Monograph Series

Volume 13—Examining Legal Writing Empirically

This article was originally published with the following citation:

Mike Madden, *Stating It Simply: A Comparative Study of the Quantitative Readability of Apex Court Decisions from Australia, Canada, South Africa, the United Kingdom, and the United States*, 23 N.C. J.L. & TECH. 270 (2021).

NORTH CAROLINA JOURNAL OF LAW & TECHNOLOGY
VOLUME 23, ISSUE 2: DECEMBER 2021

STATING IT SIMPLY: A COMPARATIVE STUDY OF THE QUANTITATIVE READABILITY OF APEX COURT DECISIONS FROM AUSTRALIA, CANADA, SOUTH AFRICA, THE UNITED KINGDOM, AND THE UNITED STATES

Mike Madden*

Even though common law courts create and articulate the law within their decisions, surprisingly little is known about the quantitative readability levels of any single national apex court's decisions, and even less is known about how any one apex court's readability levels compare to those of other similar apex courts. This Article offers new data and analysis that significantly reduces the blind spots in these areas by reporting the results of an original empirical study of the readability of judicial decisions released in 2020 from the apex courts of five English-speaking jurisdictions.

This Article draws on applied linguistics theory and Natural Language Processing techniques in order to provide both uni- and multi-dimensional readability scores for the 233 judicial decisions (comprising more than 3 million words of text) that form the corpus of this study. The results show that readability levels vary by approximately 50% between the most- and least-readable jurisdictions (the United States and Australia, respectively). This Article then analyzes the data comparatively in order to determine whether institution- or jurisdiction-specific factors are capable of explaining readability variances between the different courts. This

* Ph.D. Student, University of Ottawa Faculty of Law, and graduate student member of the Centre for Law, Technology and Society at the Faculty of Law at the University of Ottawa in Ottawa, Canada. The author wishes to thank Professor Nina Varsava, University of Wisconsin-Madison Law School, for her helpful feedback on an earlier draft of this Article. The author also wishes to thank Dr. Elizabeth F. Judge for her feedback on this Article, and her exceptional Ph.D. supervision more generally. Finally, the author is grateful for the generous support of the Social Sciences and Humanities Research Council of Canada (SSHRC), whose award of a Joseph-Armand Bombardier Canada Graduate (Doctoral) Scholarship facilitated this research.

Article concludes that certain comparative factors, such as the average panel size used by each court and the ratios of both former law professors and women who sit on panels in each jurisdiction, can explain 23.7% of the total variances in readability scores. These findings may help judicial branch and executive branch decision-makers better understand how their court's decisions "stack up" against other courts in terms of readability and offer insights into how readability levels could be enhanced.

TABLE OF CONTENTS

I. INTRODUCTION

Legal theorists and rule of law scholars generally agree that one requirement of a functioning legal system is that the system's laws must be knowable to those governed by the law.[1] In jurisdictions wherein 'law' comprises not only legislation, but also common law principles and rules that courts have set down in their judicial decisions, stakeholders may find it important to consider the extent to which the common law is ascertainable to the population. Many common law judges and scholars explicitly acknowledge their obligations to produce readable, accessible decisions and recognize that their audiences are broad: judges must not only communicate to the specific litigants, one another, and the legal profession, but also

[1] LON L. FULLER, THE MORALITY OF LAW 39 (rev. ed. 1969); JOSEPH RAZ, THE AUTHORITY OF LAW: ESSAYS ON LAW AND MORALITY 213–15 (1979); JOHN RAWLS, A THEORY OF JUSTICE 209 (1999).

to the citizenry as a whole.[2] Additionally, scholars are beginning to note how an inability to read relevant legal materials (like case law) can present significant "access to justice" barriers for citizens generally, but in particular, for self-represented litigants, who need to understand the law that applies to their cases.[3] Although one could perhaps argue for *less*-publicly understandable laws,[4] this Article begins from the assumption (grounded in "rule of law" theory)[5] that it is inherently beneficial for the common law to be more—rather than less—readable, and that, in any case, there is value in knowing how readable judicial decisions are as a baseline fact.

How well are common law courts actually achieving the objective of producing readable statements of the law? Perhaps surprisingly, very little effort appears to have been made toward answering this question, at least from an empirical or quantitative perspective. This Article makes several further steps in that direction by reporting the results of an original comparative readability study that measures the quantitative readability of apex court decisions released between January 1, 2020, and December 31, 2020, from Australia (N = 46), Canada (N = 34), South Africa (N = 30), the United Kingdom (N = 54), and the United States (N = 69), using a variety of applied linguistics metrics. Additionally, by examining or measuring how environment- and court-specific factors also differ

---

[2] *See*, *e.g.*, Gerald B. Wetlaufer, *Rhetoric and Its Denial in Legal Discourse*, 76 VA. L. REV. 1545, 1561 (1990); *see The Honourable Nicholas Kasirer's Questionnaire (Questionnaire for the Supreme Court of Canada Judicial Appointment Process),* OFF. OF THE COMM'R FOR FED. JUD. AFF., Part 10(4) (Apr. 18, 2019), https://www.fja.gc.ca/scc-csc/2019/nominee-candidat-eng.html [https://perma.cc/4MVT-MSNB]; *see also The Honourable Justice David M Paciocco's Questionnaire (Questionnaire for Judicial Appointment)*, GOV'T OF CAN., Part 11(4) (Apr. 7, 2017) http://www.canada.ca/en/department-justice/news/2017/04/the_honourable_justicedavidmpacioccosquestionnaire.html [https://perma.cc/6TG8-JBCN].

[3] Patricia Hughes, *Advancing Access to Justice through Generic Solutions: The Risk of Perpetuating Exclusion*, 31 WINDSOR Y.B. ACCESS JUST. 1, 13–15 (2013).

[4] *See* Rabeea Assy, *Can the Law Speak Directly to Its Subjects - The Limitation of Plain Language*, 38 J.L. & SOC'Y 376 (2011) (discussing how efforts to make the law more readable may compromise legal clarity and precisions, ultimately arguing that it is futile to hope that the law can be broadly accessible, without the assistance of legal professionals, to average citizens).

[5] FULLER, *supra* note 1, at 39.

across the five studied jurisdictions, this Article assesses how these comparative factors may explain readability variances across the jurisdictions.

The results of this study show that there are substantial variances in quantitative readability levels across the five jurisdictions, of approximately 50% in mean readability levels between the most- and least-readable jurisdictions, based on two different comprehensive readability formulae.[6] Furthermore, these results show that statistically significant correlations of moderate effect size exist between a decision's readability level and the number of judges on the panel,[7] the number of former law professors on the panel,[8] and the number of women on the panel.[9] Using these three factors as independent variables and the decision readability scores as the dependent variable, a multiple regression analysis yielded a statistically significant model capable of explaining 23.7% of the variance in readability scores.[10] In other words, jurisdiction-specific and court-specific factors that are particular to the different apex courts accounted for almost one quarter of the readability differences in decisions from these courts.

The results of this study can assist legal scholars in better understanding some of the factors that may be influencing readability levels of apex court decisions within a given jurisdiction. This understanding may, in turn, permit key officials within governments (such as those responsible for appointing judges) and judicial executives (such as Chief Justices who assign judges to preside on particular panels) to act in ways that could support the production of more readable judicial decisions in the future. Ultimately, however, the results of this study offer only a partial explanation for variations in readability levels across jurisdictions and suggest that other factors (perhaps more related to the identity of the author of a decision than to the court or jurisdiction from

[6] *See infra* Table 5 and associated text.
[7] *See infra* Table 6 and associated text.
[8] *See infra* Table 7 and associated text.
[9] *See infra* Table 9 and associated text.
[10] *See infra* Table 10 and associated text.

which the decision emerges) may better explain the readability variances.

This Article begins in Part II with a brief discussion of the relevant background and context of the current study. Part III describes the design and method of the study, including its limitations. Part IV presents and discusses the results of the study, including measurements and observations of the studied apex courts and their operating environments, as well as both descriptive and analytical statistics relating to readability measures of the decisions produced by these courts. This Article concludes by reiterating its principal finding: readability differences between apex courts in different countries can be explained, in part, by reference to comparative factors; however, future studies that focus more on characteristics of the individual authors of judicial decisions, as opposed to characteristics of their working environments, could provide a better understanding of additional sources of variance.

II. BACKGROUND & CONTEXT: READABILITY, LAW, AND THE COMPARATIVE METHODOLOGY

In order to situate the present study within its appropriate context, some explanations about the concept of quantitative readability and its relevance to the law would be helpful. It is also important to understand how a comparative methodology can be applied to this topic in order to generate a set of useful findings, as well as how such a methodology will differ in its focus and results from a more "law and language" methodology. Consequently, this Part: (A) introduces the concept of readability; (B) summarizes the existing literature that focuses on the readability of court decisions; and, (C) explains how a comparative methodology can be employed in order to increase our understanding of apex court readability levels.

A. Understanding Quantitative Readability

Readability, for the purposes of this study, refers to text-centered assessments of how easy or difficult, from a cognitive perspective, texts are to read and understand for non-specific readers. This sense of the term "readability" is consistent with how the term is used within the fields of education and linguistics, where "readability"

has been defined as the quality that "makes some texts easier to read than others"[11] and as "the sum total (including all the interactions) of all those elements within a given piece of printed material that affect the success a group of readers have with it."[12]

Readability, here, is not concerned with text legibility (i.e., the extent to which the visual layout and presentation of a text facilitate reader processing of that text), even though text legibility likely has an impact on the physical or optical level of effort involved in reading.[13] Instead, the readability dimensions of the present study focus on the relative level of cognitive effort that a generic reader would require to understand particular texts.

In this sense, readability is an objective but somewhat abstract concept, offering a relative and general measure of how comprehensible a text will be; readability does not, for instance, account for reader-specific factors, such as a reader's interest in the text, education level, or familiarity with the subject matter. Readability, then, says nothing about whether a particular individual will actually understand a particular text; instead, readability determines, in broad terms, whether more or fewer people are likely to understand a specific text based on the language-related properties of that text. Readability is therefore a useful concept when considering texts that are intended for broad, heterogeneous groups, or for groups whose characteristics are fluid or otherwise not well-understood, because readability measurements should determine whether one text is more or less likely to be understood than another text—even if not much is known about the world of potential readers of the texts. Since judicial decisions, written for

---

[11] WILLIAM H. DUBAY, THE PRINCIPLES OF READABILITY 3 (2004).

[12] Edgar Dale & Jeanne S. Chall, *The Concept of Readability*, 26 ELEMENTARY ENG. 19, 23 (1949).

