analysis without fine-tuning, further contributing to our understanding of computational methods in legal empirical research.

While LLMs offer significant advantages for large-scale legal document coding, their use also introduces several methodological risks that require careful validation. The most significant concern is hallucination, where LLMs confidently code content that does not actually exist in the source material. LLMs may also exhibit prompt sensitivity, where minor variations in coding instructions could yield systematically different results across similar cases.[80]

Given these risks, I drew random samples of cases from the original datasets using Python's built-in random module to audit. I then had two research assistants manually validate 100 Federal Court and 100 Federal Court of Appeal decisions by answering the same questions and comparing their results to the LLM's. This resulted in error rates of 1.3 percent and 4.1 percent, respectively. I too performed a manual validation of 100 Federal Court and 100 Federal Court of Appeal separately sampled cases, finding error rates of less than 1 percent and 2.6 percent respectively. However, I found that the LLM's average word count was consistently 10 percent higher than mine.

---

77 This involves configuring the "temperature" to zero. The "temperature" setting controls the randomness or creativity of the generated text. A temperature of zero makes the model's choices highly deterministic, consistently favoring the most probable word or phrase.

78 See e.g. Jonathan H Choi, "How to Use Large Language Models for Empirical Legal Research" (2024) 180:2 J Institutional & Theoretical Econs 214; Caleb Ziems et al, "Can Large Language Models Transform Computational Social Science?" (2024) 50:1 Computational Linguistics 237.

79 Rehaag, *supra* note 76 at 96, 97, note 92.

80 Jonathan H Choi, "Off-the-Shelf Large Language Models are Unreliable Judges", (2025) online (pdf): [perma.cc/QG9Q-2ZFG].

While the audit results show the LLM's coding was generally accurate, there were a number of limitations with the LLM which should be mentioned. I also tested identifying the standard of review on a per-issue, rather than per-case basis. Auditing the preliminary results revealed that judges do not consistently and clearly demarcate different issues. This is unsurprising as the Supreme Court had instructed courts to treat judicial review as an "organic exercise," focusing on the totality of the decision rather than segmenting individual issues for scrutiny.[81] The result, however, was that the LLM struggled to separate issues and their corresponding standards of review, particularly where there were several issues all subject to the reasonableness standard.

Another major limitation of LLMs is accurately counting words. In early testing, I asked the LLM to count the number of words judges used to analyze and determine the standard of review. My research assistants discovered that the resulting word counts were inconsistently wrong. There were no apparent patterns in the mistakes. As a workaround, we had the LLM extract the exact text of the decision that discusses standard of review. Then, we used Excel functions to perform the word count by counting spaces between words.[82]

III. RESULTS AND ANALYSIS

The primary goal of this study is to assess whether *Vavilov* simplified the process of selecting a standard of review. The data also answers two related questions: (1) whether *Vavilov* changed which standards courts select, and (2) whether any such changes affect case outcomes. I address each question in turn, beginning with the simplification question and its various metrics.

A. *VAVILOV*'S EFFECT ON SIMPLICITY

The first and most direct measure of simplification is the length of standard of review analysis. As discussed above, previous studies have adopted length of analysis as a proxy for legal complexity. Length serves as a reliable proxy for complexity because judicial decisions derive authority from the persuasiveness of their reasoning.[83] When legal frameworks are complex or unclear, judges must write more extensively to establish the legitimacy of their conclusions, often needing to reconcile conflicting precedents or work through competing factors in multi-step tests. When legal rules are straightforward and well-settled, judges can reach equally persuasive determinations more directly with concise analysis. Accordingly, if *Vavilov* successfully simplified standard of review selection, we should observe shorter standard of review analyses.

One objection to using analysis length as a proxy for complexity is that length can be influenced by many factors, including changes in judicial writing style. To test for this possibility, I separately asked the LLM to extract analyses of other standards of review

---

[81] *Newfoundland and Labrador Nurses' Union v Newfoundland and Labrador (Treasury Board)*, 2011 SCC 62 at para 14; *Mouvement laïque québécois v Saguenay (City)*, 2015 SCC 16 at para 173.

[82] The formula was: =IF(LEN(TRIM(A2))=0,0,LEN(TRIM(A2))-LEN(SUBSTITUTE(TRIM(A2)," ","""))+1).

[83] Peter Cane, *Controlling Administrative Power: An Historical Comparison* (Cambridge: Cambridge University Press, 2016) at 33.

present in the final dataset: appellate standard (*Agraira*[84]), discretionary review of the first instance judge, and procedural fairness. The length of analysis for each of these three standards of review was longer in the period from 2020 to 2024 than it was in the period from 2008 to 2019. This finding reduces the possibility that any reduction in substantive standard of review analysis length is due to external factors such as evolving judicial writing conventions. Having established that analysis length provides a reliable measure of legal complexity in this context, I now turn to the empirical results.