[13] *See* Mark Sableman, *Typographic Legibility: Delivering Your Message Effectively*, 17 SCRIBES J. OF LEGAL WRITING 9, 15–17 (2017); *see also* Mary Alton Mackey & Marilyn Metz, *Ease of Reading of Mandatory Information on Canadian Food Product Labels*, 33 INT'L J. CONSUMER STUD. 369, 371–72 (2009) (noting how typeface, color, contrast levels, and text placement can each impact elements of readability); *see generally* Khaled Moustafa, *Improving PDF Readability of Scientific Papers on Computer Screens*, 35:4 BEHAVIOUR & INFO. TECH. 319 (2016) (describing how the column-based display of text within PDF files on computer screens can inhibit readability of the text).

broad and diverse audiences, fall within this category of text, there are clear benefits to studying the readability levels of these texts.

Readability may be assessed quantitatively, qualitatively, or through some combination of these methods. For instance, in Crossley, Skalicky, and Dascalu's study, the authors first used crowd-sourced pairwise (qualitative) comparisons of side-by-side texts in order to determine the relative readability levels of approximately 600 texts.[14] Next, the authors (quantitatively) studied the linguistic properties of two-thirds of these texts to derive, through regression analysis, a readability formula that could accurately predict text readability.[15] Finally, the authors (quantitatively) tested their readability formula on the remaining one-third of the texts and found that their formula validly predicted the relative readability of these texts.[16] As this example illustrates, quantitative readability formulae can be anchored in real-world, human-involved, qualitative assessments of text readability (like a pairwise comparison, or a reader comprehension test)—although a quantitative readability formula can subsequently be applied on its own once this anchor has been established and once the formula has been validated.[17] The present study centers around the quantitative readability of apex court decisions. Accordingly, the remainder of the ensuing discussion focuses on the necessary background for understanding quantitative, rather than qualitative, readability concepts.

This Article does not explore the entire history of quantitative readability studies in English, but these studies have been numerous and varied. One of the earliest and most well-known readability

---

[14] Scott A. Crossley, Stephen Skalicky & Mihai Dascalu, *Moving Beyond Classic Readability Formulas: New Methods and New Models*, 42 J. RSCH. READING 541, 546–49 (2019) (conducting a study for the purpose of developing new readability models that identify linguistic features in texts that affect text comprehension and reading speed).

[15] *Id.* at 549–52.

[16] *Id.* at 551–57.

[17] See John C. Roberts, Robert H. Fletcher & Suzanne W. Fletcher, *Effects of Peer Review and Editing on the Readability of Articles Published in* Annals of Internal Medicine, 272:2 JAMA 119 (1994), for an example of such a study that relies on previously-validated readability formulae (concluding that peer review improves readability of manuscripts).

formulae to emerge from these studies was developed (or modified from its earlier form) in 1948 by Rudolph Flesch.[18] Flesch's "Reading Ease" formula used counts of average syllables per 100 words and average words per sentence, together with a constant, to give readability scores to texts on a scale of 1 to 100.[19] A score of 100 "corresponds to the prediction that a child who has completed fourth grade will be able to answer correctly three-quarters of the test questions to be asked about the passage that is being rated,"[20] while a score of 0 signifies a text that is "practically unreadable."[21] The Flesch Reading Ease formula continues to be used within popular word processing software applications, such as Microsoft Word and Google Docs.[22]

Many other formulae relying on similar analytical techniques to produce readability measurements were introduced after the Flesch Reading Ease formula between 1948 and 1995. Robert Gunning developed the Gunning FOG Index in 1952, which measured readability based on average sentence length, and average number of words with three or more syllables (subject to some limited exceptions) per 100 words.[23] The Automated Readability Index, developed in 1967, relied on measures of words per sentence, and characters per word, to produce an estimated reading grade level of a text.[24] The Flesch-Kincaid Grade Level Score similarly relied on measures of words per sentence and syllables per word to produce an estimated reading grade level of a text.[25] Dale and Chall

---

[18] Rudolph Flesch, *A New Readability Yardstick*, 32 J. OF APPLIED PSYCH. 221, 221–22 (1948).

[19] *Id.* at 228–29.

[20] *Id.* at 225.

[21] *Id.* at 229.

[22] *See, e.g.*, Richard Johnson, *How to Apply the Flesch Kincaid Readability Formula to Your Content*, OPTIMONK (Oct. 12, 2021), https://www.optimonk.com/how-to-apply-the-flesch-kincaid-readability-formula-to-your-content/ [https://perma.cc/5EP9-2T5F] (describing how to access this readability formula in both Microsoft Word and Google Docs).

[23] ROBERT GUNNING, THE TECHNIQUE OF CLEAR WRITING 35–37 (1952).

[24] R.J. SENTER & E.A. SMITH, AUTOMATED READABILITY INDEX 7–12 (Aerospace Med. Rsch. Lab'y's 1967).

[25] J. PETER KINCAID ET AL., NAVAL TECHNICAL TRAINING COMMAND, DERIVATION OF NEW READABILITY FORMULAS FOR NAVY ENLISTED PERSONNEL 14 (1975).

(originally in 1948, shortly after Flesch introduced his Reading Ease Formula) were unsatisfied with using word length as a proxy measurement for word difficulty and therefore developed a readability formula that relied on measures of average sentence length, as well as a ratio of difficult words in the text (i.e., words not listed within a 3,000-word list of commonly used words).[26]

As computing power increased and computers generally became more accessible, scholars began to employ (computer-based) Natural Language Processing ("NLP") techniques to derive and apply readability measures.[27] For instance, the freely-available, web-based Coh-Metrix 1.0 software tool was introduced in 2004 and was intended to analyze "texts on multiple levels of language, discourse, cohesion, and world knowledge."[28] The tool computed both the Flesch Reading Ease Score and the Flesch-Kincaid Grade Level Score, while also measuring many other novel dimensions of a text that were thought to be related to text comprehension and complexity, including scores related to lexical diversity (the ratio of unique words to total words),[29] word frequency (how commonly a word occurs in the English language, as assessed based on occurrences within large representative corpora of text),[30] concreteness ("how concrete or nonabstract a word is, on the basis of human ratings"),[31] and cohesion (the overlap of words or ideas across sentences, paragraphs, and the text as a whole).[32] The creators of Coh-Metrix 1.0 continued to refine the tool from 2002 to 2011[33] in an effort to better predict deep comprehension of texts, instead of the kind of "surface comprehension" that traditional readability

---

[26] Edgar Dale & Jeanne S. Chall, *A Formula for Predicting Readability*, 27 EDUC. RSCH. BULL. 11, 15–18 (1948).

[27] Arthur C. Graesser et al., *Coh-Metrix: Analysis of Text on Cohesion and Language*, 36 BEHAV. RSCH. METHODS, INSTRUMENTS, & COMPUTS. 193, 201 (2004).

[28] *Id.*

[29] *Id.* at 198.

[30] *Id.* at 197.

[31] *Id.* at 196.

[32] *Id.* at 199–201.

[33] DANIELLE S. MCNAMARA ET AL., AUTOMATED EVALUATION OF TEXT AND DISCOURSE WITH COH-METRIX 1–2 (2014).

formulae had predicted.[34] The current Coh-Metrix 3.0 web tool[35] now offers a Text Ease and Readability Assessor that groups together different measures into five broad categories—narrativity, syntactic simplicity, word concreteness, referential cohesion, and deep cohesion—and gives a percentile score for each category in order to show how the sample text compares with over 37,000 other texts drawn from a broad reference corpus.[36]

The Coh-Metrix tool is responsive in many ways to criticisms of older readability formulae.[37] Namely, these formulae all rely on the same two types of simple semantic and syntactic measures—vocabulary difficulty (for which word length is often a proxy) and syntactic complexity (for which sentence length is often a proxy)—that are not always, only, or equally responsible for text comprehension variances.[38] By introducing measures of narrativity and cohesion, and by relying on more direct measures of semantic difficulty (e.g., using word frequency instead of word length), the Coh-Metrix tool "is motivated by theories of discourse and text comprehension. Such theories describe comprehension at multiple levels, from shallow, text-based comprehension to deeper levels of comprehension that integrates multiple ideas in the text."[39] In other words, the Coh-Metrix tool represents an attempt to better measure the various properties of a text that reading and discourse theory suggest are actually influential in promoting or inhibiting one's understanding of a text.

New readability formulae that rely upon NLP techniques to measure linguistic properties of both subject texts (i.e., the ones that are relevant in a given study) and reference texts (i.e., the ones typically organized as large corpora that provide a basis for

---

[34] Danielle S. McNamara & Arthur C. Graesser, *Coh-Metrix: An Automated Tool for Theoretical and Applied Natural Language Processing*, *in* APPLIED NATURAL LANGUAGE PROCESSING: IDENTIFICATION, INVESTIGATION, AND RESOLUTION 188, 200 (Philip M. McCarthy & Chutima Boonthum-Denecke eds., 2012).

[35] The Coh-Metrix 3.0 web tool is available at www.cohmetrix.com [https://perma.cc/QML8-ZE3S].

[36] MCNAMARA ET AL., *supra* note 33, at 76–77, 84–95.

[37] ALAN BAILIN & ANN GRAFSTEIN, READABILITY: TEXT AND CONTEXT 53–54 (2016).

[38] *Id.*

[39] McNamara & Graesser, *supra* note 34, at 197.

comparison), together with qualitative assessments of text complexity or comprehension, now emerge regularly.[40] However, some debate continues to exist about the usefulness of such quantitative readability studies. Professors Bailin and Grafstein, for instance, abandon the idea of studying readability from a quantitative perspective altogether and instead encourage discussion of the qualitative aspects of text properties that tend to facilitate or impede comprehension.[41] Professors Davison and Kantor, in their study of four texts that were rewritten in order to simplify the texts, found that the changes did not necessarily lead to better readability scores, mainly because the readability formulae failed to account for important non-quantitative factors that contribute to comprehension.[42] More recently, scholars have begun to note that weaknesses in quantitative readability measurements may exist due to weaknesses in the underlying qualitative (human-involved) assessments of readability (or "criterion variables"), which ground

---

[40] *See, e.g.*, Scott A Crossley et al., *Predicting the Readability of Physicians' Secure Messages to Improve Health Communication Using Novel Linguistic Features: Findings from the ECLIPPSE Study*, 13 J. COMMC'N. HEALTHCARE 344, 346–53 (2020) (showing the results of a study, performed by the authors, of the linguistic properties of 724 secure messages sent by physicians to patients that had been ranked on their readability by a panel of expert raters, to derive a readability formula that validly predicts message readability in this specialized medical context); *see also* Nils Smeuninx, Bernard De Clerck & Walter Aerts, *Measuring the Readability of Sustainability Reports: A Corpus-Based Analysis Through Standard Formulae and NLP*, 57 INT'L J. BUS. COMMC'N. 52, 58–79 (2020) (assessing the readability of private-sector business reports through measures of lexical density, subordinate clause use, and passive voice use, alongside other classic readability formulae).

[41] BAILIN & GRAFSTEIN, *supra* note 37, at 53–54.

[42] Alice Davison & Robert N. Kantor, *On the Failure of Readability Formulas to Define Readable Texts: A Case Study from Adaptations*, 17 READING RESCH. Q. 187, 207 (1982) ("[T]here are features of texts which contribute to readability and that these have not been given their due as factors entering into the question of readability. They are difficult to quantify, and in many cases are only recently beginning to be understood by linguists, cognitive psychologists, and others interested in the analysis of discourse. Yet features of topic, focus, inference load, and point of view play important roles in comprehension, which are all the more crucial to identify because their effects are subtle.").

the formulae.[43] For example, if multiple-choice reading comprehension tests or fill-in-the-missing-word tests—commonly used to norm a readability formula—are not actually valid measures of comprehension, then the formula itself is also likely invalid.[44]

Notwithstanding these criticisms, quantitative readability studies continue to be conducted across a wide variety of disciplines,[45] supporting both private sector[46] and public sector[47] needs. The present study accepts that quantitative readability studies cannot provide a perfect truth on questions relating to the comprehensibility of a text, but these studies can offer some useful indicators about readability levels of texts—particularly for comparative assessments of readability between different texts or text sets.

B. What Is Already Known About the Readability of Law

Much has been written about readability and the plain language movement—an effort to promote more effective communication that, in many ways, implicates concepts of readability—in the

---

[43] James W. Cunningham, Elfrieda H. Hiebert & Heidi Anne Mesmer, *Investigating the Validity of Two Widely Used Quantitative Text Tools*, 31 READING & WRITING 813, 814–18 (2018).

[44] *Id.* at 830–31.

[45] *See, e.g.*, Matthew R. Edmunds, Robert J. Barry & Alastair K. Denniston, *Readability Assessment of Online Ophthalmic Patient Information*, 131 JAMA OPHTHALMOLOGY 1610, 1611–15 (2013) (discussing health care); *see* Scott W. Davis et al., *Say What? How the Interplay of Tweet Readability and Brand Hedonism Affects Consumer Engagement*, 100 J. BUS. RSCH. 150, 154–57 (2019) (discussing social media marketing); *see also* George R. Milne, Mary J. Culnan & Henry Greene, *A Longitudinal Assessment of Online Privacy Notice Readability*, 25 J. PUB. POL'Y & MKTG. 238, 241–45 (2006) (discussing consumer privacy).

[46] *See, e.g.*, Gene E. Burton, *The Readability of Consumer-Oriented Bank Brochures: An Empirical Investigation*, 30 BUS. & SOC'Y 21, 23–25 (1991).

[47] *See, e.g.*, Alexandre Deslongchamps, *Readability and the Bank of Canada*, BANK OF CANADA STAFF ANALYTICAL NOTE 2018–20, June 2018, https://www.bankofcanada.ca/2018/06/staff-analytical-note-2018-20 [https://perma.cc/6S8L-SGXH].

contexts of legislative drafting,[48] jury instructions,[49] and the drafting of legal forms[50] or court briefs.[51] Some of this type of work has involved quantitative readability measures,[52] and/or other forms of empirical analysis.[53] Much of the scholarship focuses on American law, but studies of readability in law also appear within works from other countries.[54] Not all of the studies view readability or plain language efforts as being useful to the law.[55]

There appears to have been far fewer studies on the readability, or plain language, of judicial decisions. Of the quantitative readability studies that have been conducted, some have analyzed the writing styles of individual judges, looking at their idiosyncrasies—particularly of judges who are thought to display

---

[48] *See, e.g.*, Ruth Sullivan, *The Promise of Plain Language Drafting*, 47 MCGILL L.J. 97, 101–08 (2001); *see also* David St. L. Kelly, *Legislative Drafting and Plain English*, 10 ADEL L. REV. 409 (1986) (discussing options for plain language reforms to Australian statutory laws).

[49] *See, e.g.*, Robert P. Charrow & Veda R. Charrow, *Making Legal Language Understandable: A Psycholinguistic Study of Jury Instructions*, 79 COLUM. L. REV. 1306, 1308–11 (1979).

[50] *See, e.g.*, Charles R. Dyer et al., *Improving Access to Justice: Plain Language Family Law Court Forms in Washington State*, 11 SEATTLE J. SOC. JUST. 1065. 1082–95 (2013).

[51] *See, e.g.*, Robert W. Benson & Joan B. Kessler, *Legalese v. Plain English: An Empirical Study of Persuasion and Credibility in Appellate Brief Writing*, 20 LOY. L.A. L. REV. 301, 305–12 (1987).

[52] *See, e.g.*, Lance N. Long & William F. Christensen, *Does the Readability of Your Brief Affect Your Chance of Winning an Appeal?*, 12 J. APP. PRAC. & PROCESS 145, 154–56 (2011).

[53] *See, e.g.*, Maria Mindlin, *Is Plain Language Better? A Comparative Readability Study of Court Forms*, 10 SCRIBES J. LEGAL WRITING 55, 58–60 (2006).

[54] *See*, *e.g.*, I. Turnbull, *Legislative Drafting in Plain Language and Statements of General Principle*, 18 STATUTE L. REV. 21 (1997) (U.K.); *see also* Jeffrey Barnes, *When Plain Language Legislation is Ambiguous – Sources of Doubt and Lessons for the Plain Language Movement*, 34 MELB. U.L. REV. 671, 704–07 (2010) (Austl.) (noting how plain language reforms, alone, cannot resolve most questions of ambiguity within Australian statutes).

[55] *See* Louis J. Sirico, Jr., *Readability Studies: How Technocentrism Can Compromise Research and Legal Determinations*, 26 QUINNIPIAC L. REV. 147, 169–70 (2007).

excellent writing styles.[56] One qualitative study examined the use of plain language techniques by the Supreme Court of the United States ("SCOTUS") during the tenure of Chief Justice Roberts.[57] DeFriez's unpublished doctoral thesis studied—both qualitatively and quantitatively—a sample of 371 Idaho Supreme Court decisions released between 1891 and 2017 and found that the decisions became more readable over time.[58] However, the only published, large-scale quantitative study to consider the readability of any national apex court's decisions was Whalen's 2015 study of 6,206 SCOTUS decisions released since 1946.[59] Whalen calculated the Simple Measure of Gobbledygook ("SMOG")[60] scores for these decisions and found that: (1) decisions were becoming less readable over time;[61] (2) individual judges' decisions became less readable the longer the judges served on the court;[62] and, (3) conservative judges wrote slightly less readable opinions than liberal judges.[63]

No study to date has quantitatively examined the readability of apex (or other) court decisions from a comparative perspective. Similarly, no published study to date has quantitatively assessed the overall readability of decisions from any of the High Court of Australia ("HCA"), the Supreme Court of Canada ("SCC"), the Constitutional Court of South Africa ("ZACC"), or the Supreme

---

[56] *See, e.g.*, Brady Coleman, *Lord Denning & Justice Cardozo: The Judge as Poet-Philosopher*, 32 RUTGERS L.J. 485 (2001); *see also* Nina Varsava, *Elements of Judicial Style: A Quantitative Guide to Neil Gorsuch's Opinion Writing*, 93 N.Y.U. L. REV. ONLINE 75, 85–106 (2018) (reporting the results of a stylometric study of Justice Gorsuch's writing and finding that Justice Gorsuch's reputation as an excellent writer is empirically borne out within the study).

[57] David A. Strauss, *The Plain Language Court*, 38 CARDOZO L. REV. 651 (2016).

[58] Brian M. DeFriez, Toward a Clearer Democracy: The Readability of Idaho Supreme Court Opinions as a Measure of the Court's Democratic Legitimacy (2017) (Ph.D. dissertation, University of Idaho) (ProQuest).

[59] Ryan Whalen, *Judicial Gobbledygook: The Readability of Supreme Court Writing*, 125 YALE L.J. F. 200, 202–10 (2016).

[60] SMOG is a quantitative readability formula that is calculated based on the number of three- (or more) syllable words within a thirty-sentence sample, first introduced in 1969. *See* G. Harry McLaughlin, *SMOG Grading – A New Readability Formula*, 12 J. READING 639, 641 (1969).

[61] Whalen, *supra* note 59, at 202–04.

[62] *Id.* at 204–06.

[63] *Id.* at 208–10.

Court of the United Kingdom ("UKSC"). Therefore, the present study offers new insight about the readability levels of decisions from individual apex courts and how these levels compare with one another across national jurisdictions.

C. Relative Readability and the Comparative Methodology

Although some readability formulae purport to suggest a reading grade level or an approximate education level needed by a reader to comprehend a given text,[64] these suggestions are somewhat unhelpful because the suggested levels can vary by several grades for a particular text depending on which formula is used.[65] Furthermore, many readability formulae do not attempt to benchmark their scores to particular education or grade levels.[66] In other words, knowing an absolute readability score for a particular text, in isolation, is not necessarily meaningful. However, knowing the readability score for a particular text (or group of texts) relative to another text (or group of texts) could be especially useful. For instance, knowing that a document scores a 78 on a readability scale does not tell one much; however, knowing that the same document scores a 78 when most other similar documents score a 35 on the same scale could show that far fewer people are likely to be able to read that particular document than a typical document in its field.

As this discussion illustrates, there are good reasons to employ a comparative methodology when assessing the readability of apex court decisions if one assumes or hypothesizes that readability results will not be identical across all apex courts. To start, some frame of reference is generally needed (or is at least useful) in order to understand the practical significance of a given set of readability measurements for any individual apex court. Should a particular court invest its limited resources in efforts to produce more readable decisions? That question can be answered—at least in part—with information about how readable that court's decisions are in

---

[64] *See* KINCAID ET AL., *supra* note 25, at 19; SENTER & SMITH, *supra* note 24, at 7–12.

[65] *See* McLaughlin, *supra* note 60, at 645 ("Comparisons show that SMOG Grades are generally two grades higher than the corrected Dale-Chall levels.").