The aggregate data from both courts supports the hypothesis that *Vavilov* simplified standard of review analysis. Between 2008 and 2019, the mean word count for determining the substantive standard of review in the Federal Court was 153 words (median: 88 words).[85] Between 2020 and 2024, this dropped to 80 words (median: 49 words). The Federal Court of Appeal is much the same story, with a mean analysis length of 259 words before (median: 122 words) and 178 words (median: 105 words) after. These substantial decreases — 48 percent for the Federal Court and 31 percent for the Federal Court of Appeal — suggest significant simplification occurred.

To determine whether there was a statistically significant difference in analysis length before and after *Vavilov*, I used the Mann-Whitney U test.[86] The test found the differences in both courts statistically significant.[87] However, the effect size is modest: in the Federal Court, $r = 0.27$ approaches a medium effect, and in the Federal Court of Appeal, $r = 0.07$ indicates a small effect. Generally, 0.1 is considered a small effect, 0.3 is a medium effect, and 0.5 is considered to be a large effect.[88]

These differing effect sizes reflect the distinct trajectories of the two courts. Their modest effect sizes — despite the large differences in mean and median values — demonstrate that *Vavilov* is just one of many factors influencing analysis length. This reality becomes even more apparent when the mean and median analysis lengths are viewed on an annual basis.

Figure 1 plots the mean and median standard of review analysis lengths on a time series. The data shows that median analysis length in the Federal Court has consistently, albeit gradually, declined since 2008. The mean analysis length also follows the same general trend of decline, with two exceptions: increases from 2012 to 2014 and the 12-month period after *Vavilov* was released.

---

84 *Agraira v Canada (Public Safety and Emergency Preparedness)*, 2023 SCC 36.

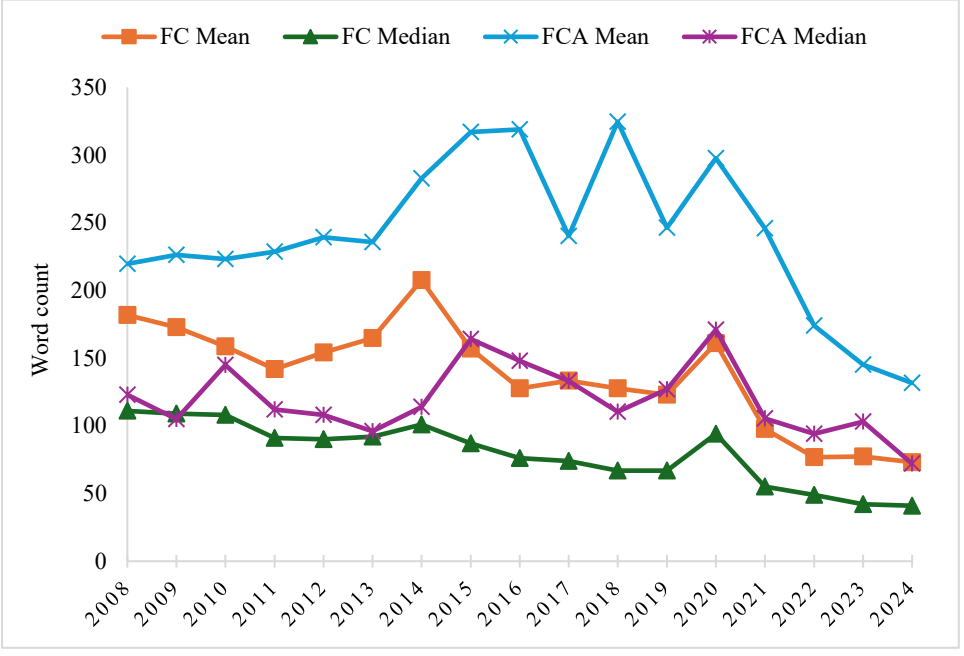
85 This figure excludes all cases where there was no analysis; in other words, cases where the analysis length was zero.

86 The data was left-skewed (most word counts falling between one and 150 words in the FCA and between one and 110 in the FC, but with some into the thousands). The Mann-Whitney U does not require a normal distribution and tests whether one group tends to have systematically higher or lower values than the other by comparing the ranks of all observations rather than their actual values. I performed the test using the Real Statistics Excel Add-in: [perma.cc/45PZ-S8Q2].

87 The resulting test statistic for the FCA was $U = 141{,}404$ and the p-value was 0.0129 (two tail); for the FC the test statistic was $U = 17{,}795{,}063$ and the p-value was 0.001.

88 Charles Zaiontz, "Mann-Whitney Test for Independent Samples", online: [perma.cc/ZGK2-LQUX].

FIGURE 1:
LENGTH OF STANDARD OF REVIEW ANALYSIS IN WORDS



One may ask why standard of review analysis length was decreasing in the years leading up to *Vavilov*, a time when academic and judicial criticism of the *Dunsmuir* framework was reaching its peak. If the doctrine was truly as confused and unworkable as critics claimed, why were Federal Court judges writing shorter, not longer, analyses?