[66] *See* Flesch, *supra* note 18, at 225; Crossley, Skalicky & Dascalu, *supra* note 14, at 552–54; MCNAMARA ET AL., *supra* note 33, at 60–77.

comparison with the decisions of other similar courts. To demonstrate this point, an investment may not be sensible if a court is already producing the most readable decisions, but an investment may be well-advised if a court is clearly lagging behind other comparable courts.

However, a comparative approach facilitates more than just a contextualized understanding of raw readability numbers; this approach also opens up the possibility of understanding the factors that explain differences in readability scores. Where different common law jurisdictions all have apex courts that perform essentially the same legal functions (i.e., disposing finally of appeals, developing the law, and standardizing how the law is to be applied)[67] but operate within somewhat different social, political, legal, and institutional environments, there is reason to ask whether any (and if so, which) environmental factors are capable of explaining readability variances across the jurisdictions.

Finally, if such environmental factors are found to be driving readability scores, then a comparative methodology may reveal a type of best solution, or a path forward, for those interested in improving readability scores within a given jurisdiction. For instance, if a study of multiple jurisdictions—each possessing different levels of factor X—shows that factor X correlates strongly with higher readability scores, then a poorly-performing jurisdiction should explore law reform interventions that foster growth of factor X. While factor X may not actually cause changes in readability scores (since correlation merely represents the existence of a dependence or a relationship between two factors, but not necessarily a causal relationship), exploring strongly correlated factors as potential sources of positive change at the beginning is more efficient than deciding on law reform interventions without regard for the relevant data (essentially, by guessing).

The present study leverages the benefits of using a comparative methodology to achieve the study's goals of reporting the readability levels of apex court decisions released in 2020 from five English-speaking jurisdictions. Specifically, the study looks more closely at jurisdiction-specific factors to ascertain whether any of

[67] PAUL DALY, APEX COURTS AND THE COMMON LAW 4–10 (Paul Daly ed., 2019).

them—alone or in combination—can explain readability variances across the jurisdictions.

III. STUDY DESIGN AND METHOD

Because the present study is one of the first of its kind, and because the study employs natural language processing and quantitative linguistics tools not commonly seen in legal scholarship, the following subparts offer detailed descriptions of the design and methods of the study, as well as some of its limitations.

A. Study Design

This subpart describes the design choices that were made in creating the current study and the rationale for those choices. Given the minimal amount of research that has been undertaken about readability levels of court decisions from around the world, a variety of approaches could be used to comparatively measure the readability levels of decisions from courts in different countries. However, as this section illustrates, the design choices that underpin the present study were made in order to facilitate specific, targeted comparisons between jurisdictions that appear to offer a sound basis for studying readability variances between national apex courts.

1. Selecting the Level of Court to Study

Understanding the readability levels of common law court decisions is useful primarily because these decisions declare the law that governs the population, and because the law, as stated by the courts, can be of interest to broader groups of stakeholders. Apex courts tend to declare the law in the most geographically and hierarchically definitive manner (i.e., throughout a jurisdiction's entire territory, and for the benefit of all lower courts within the jurisdiction), so their decisions are likely significant to a wider range and larger number of potential domestic readers than the decisions of lower courts. Additionally, courts and scholars from outside the jurisdiction tend to cite foreign apex court decisions more than trial or appellate court decisions, making these decisions more globally

significant than other court decisions.[68] For these reasons, the present study focuses solely on the readability of apex court decisions.

2. Selecting Jurisdictions for the Study

To begin, this study is concerned with the readability of judicial decisions *as sources of law*. Consequently, each jurisdiction selected for the study needed to form part of the common law "family" that recognizes the precedential value of judicial decisions and their status as sources of law—even if the selected jurisdiction also drew upon other legal traditions.[69] Along related lines, most readability and quantitative linguistic measures are language-specific, so selecting apex courts that all produce decisions in the same language was necessary for this study in order for these decisions to be compared on a common basis.

This study also sought to compare courts that perform similar functions under similar procedural circumstances, so only apex courts that sit at the pinnacle of, at a minimum, a three-tiered court system (consisting of at least one level each of a trial and an intermediate appellate court below the apex court) were included in the study.[70] Specifically, jurisdictions with somewhat equivalent caseloads were compared[71] to ensure that any statistical analysis of the global pool of cases within the study would not be disproportionately affected by decisions from a single jurisdiction. Additionally, jurisdictions from across a broad geographic spectrum were included: North America, Europe, Africa, and Oceania. In

---

[68] *See, e.g.*, TANIA GROPPI & MARIE-CLAIRE PONTHOREAU, THE USE OF FOREIGN PRECEDENTS BY CONSTITUTIONAL JUDGES 418–20 (Tania Groppi & Marie-Claire Ponthoreau eds., 2013) (noting how SCOTUS, SCC, ZACC, and German Constitutional Court decisions have been observed to be the most frequently cited foreign courts within the domestic jurisprudence of other countries).

[69] For instance, Canada's legal system, when viewed as a whole, incorporates elements of common, civil, and Indigenous law. South Africa's legal system similarly incorporates elements of common, civil, and customary law.

[70] This criterion excludes jurisdictions like Singapore and Seychelles, which both use courts of appeal as their apex courts.

[71] The selected jurisdictions within the present study have between thirty and seventy cases per apex court per year. India, in contrast, had 696 cases, and was excluded on that basis.

choosing jurisdictions from within these continents, selections were driven in part by ease of access to the raw data (e.g., the judicial decisions) and ease of effort in manipulating the raw data as required for processing.[72] Finally, the selected jurisdictions needed to have some meaningful variance across jurisdiction- and institution-specific factors, in order to permit a comparative assessment of whether these factors explain any readability variances across the jurisdictions.

Applying these selection criteria to the list of potential jurisdictions, the present study was ultimately designed to include Australia, Canada, South Africa, the United Kingdom, and the United States.

3. Identifying Relevant Linguistic and Readability Measures

One of the most obvious dimensions of a decision's readability is the length of the decision: longer texts take more total time and effort to read than shorter texts. Thus, the present study includes measures of *decision length*, in words, by apex court. Recently, however, applied linguists have realized that several other factors influence understanding, processing effort, and overall readability of texts. For instance, linguistic studies have shown that readability is affected by the extent to which words are imageable or concrete;[73] the present study therefore includes measures of *average concreteness for content words*[74] within decisions.

---

[72] Thus, as between Australia and New Zealand, and as between the United Kingdom and Ireland, Australia and the United Kingdom were selected because their decisions can be easily downloaded and manually converted to plain text files (for processing by readability software) at a rate of about three to five minutes per decision. In contrast, New Zealand's and Ireland's cases are only available in PDF format. When converting these PDFs to plain text files, it would have been necessary to manually remove each line break at the end of a line of text—a process that would take approximately twenty to thirty minutes extra per file.

[73] *See* Max Coltheart, *The MRC Psycholinguistic Database*, 33 Q. J. EXPERIMENTAL PSYCH. 497, 497 (1981).

[74] Content words (or lexical words) are words that contribute more information to a text and make up the overwhelming majority of words in the English language. Content words can be contrasted with function words (or grammatical words) like pronouns, prepositions, and conjunctions that do not add meaning inasmuch as they provide grammatical and relational structures for content words.

Linguists also recognize that readability and mental processing times are affected by "the degree of cognitive entrenchment of particular words / grammatical patterns" that are used within a text.[75] One quantitative technique for measuring such entrenchment involves comparing a sample text (a judicial decision, for example) to a reference corpus (a large body of representative texts) to see how often words used within the sample appear within the reference corpus—a raw frequency measure.[76] The present study reports one such linguistic measure that influences readability: the *average frequency for function words*, using the Corpus of Contemporary American English ("COCA")[77] Academic corpus as the reference corpus.[78] Applied linguists have also recognized that concepts of frequency can be applied, not only to individual words, but also to multi-word phrases[79]—where more common phrases are processed by audiences more easily than less common phrases.[80] Consequently, this study includes a frequency measure of the

---

See RONALD CARTER, VOCABULARY: APPLIED LINGUISTIC PERSPECTIVES 8 (2d ed. 1998).

[75] Stefan Th. Gries, *Dispersions and Adjusted Frequencies in Corpora*, 13 INT'L J. CORPUS LINGUISTICS 403, 403 (2008).

[76] Xiaobin Chen & Detmar Meurers, *Word Frequency and Readability: Predicting the Text-Level Readability with a Lexical-Level Attribute*, 41 J. RES. IN READING 486, 488–91 (2018).

[77] Mark Davies, *The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English*, 25 LITERARY & LINGUISTIC COMPUTING 447, 453–54 (2010).

[78] One would expect frequent use of academic language in a judicial decision to render the decision less readable.

[79] Multi-word phrases are also often called N-grams (where "N" represents the size of the phrase). Two- and three-word phrases are also often called bigrams and trigrams, respectively. To offer a concrete example, consider the following sentence: "I am hungry today." The sentence contains three distinct bigrams (I am; am hungry; hungry today) and two distinct trigrams (I am hungry; am hungry today).

[80] *See* Inbal Arnon & Neal Snider, *More Than Words: Frequency Effects for Multi-Word Phrases*, 62 J. MEMORY & LANGUAGE 67, 76 (2010). The theory that common phrases are more easily processed makes intuitive sense: readers of this footnote will likely process the phrase "stop at the red light" much faster than they would process the phrase "go at the red light." The latter phrase jars on the reader because of its unfamiliarity and may require re-reading for confirmation of the contents of the phrase.

proportion of bigrams in judicial decisions that appear in the top 20,000 most common bigrams from within the COCA Fiction corpus.[81]

Each of the above linguistic measures offer unidimensional assessments of a factor that likely influences text readability. However, readability is understood to be affected simultaneously by many different factors. Therefore, this study also—and perhaps most importantly—includes comprehensive (or multidimensional) readability scores. Specifically, this study reports on *Flesch-Kincaid* scores,[82] *SMOG* scores,[83] and *Crowd-sourced Reading Comprehension-Modified* ("*CAREC-M*")[84] scores. *Flesch-Kincaid* scores are included because this measure of readability is arguably the most widely known across all disciplines, perhaps because of its inclusion within common word processing software packages.[85] *SMOG* scores are included because this measure has already been used in legal scholarship as part of a large-scale study looking at SCOTUS decisions.[86] Finally, *CAREC-M* scores are included because this new measure leverages NLP techniques in order to derive a comprehensive readability formula from observations of several hundred text-related features (relating to word, phrase, and sentence properties; sentiment; cohesion; and, numerous other linguistic and grammatical properties of texts).[87] In this sense, the *CAREC-M* score is perhaps the most sophisticated general readability measure currently available: the score is "based on

---

[81] Davies, *supra* note 77, at 453–54. Where the COCA Fiction corpus is a general and non-specialized corpus, one would expect that judicial decisions using a high proportion of the top 20,000 bigrams from that corpus to be more readable than decisions using a low proportion of such bigrams. *See id.*

[82] KINCAID ET AL., *supra* note 25, at 14.

[83] McLaughlin, *supra* note 60, at 639.

[84] Crossley, Skalicky & Dascalu, *supra* note 14, at 553; Joon Suh Choi & Scott A. Crossley, *Assessing Readability Formulas: A Comparison of Readability Formula Performance on the Classification of Simplified Texts*, EASYCHAIR (July 13, 2020), https://easychair.org/publications/preprint_download/Glkz [https://perma.cc/3T87-X25Z].

[85] *See* Norman Otto Stockmeyer, *Using Microsoft Word's Readability Program*, 88 MICH. BAR J. 46, 46 (2009).

[86] *See* Whalen, *supra* note 59, at 202–10.

[87] *See* Crossley, Skalicky & Dascalu, *supra* note 14, at 549–51.

linguistic features that better represent theoretical and behavioural accounts of the reading process [and] significantly outperformed classic readability formulas" in a validating study.[88]

The above measures, when calculated for each of the large number of full-text judicial decisions within the current study, provide a robust and informed picture of the readability of those judicial decisions. Although each measure offers distinct insights into decision readability levels, *CAREC-M* scores were chosen within this study as the most comprehensive measure of readability. Consequently, all comparative and statistical analyses examining jurisdiction- and court-specific variables as potential sources of explanation for readability variances within the present study were performed using *CAREC-M* scores as the relevant readability measure.

4. Identifying Variables for Comparison

Initial research into apex courts from the selected jurisdictions, as well as research of the environments within which these courts operate, revealed many differences that could provide a useful basis for comparison. In particular, the overall education levels of the populations in the different jurisdictions varied substantially. One might expect that judges would write decisions with some sense of the population's education levels in mind, such that readability levels would be higher in jurisdictions with lower general education levels. Accordingly, this variable was included in the study, with a single measurement of *adult secondary school completion rate* for each jurisdiction.

On a related point, initial findings showed that judges within different jurisdictions possessed widely divergent levels of post-secondary education, and that former law professors were appointed to the apex courts more often in some jurisdictions than others. One could logically assume that courts comprised of more-educated judges would tend to produce less-readable decisions (since these judges presumably have access to broader academic vocabularies and have more experience with complex writing styles). Thus, these variables were also included within the study—with discrete

[88] *Id.* at 557.

measures of (educational) *degrees per judge* and *former law professors per judge* for each case within the study. Additionally, in terms of demographic characteristics of judges, some courts had a significantly greater proportion of female judges than other courts. Some research has suggested that women write more readably than men,[89] so one might expect that courts with a greater relative representation of women would produce more readable decisions; this variable—*women per judge*—was included in the study on that basis.

On a more institutional-procedural level, different apex courts hear cases with different panel sizes: in Australia some cases were decided by a single judge, while in South Africa, one case was decided by a panel of eleven judges (and many were decided by panels of ten judges).[90] One might expect that decisions would be more readable where panel sizes are larger, on the assumption that more effort would need to be expended to communicate clearly to fellow judges on the larger panel for the purpose of building a majority or consensus view. For this reason, *panel size* was included as a variable within the study, with unique measurements for each decision.

In addition, judicial law clerks were used to differing extents within the selected jurisdictions, with clerks heavily involved in drafting decisions in some jurisdictions and not involved at all in other jurisdictions. One might expect that a more collaborative decision-drafting jurisdiction that involves clerks and judges (instead of only judges) would produce more readable decisions. On that basis, *clerk involvement* was included as a variable, with a single subjective and relative ranking included for each jurisdiction.

Finally, the judicial appointment processes, and overall role of the apex courts, are politicized to different extents in each jurisdiction. One might expect judicial decisions to be more readable in places with higher levels of politicization since a court's legitimacy in such places likely depends more heavily on approval of the broad population. This factor might accordingly drive judges

---

[89] *See* Erin Hengel, *Publishing While Female*, *in* WOMEN IN ECONOMICS 80, 80–82 (Shelley Lundberg ed., 2020).

[90] *See infra* Table 6.

to write more accessible decisions aimed at the general population in such jurisdictions. The variable of *court politicization* was therefore included within the study, again with a single subjective and relative ranking included for each jurisdiction.

Although the apex courts within the present study all fill similar roles and perform similar functions, the differences in the above-listed factors across the jurisdictions provide ample basis for meaningful comparison. By studying readability variances alongside differences in each of the above variables, the study can assess the extent to which any of the variables alone, or in combination, can explain the readability variances of the apex courts.

B. Method

This subpart provides a detailed explanation of the way in which the present study was conducted. Ideally, the results that the study has produced should be replicable by anyone who follows the method described below.

1. Case Selection and Acquisition

The data for this study was collected by first identifying all decided cases for each apex court in 2020 from the respective courts' websites.[91] Australian cases were downloaded in rich-text format ("RTF") and then batch converted to plain text ("TXT") format using the Mac OS 11 *Text Utility*.[92] For each Canadian, South

[91] *Judgments, Ordered By Date, Browsing By Year (2020)*, AUSTL. HIGH CT., http://eresources.hcourt.gov.au/browse?col=0&facets=dateDecided&srch-term=2020 [https://perma.cc/SK3H-ALLB]; *Supreme Court Judgments*, SUP. CT. OF CAN., https://scc-csc.lexum.com/scc-csc/scc-csc/en/2020/nav_date.do [https://perma.cc/Z3QE-4VHQ]; *2020 South Africa: Constitutional Court Decisions*, S. AFR. LEGAL INFO. INST., http://www.saflii.org/za/cases/ZACC/2020/ [https://perma.cc/H5WK-SETY] (last visited Sep. 22, 2021); *Decided Cases*, THE SUP. CT., https://www.supremecourt.uk/decided-cases/2020.html [https://perma.cc/5FBK-JGV5]; *Opinions of the Court – 2020*, SUP. CT. OF U.S., https://www.supremecourt.gov/opinions/slipopinion/20 [https://perma.cc/594J-BYAY].

[92] See *How to Batch Convert DOCX Files to TXT Format with Textutil in Mac OS X*, OSXDAILY, (February 20, 2014), https://osxdaily.com/2014/02/20/batch-convert-docx-to-txt-mac/ [https://perma.cc/PJD4-3YBW], for a description of how to perform this conversion.

African, and United Kingdom decision, the full text of the decision was "copied" from the HTML webpage containing the decision, and "pasted" into a new TXT file using the Mac OS 11 *Text Editor* application. For the United States, each decision's citation was used to search the decision in the Google Scholar database, where HTML versions of the decisions were available. The text of each decision from its HTML webpage was then "copied" and "pasted" into a text file (following the same process as for Canadian decisions). In this manner, individual TXT files were created and stored for all decisions from 2020 from each of the selected apex courts.

2. Data Pre-Processing

Each TXT file was opened individually. Once opened, all front-end matter preceding the text of the decision, other than core identifying information (e.g., style of cause, date, judges present, etc.), was manually deleted, such as the names of counsel, headnotes, case summaries, cases cited, authors cited, and other similar front-end information. Similarly, all back-end information following final statements of disposition of the cases or other conclusions were also deleted manually. The back-end information that was deleted included footnotes, annexes or appendices, copies of orders issued by the courts, and other similar information. The extent of information that preceded or followed the actual decision varied greatly from one decision to another (e.g., some SCOTUS cases had extensive footnotes, while others had no footnotes), and from one jurisdiction to another (e.g., most SCC cases had lengthy headnotes and case summaries, but most UKSC cases had no such information). After manually deleting the front- and back-end text from each file, the decisions were then saved and ready for processing by NLP software with only a common and approximately equal amount of extra (case-identifying) text included in each file.

3. Processing Tools, NLP Computations, and Exclusions

All decisions were processed through the Simple Natural Language Processing ("SiNLP") software application,[93] which is freely available for both Mac and Windows operating systems,[94] to measure *decision length* (in words) for each decision. At this stage, any decisions that contained less than 260 words (including any remaining front- and back-end text) were excluded from the study (N = 16).[95] These decisions were excluded because both the *SMOG* and *CAREC-M* comprehensive readability formulae are intended for use with larger text samples.[96] The remaining decisions were then processed through the Tool for the Automatic Analysis of Lexical Sophistication ("TAALES"),[97] which is also freely available for

---

[93] Scott Crossley et al., *Analyzing Discourse Processing Using a Simple Natural Language Processing Tool (SiNLP)*, 51 DISCOURSE PROCESSES 511, 520–24 (2014). This application provides seven different simple linguistic measures, such as number of words, sentences and paragraphs, and average word and sentence lengths, for all text files processed by the software. *See id.*

[94] *See NLP Tools for the Social Sciences – SiNLP: The Simple Natural Language Processing Tool*, NLP TOOLS FOR THE SOC. SCIS. [hereinafter NLP Tools for the Social Sciences], https://www.linguisticanalysistools.org/sinlp.html [https://perma.cc/5MRG-F74N] (last visited Oct. 4, 2021).

[95] From Australia, N=1; from Canada, N=13; from the United States, N=2; and from both South Africa and the United Kingdom, N=0. The excluded decisions tended to be ones wherein a lower court's decision was upheld or overturned by the apex court in a very short opinion that expressed full agreement with the lower court (or a judge of that lower court) without further explanation.

[96] SMOG calculations are based on a minimum of thirty sentences of text (which would equate to approximately 600–900 words of text from a typical judicial decision). *See* McLaughlin, *supra* note 55. CAREC-M calculations are intended for text samples of more than 200 words. *See* J.S. Choi & S.A. Crossley, *NLP Tools for the Social Sciences - ARTE: Automatic Readability Tool for English*, NLP TOOLS FOR THE SOC. SCIS., https://www.linguistic analysistools.org/arte.html [https://perma.cc/E5G9-LT9L] (last visited Oct. 4, 2021). Given that front- and back-end matter comprised approximately 60 words in many decisions within the present study, a minimum threshold of 260 words was selected as an inclusion criterion.