I would suggest that the Federal Court was effectively insulated from the broader doctrinal problems due to the repetitive nature of its caseload — particularly immigration cases, with limited appeals — which enabled the Court to develop internal consensus on standard selection that operated independently of Supreme Court doctrine. Two metrics strongly support this theory. First, the Federal Court virtually abandoned the four-factor contextual analysis that was supposedly central to *Dunsmuir*. The rate of performing this analysis peaked in 2008 at 10 percent, steadily falling to just 1 percent by the time of *Vavilov*. More tellingly, the rate of relying on Federal Court or Federal Court of Appeal precedent in determining the standard increased from 66 percent to 80 percent by the time of *Vavilov* — suggesting the Court was essentially creating its own internal doctrine. This pattern suggests that institutional practice can diverge significantly from formal doctrine, with courts finding practical solutions that bypass rather than resolve theoretical problems.

This institutional learning pattern is further evidenced by the courts' response to *Vavilov* itself. A spike in mean and median analysis lengths in both courts occurred in 2020. If *Vavilov* simplified the standard of review analysis, why do we see a significant and immediate increase in the length of analysis? Closer examination of the 2020 cases reveals that the increases stem from judges explaining the new *Vavilov* framework in their reasons. The next

year, all four trend lines show a precipitous drop-off. This pattern reveals how legal knowledge becomes encoded within the judicial community — what initially required extensive explanation becomes so familiar that judges view detailed exposition as unnecessary, having already explained *Vavilov* numerous times in previous decisions. Although this study is limited to the federal courts, I would hypothesize the provincial superior courts that do not frequently see the same volume of administrative law cases might have a longer period of explanation. Presumably because administrative law is such a significant part of the federal courts' workload, they felt comfortable justifying their decisions without explaining *Vavilov*.

Whereas the Federal Court saw consistent decreases in analysis length pre-*Vavilov*, the Federal Court of Appeal analysis lengths are more consistent with the *Dunsmuir*-era problems that critics identified. Between 2008 and 2016, the mean standard of review analysis length was increased (as one would expect with the increasing complexity and uncertainty). However, the data becomes puzzling in the final years before *Vavilov*. The mean sharply fell in 2017, then rebounded in 2018 before falling again in 2019. The median analysis length did not exhibit the same volatility in the 2017 to 2019 period, indicating that the mean was being skewed by outlier decisions with long standard of review analyses.

My initial hypothesis was that the 2018 outliers would be judges' attempts to influence the development of standard of review doctrine, given the Supreme Court's announced intention to reconsider the law in *Vavilov*. However, closer examination of the longest decisions revealed this was not the case. The longest analyses were standard administrative law on topics like true questions of jurisdiction.[89]

The Federal Court of Appeal's response to *Vavilov* followed a similar pattern to the Federal Court. Analysis length initially increased in 2020 as judges explained the new framework. However, since that time, the mean length has fallen dramatically. More significantly, the mean and median values are now converging, indicating the virtual elimination of outlier cases that previously required extensive analysis.

The combined evidence from both courts reveals *Vavilov*'s ultimate success in simplification, despite different pre-2020 trajectories. While the Federal Court showed steady pre-*Vavilov* improvement and the Federal Court of Appeal experienced volatility, both courts have now achieved uniform declines in analysis length. Notably, 2024 produced the lowest mean and median analysis statistics in both courts. The median length of analysis in the Federal Court is now just 41 words. For perspective, a 41 word analysis looks like this: "The sole issue for determination is whether the PRRA Officer's decision was reasonable. Reasonableness is the presumptive standard of review of the merits of an administrative decision. None of the circumstances warranting a departure from this presumption arise in this case."[90] This demonstrates that standard of review determination has become genuinely routine for the median case.

This routinization of standard of review selection is further reflected in how the parties themselves approach these issues. When both parties agree on a standard of review, one party

---

89 See e.g. *Laurentian Pilotage Authority v Corporation des Pilotes du Saint-Laurent Central Inc*, 2018 FCA 117; *Bell Canada v 7262591 Canada Ltd*, 2018 FCA 174.