[97] Kristopher Kyle, Scott Crossley & Cynthia Berger, *The Tool for the Automatic Analysis of Lexical Sophistication (TAALES): Version 2.0*, 50 BEHAV. RES. METHODS 1030, 1032–37 (2018). This application provides over 250 different linguistic measures, including range and frequency for words and N-gram from multiple corpora, psycholinguistic properties of words, and many other related measures, for all text files processed by the software. *Id.*

both Mac and Windows operating systems,[98] in order to measure the following text dimensions: *average concreteness for content words*; *average frequency for function words (COCA Academic)*; and, *average proportion of bigrams in top 20K (COCA Fiction)*. Finally, all decisions were processed through the Automatic Readability Tool for English ("ARTE") software application,[99] which is freely available for both Mac and Windows operating systems,[100] to compute *Flesch-Kincaid*, *SMOG*, and *CAREC-M* comprehensive readability scores.

4. Data Collection and Coding of Variables

Data for each jurisdiction for *adult secondary school completion rate* was taken from an Organisation for Economic Co-operation and Development ("OECD") database[101] using figures from 2018 (the most recent year with reported figures for all five of the jurisdictions forming part of the present study). This data point is reported in terms of the percentage of the adult population in the jurisdiction with less than a completed upper secondary school level of education.[102]

Data for *panel size* was compiled by identifying, via a manual count for each decision, the number of judges who participated in the decision. Similarly, data for *degrees per judge*, *former law professors per judge*, and *women per judge* was collected first by identifying which judges participated in the decision. This information was used, together with publicly available biographical information about each judge relating to their educational and professional experiences (drawn primarily from official court websites), in order to produce the relevant measures. For the variable *degrees per judge*, each post-secondary degree possessed by a judge was counted (regardless of whether the degree was at the

---

[98] *See NLP Tools for the Social Sciences*, *supra* note 94.

[99] Choi & Crossley, *supra* note 84. This application provides comprehensive readability scores for all text files processed by the software, based on nine different formulae (e.g., Dale-Chall, SMOG, ARI).

[100] *Id.*

[101] *Adult Education Level*, ORG. ECON. COOPERATION & DEV, https://data.oecd.org/eduatt/adult-education-level.htm [https://perma.cc/S8J8-FY6Y] (last visited Apr. 2, 2021).

[102] *Id.*

undergraduate, masters, or doctoral level), but diplomas and certificates were not counted. The variable was calculated by dividing the total number of degrees possessed by all judges on a panel by the number of judges on the panel. For the variable *former law professors per judge*, any judge who had worked full-time as a law professor for at least two years was counted, but judges who had taught on a part-time basis as sessional or adjunct faculty were not counted. The variable was calculated by dividing the number of judges on a panel who had formerly been law professors by the total number of judges on the panel. For the variable *women per judge*, any judge who was biographically described using "she/her" pronouns was counted, and the variable was calculated by dividing the total number of women on a panel by the total number of judges on the panel.

Quantitative data for *clerk involvement* and *court politicization* were not yet available for use within the present study. This study therefore relied on secondary sources discussing each of these variables to derive subjective relative scores for each jurisdiction. The results and the sources relied upon to derive the results are identified in detail below.[103]

5. Descriptive and Analytical Statistical Techniques

A number of approaches were used in order to determine how readability variances across jurisdictions may be explained by jurisdiction- or court-specific variables. With respect to variables for which only national data is available and for which there are not discrete measurements specific to each case being analyzed (e.g., *adult secondary school completion rate*, *clerk involvement*, and *court politicization*), comparative analysis was undertaken by comparing mean readability scores in each jurisdiction with national levels of the relevant variable—in a largely descriptive manner. In the case of variables for which there are discrete measurements for each case (e.g., *panel size*, *degrees per judge*, *former law professors per judge*, and *women per judge*), statistical analysis was performed using SPSS software to compute Pearson correlations between each variable and *CAREC-M* readability scores. Additionally, SPSS

[103] *See infra* Part IV.B.2 ("*Clerk Involvement*") and Part IV.B.3 ("*Court Politicization*").

software was used to run a multiple regression analysis as a means of modeling the extent to which a combination of variables might explain overall readability variances across jurisdictions.

C. Limitations of the Study

The present research sheds meaningful light on an understudied topic: the readability of apex court decisions; however, the limitations of this work should be recognized. First, the study does not identify what factors *cause* readability variances across jurisdictions; rather, this study only illustrates associations, correlations, or regression coefficients between different variables and associated readability levels to show the relationships between these properties. Second, the present study does not purport to exhaustively survey all of the potential jurisdiction- or court-specific variables that may correlate with, or explain, readability levels. For instance, one might hypothesize that readability levels would be affected by differences in rates of litigants' self-representation across the different national jurisdictions (on the assumption that courts would produce more readable decisions in jurisdictions where litigants more frequently ascertain the law for themselves, without the assistance of counsel). However, preliminary research quickly revealed that such data is not readily available for each selected jurisdiction and is collected inconsistently (if at all) in many places.[104] As a result, the variable that reflects the rates of litigant

[104] *See, e.g.*, GOV'T. OF S. AFR., GOVERNANCE, PUBLIC SAFETY AND JUSTICE SURVEY GPSJS 2018/19 1, 46 (2020), http://www.statssa.gov.za/publications/P0340/P03402019.pdf [https://perma.cc/EX98-ENRK] (providing self-reported information from justice system participants, suggesting that 48% of accused persons are unrepresented, and 77% of "litigants" are unrepresented, whereby the term "litigants" is not defined in that context); *see also* Mark D. Gough & Emily S. Taylor Poppe, *(Un)Changing Rates of Pro Se Litigation in Federal Court*, 45 L. & SOC. INQUIRY 567, 574–75 (2020) (noting that at least one party was self-represented in 27% of U.S. federal district court cases based on a 2018 study, which provided no data on the extent of self-representation in state courts); *see also* JULIE MACFARLANE, THE NATIONAL SELF-REPRESENTED LITIGANTS PROJECT: IDENTIFYING AND MEETING THE NEEDS OF SELF-REPRESENTED LITIGANTS 1, 8 (2013), http://representingyourselfcanada.com/wp-content/uploads/2015/07/nsrlp-srl-research-study-final-report.pdf [https://perma.cc/VEP5-KDN8] (highlighting rates of self-representation in civil

self-representation, as well as other variables where information could not readily be found, were not considered within the present study. Finally, the present study applies a comparative methodology that deliberately excludes (or at least very significantly dilutes) the consideration of variables related to authorship of judicial decisions, in order to focus attention on jurisdiction- and court-specific factors that may explain readability variances across different apex courts. Notably, variables related to authorship of judicial decisions are likely correlated with readability scores; however, those variables are beyond the scope of the present comparative study.

IV. RESULTS AND DISCUSSION

This Part presents and discusses the results from this Article's original comparative study of the readability of apex court decisions released in 2020 from Australia, Canada, South Africa, the United Kingdom, and the United States. The study involved 233 decisions, consisting of over 3 million words of text. The results offer an up-to-date and comprehensive account of the readability levels of judicial decisions produced by the selected apex courts.

Subpart A describes the readability results for each linguistic measure within each jurisdiction, and Subpart B discusses the jurisdiction- and institution-specific environments within which each apex court operates. Specifically, the comparative analysis in Subpart B uses descriptive statistics to discuss how *adult secondary school completion rate*, *clerk involvement*, and *court politicization* levels relate to average decision readability levels in the different jurisdictions. The comparative analysis also employs analytical statistics to illustrate through correlations and a regression model how readability variances across jurisdictions may be explained by *panel size*, *former law professors per judge*, and *women per judge* variables, but these variances cannot appreciably be explained by *degrees per judge*.

---

and family court cases as reported from a study of 259 self-represented litigants drawn from Alberta, British Columbia, and Ontario; rates of self-representation in other types of court cases, and in other provinces, were not considered within this study).

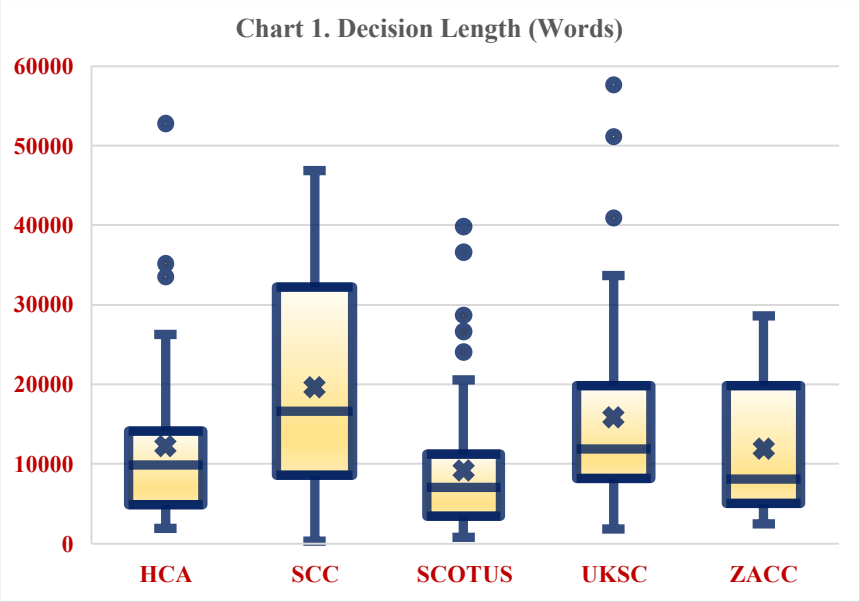
A. Results: Reporting Linguistic and Readability Measures

This subpart provides a snapshot of how apex courts perform in terms of several isolated linguistic measures that may impact readability of their decisions, and also presents apex courts' performances in terms of more comprehensive readability scores. Through visual, tabular, and descriptive accounts of how each apex court communicates its decisions, this subpart aims to illustrate how readable each court's decisions are relative to one another.

1. Decision Length

Boxplots of all decision lengths are shown in Chart 1, below. For each jurisdiction, the box represents the interquartile range[105] for *decision length* in that jurisdiction. The middle horizontal line represents the median value. The "x" represents the average or mean value for the jurisdiction. The "whiskers" extending above and below each box extend to show the full range of the decision lengths in the jurisdiction, or to 1.5 times the size of the interquartile range—whichever is greater. Individual data points extending above or below the whiskers represent outlier values that are noteworthy for their distance away from the central tendency (mean/median) of the data. Similar boxplots are shown for various other data in subsequent Charts and should be read in the same manner as the current boxplot.

[105] "Interquartile range" refers to the range between the twenty-fifth and seventy-fifth percentiles; it offers a view of the middle fifty-percent of the data points and is less sensitive to outlier points than other dispersion measures. MICHAEL O. FINKELSTEIN & BRUCE LEVIN, STATISTICS FOR LAWYERS 25 (3d ed. 2015).