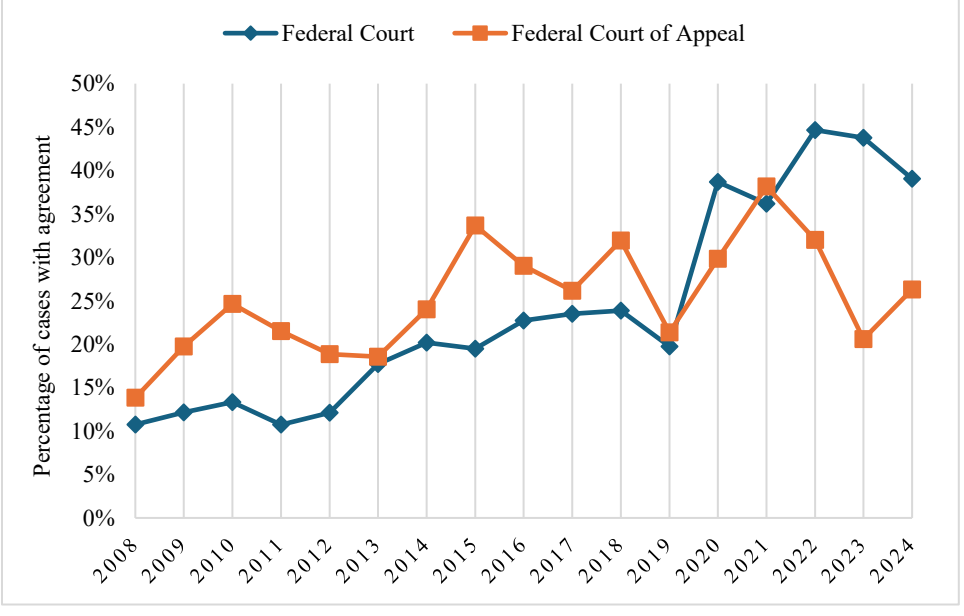
90 *Nansobya v Canada (Immigration, Refugees and Citizenship)*, 2024 FC 1049 at paras 8–9.

is conceding to a standard that is less beneficial to them. This concession suggests that the law is so clear that arguments to the contrary are not a worthwhile use of advocacy resources or credibility capital. Thus, I would hypothesize that if *Vavilov* successfully simplified standard of review selection, the percentage of cases where there is consensus on the standard of review would increase.

The optimal way to assess the rate of party agreement would be to assess the parties' submissions directly. As the data for this study comes from reasons for decision, I am limited to instances of agreement that are specifically highlighted by judges. Nonetheless, the results are striking. In the Federal Court, the percentage of cases where the Court explicitly acknowledged that the parties agreed on standard of review rose from 17 percent in the pre-*Vavilov* period to 41 percent. The Federal Court of Appeal also saw an increase, albeit to a lesser degree, from 24 percent to 30 percent.

Figure 2 plots the agreement rate over time. Like word count, the post-*Vavilov* improvements are part of a longer-term trend, with a general upward trend in both courts. Nonetheless, 2020 marks a stark jump with a 19 percent increase in the Federal Court.

FIGURE 2:
PARTIES EXPRESSLY AGREE ON STANDARD OF REVIEW



In contrast to party agreement, we can also look at judicial disagreement — the rate at which the judges on a panel disagree on the appropriate standard of review. The rates of disagreement are very low in the Federal Court of Appeal. Only 7 percent of cases have more than one set of reasons, and usually the division is not related to standard of review. All ten of the cases that involved standard of review disagreements were in the pre-2020 period. Given the small numbers involved, this difference may appear negligible, but the pattern is consistent with the broader simplification trends.

A final metric that I used to track complexity was avoidance. Courts may avoid determining the standard of review when the analysis is complex, uncertain, or when they anticipate the determination would be controversial or difficult to justify. This avoidance allows judges to sidestep thorny doctrinal questions while still reaching a decision on the merits. A court may avoid determining standard of review analysis in two ways. First, it may explicitly find that no conclusion on standard of review is necessary because the result would be the same under any standard. Second, it may ignore standard of review altogether and instead use language other than reasonableness or correctness.

The data reveals declining avoidance rates in both categories. Explicit refusal to determine the standard of review dropped modestly in both courts — from 3 percent to 1 percent in the Federal Court and from 6 percent to 4 percent in the Federal Court of Appeal. More significantly, implicit avoidance fell substantially: cases where the LLM could not identify any standard of review decreased from 17 percent to 10 percent in the Federal Court and from 37 percent to 24 percent in the Federal Court of Appeal. These reductions suggest that *Vavilov* has made standard of review determination sufficiently straightforward such that courts less often feel the need to circumvent the analysis entirely.

Multiple metrics thus confirm that *Vavilov* has achieved meaningful simplification of standard of review analysis. Beyond the dramatic reduction in analysis length, party agreement rates have more than doubled in the Federal Court (from 17 percent to 41 percent) and both courts were less likely to avoid the standard of review analysis altogether. While the Federal Court and Federal Court of Appeal followed different pre-*Vavilov* trajectories, both have converged on streamlined approaches that treat standard of review selection as routine rather than complex.

B. *VAVILOV*'S EFFECT ON STANDARD OF REVIEW

Beyond simplifying the standard of review analysis, we may also ask whether the *Vavilov* framework has altered the substantive results of that analysis — that is, which standards the courts ultimately apply. As discussed above, Green tracked the standard of review selected in the Federal Court from just before *Dunsmuir* to just after *Vavilov*. To isolate the effect of the Supreme Court's doctrinal changes from other factors, he performed a logistic regression comparing pre-*Dunsmuir* decisions with post-*Vavilov* decisions. The regression results showed that both the nature of the question and the post-*Vavilov* period were significant predictors of reasonableness selection. Discretion and fact-based issues were strongly associated with reasonableness review, while the post-*Vavilov* period independently made judges about 30 percent more likely to choose reasonableness compared to the pre-*Dunsmuir* era. By contrast, the type of decision-maker and area of law had no significant effect on standard selection.[91]

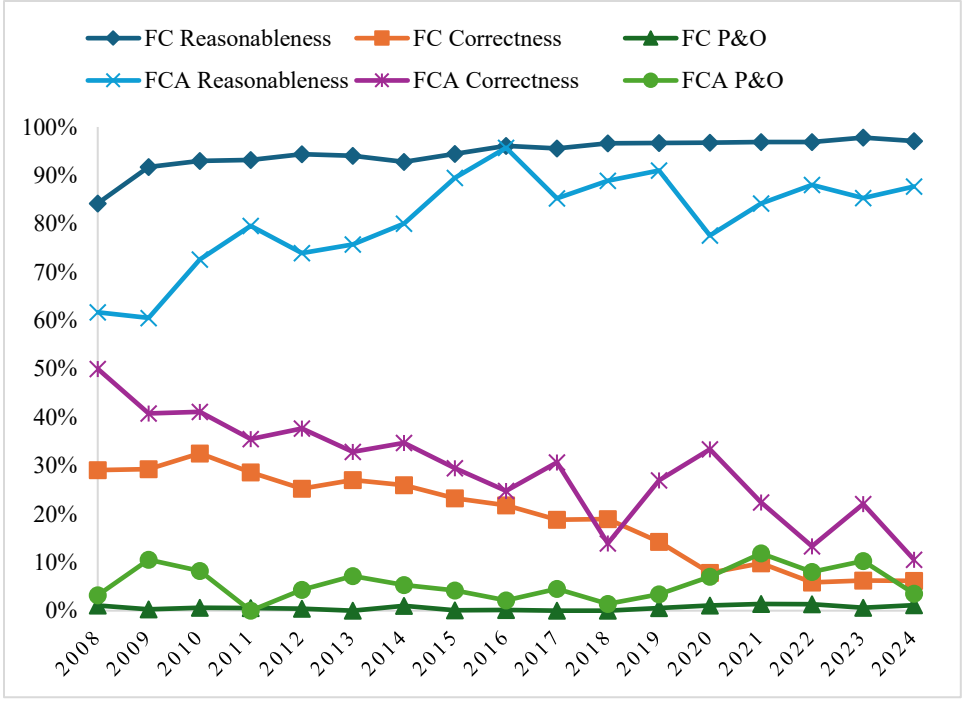
However, when Green narrowed his analysis and performed separate regressions to isolate the individual effects of *Dunsmuir* and *Vavilov*, he found that only *Dunsmuir* had a statistically significant impact on standard selection. *Vavilov*, despite the Supreme Court's major doctrinal reforms, showed no measurable independent effect. Green hypothesized that

[91] Green, "Groundbreaking Cases", *supra* note 2 at 446–47.

this counterintuitive result occurred because reasonableness review had already become dominant by 2019.[92]

Figure 3 displays annual trends in standard of review selection in my dataset. It shows the percentage of decisions that applied reasonableness, correctness, or palpable and overriding error (P&O) to any issue in the case. It confirms Green's findings regarding the dominance of reasonableness by the time of *Vavilov*.[93]

FIGURE 3:
RATE OF STANDARD OF REVIEW SELECTION



To isolate *Vavilov*'s specific impact, I analyzed changes in reasonableness selection rates in the five years before and after the decision. The Federal Court applied the reasonableness standard to at least one issue in 96 percent of cases between 2015 and 2019. This rose slightly to 97 percent in the 2020 to 2024 period. The Federal Court of Appeal applied reasonableness

[92] *Ibid* at 454.

[93] Although Green coded the standard of review differently — allowing only for one standard per case (reasonableness, correctness, mixed, or none) — the results of this study closely track his. When adjusted to match Green's coding methodology (reasonableness-only cases), the comparative rates are: 65.3 percent (Green: 64 percent), 2012: 74.1 percent (Green: 76.9 percent), 2019: 84.6 percent (Green: 90.2 percent) and 2020: 91.3 percent (Green: 88.3 percent). For correctness review, 2008: 10.9 percent (Green: 16.3 percent), 2012: 5.3 percent (Green: 8.4 percent), 2019: 2.4 percent (Green: 5.8 percent) and 2020: 1.3 percent (Green: 5.2 percent). Green's consistently higher correctness rates likely reflect the inclusion of procedural fairness-only cases, which I excluded from this study.

to at least one issue in 90 percent of cases between 2015 and 2019, with the same figure dropping to 85 percent after *Vavilov*.