The longest recorded *decision length* measurement was for a UKSC decision that was 57,632 words.[106] The shortest decision was from the SCC, at 316 words.[107] For greater precision and ease of comparison, average *decision lengths* by jurisdiction, reported in number of words, are shown in Table 1, below. Standard deviations are also included for each jurisdiction. Jurisdictions are ranked based on average *decision length*, from shortest to longest.

[106] Test Claimants in the Franked Inv. Income Grp. Litig. & Others v. Comm'rs of Inland Revenue [2020] UKSC 47 (ruling on a corporate taxation case).

[107] R. v. Kishayinew, 2020 SCC 34 (ruling on a criminal case wherein reasons were delivered orally, substantially supporting the reasons given by the dissenting judge from the Court of Appeal below). It should be recalled that decisions of less than 260 words were excluded from the study.

Table 1. *Decision Length* by Jurisdiction			
Rank	**Apex Court**	**Average *Decision Length* (words)**	**Standard Deviation (words)**
1	SCOTUS	9,215	8,037
2	ZACC	11,929	8,366
3	HCA	12,250	10,252
4	UKSC	15,860	11,838
5	SCC	19,680	13,447

The above data on *decision length* presents several findings. First, from a total reading time (or absolute level of processing effort) perspective, one would expect that American decisions would be ranked the lowest, and Canadian decisions would be the highest, because one expends more time and effort to read longer texts than shorter texts. However, shorter texts are not necessarily more easily understood than longer texts. Moreover, this study focuses on how well judicial decisions may facilitate comprehension of the common law, rather than on how efficiently or succinctly judicial decisions communicate their points. Thus, *decision length* scores provide some useful information about the level of effort that would be required to read a court's decisions and may therefore serve as practical indicators of whether individuals are likely to even attempt reading that court's decisions. However, these scores provide little information about how readable or comprehensible a court's decisions are likely to be for those individuals who decide to read the texts.

Second, the large standard deviations[108] in each jurisdiction (shown in Table 1) and the dispersion of measurements (shown in

[108] Standard deviation is a statistic describing how dispersed the measurements within a sample are, relative to the average measurement for that sample: a low standard deviation indicates that measurements within the sample are generally close in size to the average, while a high standard deviation indicates that measurements within the sample are generally farther in size from the average. FINKELSTEIN & LEVIN, *supra* note 105, at 21–23.

Chart 1) indicate that *decision lengths* in all the studied jurisdictions are spread widely. Thus, this data suggests that there are not strong guiding norms in any of the jurisdictions as to the ideal apex court *decision length* for all cases. Some decisions are very short, and some are exceedingly long. The SCC, for instance, produced three decisions that were less than 1,000 words each,[109] and two decisions that were more than 43,000 words each.[110] Similarly, SCOTUS produced five decisions that were less than 2,000 words each,[111] and two decisions that were more than 36,000 words.[112] The variations in *decision lengths* are perhaps to be expected, given how different (factually and legally) each case that comes before a given apex court might be from all other cases heard by that court. Regardless, considering that similar dispersions exist across all of the studied apex courts,[113] one might conclude that this dispersion phenomenon

[109] R. v. Kishayinew, 2020 SCC 34; R. v. Doonanco, 2020 SCC 2 (ruling on a criminal case wherein the SCC agreed with the reasons of the dissenting judge from the Court of Appeal below but ordered a different disposition of the case); R. v. Li, 2020 SCC 12 (ruling on a criminal case wherein the SCC noted that one of its recent prior decisions—released after both the trial and appeal court decisions had been rendered—supplied the correct legal framework). The SCC briefly applied the framework to the *Li* case and disposed of the appeal). *See id.*

[110] Uber Technologies Inc. v. Heller, 2020 SCC 16 (ruling on a contract case involving forum selection and class action issues); Conseil scolaire francophone de la Colombie-Britannique v. B.C., 2020 SCC 13 (ruling on a constitutional language rights case).

[111] Davis v. United States, 140 S. Ct. 1060, (2020) (ruling on a criminal sentencing appeal); McKesson v. Doe, 141 S. Ct. 48 (2020) (granting a petition for a writ of certiorari in a civil damages case and remanding the case to the appeals court); Taylor v. Riojas, 141 S. Ct. 52 (2020) (granting a petition for a writ of certiorari in a prisoner's case and remanding the case to the Court of Appeals); Lomax v. Ortiz-Marquez, 140 S. Ct. 1721 (2020) (dismissing a case due to a pattern of previous unmeritorious litigation by the prisoner); Rodriguez v. FDIC, 140 S. Ct. 713 (2020) (vacating and remanding a federalism case involving a tax refund).

[112] Bostock v. Clayton Cnty., 140 S. Ct. 1731 (2020) (ruling on an LGBT employment discrimination case); June Medical Servs. LLC v. Russo, 140 S. Ct. 2103 (2020) (ruling on an abortion case).

[113] The coefficients of variance (calculated by dividing each jurisdiction's standard deviation by its average) for each jurisdiction are as follows: United States: 0.87; South Africa: 0.70; Australia: 0.84; United Kingdom: 0.75; Canada: 0.67. The similar values of these statistics suggest that, relative to one another, each apex court has a similar level of variance around its own mean.

is either desirable, or at least somewhat inevitable, and therefore not a necessary subject for further study.

Third, there is substantial variation in average *decision lengths* across jurisdictions, with the average SCC decision having more than twice the length of the average SCOTUS decision. Where the studied apex courts all serve essentially the same functions, some courts are tellingly capable of communicating their legal reasoning with far fewer words than other courts. If making the common law, as set out in judicial decisions, accessible to more people, or accessible to people more quickly (with less total reading time/effort), is an important aim of apex courts, then perhaps the UKSC and the SCC should consider how their decisions could be pared down. For instance, perhaps these courts could shorten their summaries of relevant facts[114] or their accounts of how the lower courts treated the case at issue to reduce overall decision lengths and hopefully increase readability.

Finally, comparing the *decision lengths* reported above in Table 1 with *decision lengths* for the selected apex courts from previous studies would be helpful. However, no compatible studies have been conducted that also used "number of words" as the relevant measure of length for the entire content of the judicial decisions,[115] so such comparisons are—for the moment—unavailable.

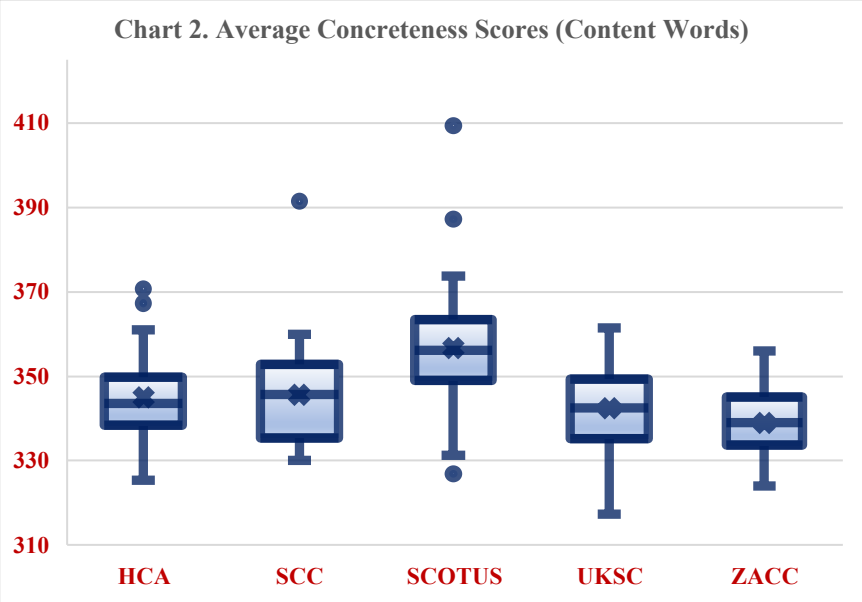
---

[114] On this point, the Court of Appeal for Ontario recently leveled some harsh criticism against judges who employ excessive length in their recitations of facts. *See* Welton v. United Lands Corp. Ltd., 2020 ONCA 322, at para. 56 ("I conclude by expressing a concern about the length of the reasons for decision in this case, which is reflective of an unfortunately growing trend.") and para. 63 ("Digesting unduly lengthy reasons consumes far too much time . . . . A data dump does not constitute fact-finding.").

[115] See Stephen M. Johnson, *The Changing Discourse of the Supreme Court*, 12 U.N.H. L. REV. 29, 57–58 (2014), for a study of SCOTUS decisions from 2009 to 2011 that reports decision length in "number of pages." See Ryan C. Black & James F.II. Spriggs, *An Empirical Analysis of the Length of U.S. Supreme Court Opinions*, 45 HOUS. L. REV. 621, 630–31 (2008), for a study of SCOTUS cases from 1971 to 2005 that reports decision length in "number of words" and in which the authors report "opinions" of each judge separately, such that there is no easy way of comparing overall decision lengths (i.e., the sum length of all opinions that make up a single decision) from their study with decision lengths from the present

2. Average Concreteness for Content Words

Boxplots of all concreteness scores for individual decisions are shown in Chart 2, below.



The lowest recorded value of concreteness was from the UKSC, for a decision scoring a 317.[116] The decision with the highest average content word concreteness score was from SCOTUS, at 409.[117]

The *average concreteness for content words* in decisions, reported in raw scores,[118] are shown in Table 2, below. This measure

---

study. See David J. Carter et al., *Reading the High Court at a Distance: Topic Modelling the Legal Subject Matter and Judicial Activity of the High Court of Australia, 1903-2015*, 39 U.N.S.W.L.J. 1300, 1315 (2016), for a study on decision lengths at the High Court of Australia, but in terms of average characters per decision.

[116] Comm'rs for Her Majesty's Revenue & Customs v. Parry and others, [2020] UKSC 35 (a pension inheritance case).

[117] U.S. Forest Serv. v. Cowpasture River Pres. Ass'n., 140 S. Ct. 1837 (2020) (a natural gas pipeline case).

[118] The Medical Research Council (MRC) Psycholinguistic Database from which concreteness scores are drawn reports scores on an integer scale ranging from 100 to 700, with the lowest reported word value having a score of 158, the

takes the sum of each *average concreteness for content words* score from every individual apex court decision from a jurisdiction (which are computed using TAALES),[119] and divides that figure by the number of apex court decisions in that jurisdiction, to produce a jurisdiction-wide average score. Standard deviations are also included. Jurisdictions are ranked from the highest *average concreteness for content words* score to the lowest score.