The picture changes slightly when examining cases that applied reasonableness exclusively. Here, the Federal Court showed substantial improvement from 80 percent to 92 percent. However, the Federal Court of Appeal continued to decline, dropping from 72 percent to 66 percent.

These contrasts between the Federal Court and the Federal Court of Appeal are the product of *Vavilov*'s instruction that statutory appeals should proceed according to the *Housen* standards of correctness or palpable and overriding error.[94] Statutory appeals are more common in the Federal Court of Appeal (6 percent of cases) than the Federal Court (less than 2 percent). When I removed statutory appeals from the analysis, and only included judicial reviews, the rate of exclusive reasonableness review increased after *Vavilov* (from 62 percent to 64 percent) as did reasonableness applied to at least one issue (77 percent to 82 percent). Alternatively, when we consider judicial reviews and appeals together, but group both deferential standards (reasonableness and palpable and overriding error), applying at least one deferential standard remained constant at 93 percent of Federal Court of Appeal decisions before and after *Vavilov*, while applying only deferential standards rose from 68 percent to 73 percent.

To test whether these differences were statistically significant, I performed two-proportion Z-tests.[95] The differences in rates of reasonableness review in the Federal Court were statistically significant at the 99 percent confidence level. In the Federal Court of Appeal, only the decrease in reasonableness applied to at least one issue reached statistical significance (at a 95 percent confidence interval).

The statistical analysis reveals that *Vavilov*'s impact on standard selection has been modest and uneven across courts, largely confirming that reasonableness review was already dominant by 2019. An equally significant question is whether *Vavilov* has altered the practical consequences of judicial review, a question to which I now turn.

C. *VAVILOV*'S EFFECT ON OUTCOME

The most surprising conclusion from Green's study was that standard of review did not appear to have a significant effect on case outcome; in other words, whether the judicial review was granted or statutory appeal allowed. Despite the dramatic increase in the use of reasonableness review — from roughly 60 percent pre-*Dunsmuir* to 90 percent post-*Vavilov* — case outcomes remained stable. Federal Court judges granted judicial review at essentially the same rate they did before *Dunsmuir*. Thus, when Green ran a logistic regression, the use of a reasonableness or patent unreasonableness standard was not a statistically significant predictor of outcome.[96]

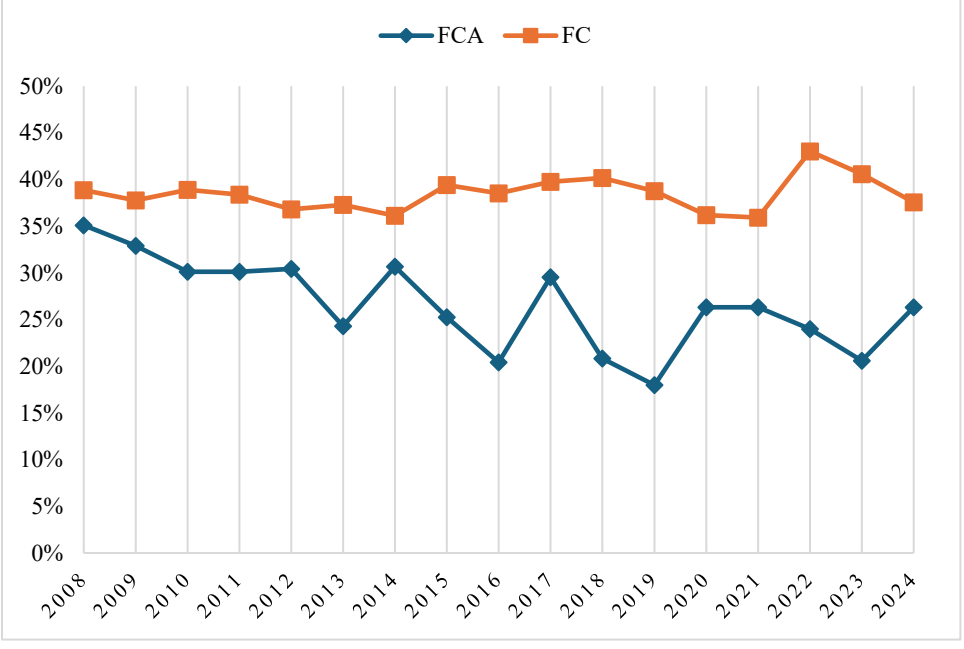
Figure 4 plots the rate of granting judicial review and allowing statutory appeals over time for both the Federal Court and Federal Court of Appeal based on the present study's data. The

---

94 *Vavilov*, *supra* note 1 at para 37, citing *Housen*, *supra* note 3.
95 Analysis was performed in Excel using the Real Statistics Resource Pack, *supra* note 86.
96 Green, "Groundbreaking Cases", *supra* note 2 at 447.

observed grant rate tracks closely with the rate found in Green's study, with the percentage of cases where judicial review was granted hovering just under 40 percent for the entire period.[97] The before and after *Vavilov* figures are 38.4 percent and 39.5 percent, respectively. It is notable, however, that 2022 (after Green's study period) saw an unusual jump to 45 percent, which has slowly trended back down to 39 percent in 2024.

FIGURE 4:
RATE OF GRANTING JUDICIAL REVIEW



The Federal Court of Appeal data tells a different story. Rather than stability, it shows a downward trend in grant rates, declining from 35 percent in 2008 to 20 percent in 2019, rebounding slightly to 26 percent in 2024. This counterintuitive pattern — where the Federal Court of Appeal applies correctness review at three times the rate of the Federal Court, yet quashes administrative decisions less frequently — suggests that the relationship between standards of review and outcomes is more nuanced than traditional doctrine assumes.

A few factors may explain this divergence. First, a large portion of the Federal Court's docket consists of immigration cases that are pre-screened through the leave process, filtering out frivolous claims before they reach the review stage. Second, the Federal Court of Appeal hears many appeals from Federal Court decisions. I suspect applicants are less disciplined in their appeals than the Attorney General, which (if true) leads to the Federal Court of Appeal disproportionately seeing weaker cases that have already been rejected at the Federal Court.

[97] The comparative figures are 2008: 38.8 percent (Green: 43 percent), 2012: 36.8 percent (Green: 38.8 percent), 2019: 38.8 percent (Green: 38.4 percent) and 2020: 38.6 percent (Green: 40.9 percent). Differences may be due to the fact that my figures represent only cases where there is a standard of review analysis or clear standard of review.

The Federal Court of Appeal's different pattern of granting judicial review inspired me to run logistic regressions to test whether there is a relationship between standard of review and outcome, and whether that relationship has changed post-*Vavilov*. I ultimately ran two logistic regressions, one for each court. A logistic regression is the appropriate form of regression analysis when predicting a binary outcome — here, whether or not a judicial review is granted.[98]

The regression equation here is:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1(\text{Post-}\textit{Vavilov}) + \beta_2(\text{Deferential SOR}) + \beta_3(\text{Deferential} \times \text{Post-}\textit{Vavilov})$$

where p represents the probability of granting the judicial review or statutory appeal, and the β coefficients measure how each factor affects this probability. The first independent variable captures when the case was decided. It was coded 1 where the decision was from 2020 or later (that is, post-*Vavilov*), and 0 for earlier decisions. The second independent variable captures the standard of review applied and was coded 1 where only deferential standards were applied in the case, and 0 where any non-deferential standards were used. The third variable represents the interaction between the post-*Vavilov* period and deferential review and was coded 1 for post-2020 cases that applied only deferential standards of review. The purpose of the interaction variable is to capture whether the effect of applying a deferential standard differs between the pre- and post-*Vavilov* periods. As *Vavilov* altered the methodology of reasonableness review, it tests whether deference operates differently after *Vavilov*.

Table 1 displays the results, conveyed in marginal effect. The marginal effect tells us how much the probability of granting judicial review changes when we alter one variable, holding other variables constant, relative to the baseline of pre-*Vavilov* cases applying correctness review. In both courts, the regression results show that standard of review significantly affects outcome and *Vavilov* did not change this relationship.

TABLE 1:
LOGISTIC REGRESSION RESULTS

Variable	**Federal Court**	**Federal Court of Appeal**
	Marginal effects	Marginal effects
Post-*Vavilov*	−0.016 (0.027)	−0.043 (0.066)
Deferential SOR	−0.085*** (0.011)	−0.184*** (0.034)
Interaction	0.042 (0.029)	0.052 (0.078)
Observations	15,573	1,321

Table 1: Post-*Vavilov* is a dummy for cases decided from 2020 to 2024. Deferential indicates judge only applied one or more deferential standards (reasonableness and/or palpable and overriding error). Interaction tests whether the effect of reasonableness changed post-*Vavilov*. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Notes: Standard errors in parentheses. The dependent variable is Grant (whether judicial review was granted).

When a court is performing a non-deferential review, the model sets the probability of granting the judicial review in the Federal Court at 44.5 percent. In the Federal Court of

[98] David Rindskopf, "Trends in Categorical Data Analysis: New, Semi-New, and Recycled Ideas" in David Kaplan, ed, *The SAGE Handbook of Quantitative Methodology for the Social Sciences* (Thousand Oaks, California: SAGE Publications, 2004) 137 at 144.

Appeal, the model sets the probability at 36.9 percent. Those probabilities shift to 36.3 percent and 21.7 percent respectively when a deferential standard is applied.

These results contradict Green's finding that standard of review is not a statistically significant predictor of outcome in the Federal Court. Several methodological differences may explain the divergent findings, though the significance of these differences remains unclear.[99] However, one thing that is clear is that the model fit statistics in Appendix A1 show that my logistic regression has very poor predictive power. The pseudo R-squared values indicate the models explain less than 1 percent and 3 percent of variance in the Federal Court and Federal Court of Appeal respectively. Similarly, the area under the curve values of 0.53 (Federal Court) and 0.58 (Federal Court of Appeal) barely exceed random chance (0.50).[100] The regression thus contradicts Green's finding regarding statistical significance, but affirms his broader conclusions that standard of review alone cannot meaningfully predict case outcomes and that *Vavilov* has not altered the fundamental relationship between standards of review and case outcomes.

IV. CONCLUSION

With this study, I aimed to provide a comprehensive empirical assessment of *Vavilov*'s impact on standard of review analysis after five years in action. The data confirms that *Vavilov* has delivered on its central promise to simplify standard of review analysis. The dramatic reduction in analysis length — 48 percent in the Federal Court and 31 percent in the Federal Court of Appeal — will likely translate into tangible benefits across the legal system: reduced costs for litigants, more efficient judicial decision-making, and clearer guidance for administrative tribunals.

Yet the data also hints that the story of *Vavilov* is one of evolution rather than revolution. The federal courts had already begun simplifying their approach years before the Supreme Court of Canada intervened, with analysis length declining steadily from 2008 onward. By 2019, these courts had largely abandoned *Dunsmuir*'s complex contextual analysis in favor of precedent-based decision-making. This finding carries important implications for how we understand legal reform. The Supreme Court's repeated interventions in administrative law — from *CUPE* through to *Pushpanathan*, *Dunsmuir*, and now *Vavilov* — have assumed that clearer doctrinal frameworks will reshape judicial behavior. These findings suggest a more complex reality: lower courts develop their own institutional practices that may diverge from Supreme Court doctrine.

The study also reveals the limits of doctrinal reform in other ways. Despite dramatic changes in standard selection, grant rates remain stable and the relationship between standards and outcomes persists unchanged. While applying exclusively deferential standards reduces the probability of quashing an administrative decision by 8.5 percent in the Federal

---

[99] Methodological differences include: this study's substantially larger sample size (over 15,000 cases versus 1,076), which provides greater statistical power; Green's use of judge clustering and robust standard errors, which represents more sophisticated modelling; his straddle methodology focusing on immediate before-and-after effects (ending just after *Vavilov*) versus this study's 2008-2025 timeframe; finally, his exclusion of the palpable and overriding error standard.

[100] As Green does not provide any measures of goodness of fit, we cannot tell whether his model achieved better predictive power than the present analysis.

Court and 18.4 percent in the Federal Court of Appeal, this relationship predates *Vavilov* and appears unaffected by the Supreme Court's doctrinal reforms. These findings underscore that while new judicial frameworks clearly influence judicial behaviour, they operate within complex institutional environments where deeper structural forces continue to shape case results regardless of doctrinal changes.

Vavilov has given Canadian administrative law a simpler framework, reducing transaction costs and increasing certainty. The first five years of data in the federal courts support the idea that *Vavilov* has achieved durable simplification without triggering the doctrinal battles that plagued its predecessors. But *Dunsmuir*'s first five years were far more promising than its last five.

V. APPENDIX

Table A1 provides the complete statistical output from the logistic regression analyses discussed in the main text. The coefficient values represent the change in log-odds for each variable. The standard errors (shown in parentheses) measure the precision of these estimates — smaller standard errors indicate more precise estimates. The p-values indicate the probability that the observed relationship occurred by chance.

Pseudo R^2 measures how well the model explains the variation in outcomes, similar to R^2 in linear regression. This pseudo R^2 was calculated using Nagelkerke's method. Values closer to 1.0 indicate better explanatory power, while values near 0 suggest the model explains little variance. AUC (area under the curve) measures the model's capacity to distinguish between cases where judicial review is granted versus denied. An AUC of 0.50 means the model predicts at a rate equal to random chance. Classification accuracy shows the percentage of cases the model correctly predicts. However, in this case, the figures are misleading because both models classified all cases as "denied" and achieving accuracy rates identical to the rate of granting judicial review in the two courts.

TABLE A1:
LOGISTIC REGRESSION PREDICTING THE GRANTING OF A JUDICIAL REVIEW OR ALLOWING OF A STATUTORY APPEAL

Variable	**Federal Court of Appeal**			**Federal Court**		
	Coefficient (Standard Error)	***p-value***	**Odds Ratio**	**Coefficient (Standard Error)**	***p-value***	**Odds Ratio**
Intercept	-0.535 (0.109)	1E-06	0.586	-0.220 (0.039)	2E-08	0.803
Post-Vavilov	-0.180 (0.277)	0.5155	0.835	-0.065 (0.109)	0.55027	0.937
Deferential SOR	-0.749 (0.146)	3E-07	0.473	-0.342 (0.046)	7.7E-14	0.710
Interaction	0.218 (0.328)	0.5062	1.244	0.169 (0.115)	0.14306	1.184
Observations	1,321			15,573		
Pseudo R^2	0.033			0.005		
AUC	0.584			0.531		
Classification accuracy	73.4%			61.4%		