Table 2. *Average Concreteness for Content Words* by Jurisdiction			
Rank	**Apex Court**	**Average *Concreteness* (Raw Score)**	**Standard Deviation (Raw Score)**
1	SCOTUS	356.7	13.1
2	SCC	345.6	11.8
3	HCA	345.0	11.1
4	UKSC	342.4	9.4
5	ZACC	338.9	7.6

Where more concrete language is thought to facilitate a reader's comprehension of a text, the above data suggests that SCOTUS decisions may be the most readable, and ZACC decisions the least readable, with SCC, HCA, and UKSC decisions clustered more closely together in the middle. The *average concreteness* score for all words in the reference database is 438, and the decision with the highest *average concreteness* score from the entire study scored only 409.[120] This information suggests that all the studied apex courts may use words that are, on average, less concrete (more

highest word value having a score of 670, and the average word value at 438. There are 8,228 words in the database with concreteness scores. *See MRC Psycholinguistic Database Version 2.0*, UNIV. OF W. AUSTL. SCH. OF PSYCH. (Apr. 1, 1987), https://websites.psychology.uwa.edu.au/school/mrcdatabase/mrc2.html [https://perma.cc/A7WZ-UC22].

[119] *See NLP Tools for the Social Sciences*, *supra* note 98.

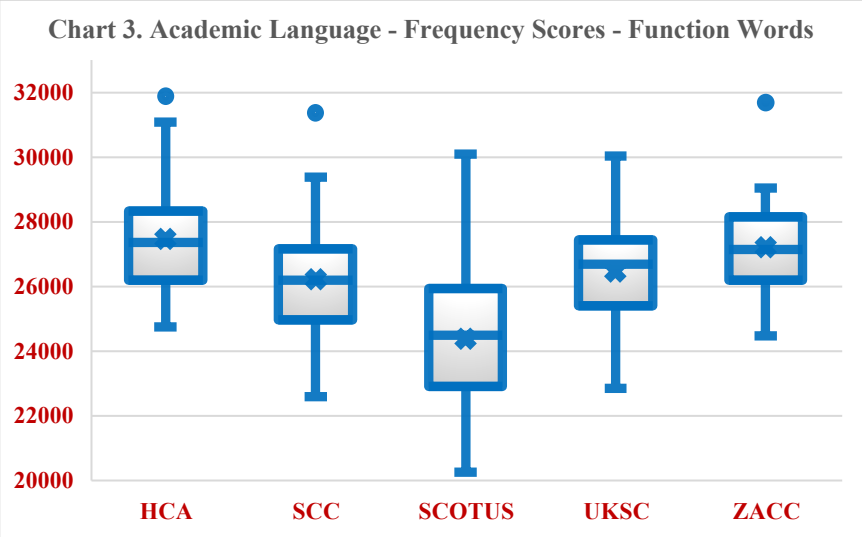
[120] *U.S. Forest Serv.*, 140 S. Ct. at 1837.

abstract) than words used in more common forms of English communication.

This data importantly points to a specific linguistic feature upon which different apex courts show variance. In other words, there are clear indications from this data (in terms of more concrete language and word choice) that show how some apex courts might make their language more accessible to readers—perhaps by following the example of SCOTUS.

3. Average Frequency for Function Words—COCA Academic Corpus

Boxplots of all *average frequency for function words* scores, reflecting each individual decision within the present study, are shown in Chart 3, below. This measure is intended to illustrate the extent to which language in judicial decisions overlaps with language in academic (as opposed to popular, media, news, or other simpler) texts.



The highest reported value was from an HCA decision[121] with a score of 31,879 occurrences-per-function-word in the reference corpus—the greatest overlap of function word use in an apex court decision compared to an academic reference corpus. The decision with the lowest average function word frequency score was from SCOTUS,[122] at 20,242 occurrences in the academic reference corpus.

The *average frequency for function words*, reported in terms of the average number of occurrences-per-function-word in the reference corpus, are shown in Table 3, below. The calculation of this score is explained in the next paragraph.

For each judicial decision, an average function word frequency score is calculated by summing the total number of times that all function words within the judicial decision appear within the reference corpus, and then dividing that sum by the total number of function words that contributed to the sum. This process (which is computed using TAALES)[123] produces an average frequency score for function words for each judicial decision. These individual decision scores are then summed for each jurisdiction and divided by the total number of decisions within that jurisdiction, to produce jurisdiction-wide *average frequency for function word* scores. Standard deviations are also included. Jurisdictions are ranked from the lowest *average frequency for function words* score to the highest score.

---

[121] *Kadir v. The Queen; Grech v. The Queen*, [2020] HCA 1 (Austl.) (ruling in a criminal case involving cruelty to animal charges against two individuals who were jointly tried).

[122] Rutledge v. Pharm. Care Mgmt. Ass'n, 141 S. Ct. 474 (2020) (deciding a statutory interpretation appeal involving pharmacy and prescription insurance benefit questions).

[123] *See NLP Tools for the Social Sciences*, *supra* note 98.

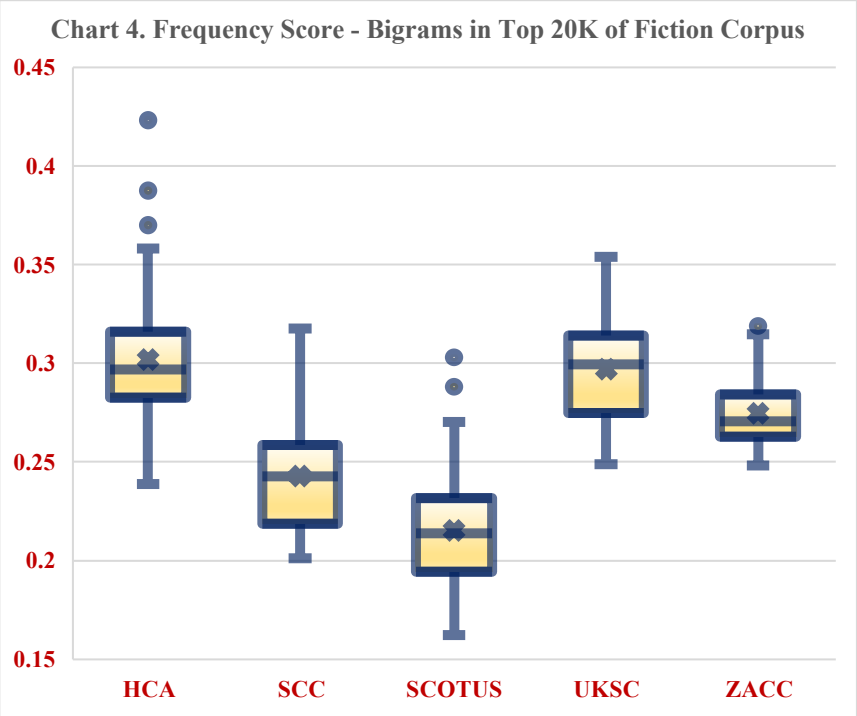
Table 3. ***Average Frequency for Function Words: COCA Academic Corpus*** **by Jurisdiction**			
Rank	**Apex Court**	***average frequency*** **(Number of Occurrences)**	**Standard Deviation (Number of Occurrences)**
1	SCOTUS	24,378	1,973
2	SCC	26,214	1,661
3	UKSC	26,464	1,517
4	ZACC	27,202	1,566
5	HCA	27,471	1,393

Where this measure identifies the use of academic language that is typically more complicated, specialized, or otherwise difficult to read than other forms of language, one would expect from the data that SCOTUS would produce more readable decisions than the HCA. SCOTUS leads the other apex courts in terms of avoidance of academic language by a very strong margin: the difference between SCOTUS and the SCC (ranked number one and number two respectively) is greater than the difference between the SCC and the HCA (ranked number two and number five).

This data suggests that all of the studied apex courts, other than SCOTUS, might benefit from efforts to use less academic language within their decisions as a means of communicating more effectively with their audiences.

4. Average Proportion of Bigrams – Top 20,000 – COCA Fiction Corpus

Boxplots for all proportion scores of bigrams (two-word phrases), reflecting each individual decision within the present study, are shown in Chart 4, below. This measure is intended to illustrate the extent to which bigrams in judicial decisions overlap with the most commonly-used bigrams from a general literary corpus.



The highest recorded measurement for *average proportion of bigrams* was from an HCA decision,[124] wherein 42.3% of all bigrams contained in that decision could also be found in the top 20,000 bigrams within the literary reference corpus, indicating that the decision tended to use two-word phrases that are more common in English. The lowest value was from a SCOTUS decision,[125] wherein only 16.2% of all bigrams used in the decision were found in the top 20,000 list from the reference corpus. Chart 4 also shows that the UKSC and the HCA have relatively similar interquartile ranges on this measurement. However, the HCA has several high outliers that increase its jurisdiction average (indicated by the "x") above the UKSC's average.

---

[124] *Coughlan v. The Queen*, [2020] HCA 15 (Austl.) (deciding a criminal appeal involving arson and fraud charges).

[125] McKinney v. Arizona, 140 S. Ct. 702 (2020) (deciding an historical sentencing appeal).

The *average proportion of bigrams*, reported in terms of the average percentage of all bigrams from judicial decisions that appear within the list of the top 20,000 most frequent bigrams in the reference corpus, are shown in Table 4, below. The calculation of this score is explained in the next paragraph.

For each judicial decision, an average proportion of bigrams in the top 20,000 reference list is calculated by summing the total number of bigrams from that judicial decision that appear within the reference list, and then by dividing that sum by the total number bigrams contained within the judicial decision. This process (which is computed using TAALES)[126] produces an average proportion of bigrams (top 20K) score for each judicial decision. These individual decision scores are then summed for each jurisdiction and divided by the total number of decisions within that jurisdiction to produce jurisdiction-wide *average proportion of bigrams—Top 20K* scores. Standard deviations are also included. Jurisdictions are ranked from highest *average proportion of bigrams—Top 20K* to the lowest.

Table 4. *Average Proportion of Bigrams – Top 20K* by Jurisdiction

Rank	Apex Court	Average proportion (Percent)	Standard Deviation (Percent)
1	HCA	30.2	3.7
2	UKSC	29.7	2.5
3	ZACC	27.5	1.7
4	SCC	24.3	2.6
5	SCOTUS	21.5	2.7

To the extent that using common two-word phrases will facilitate reading comprehension and reduce mental processing loads, the above data suggest that HCA decisions will have the

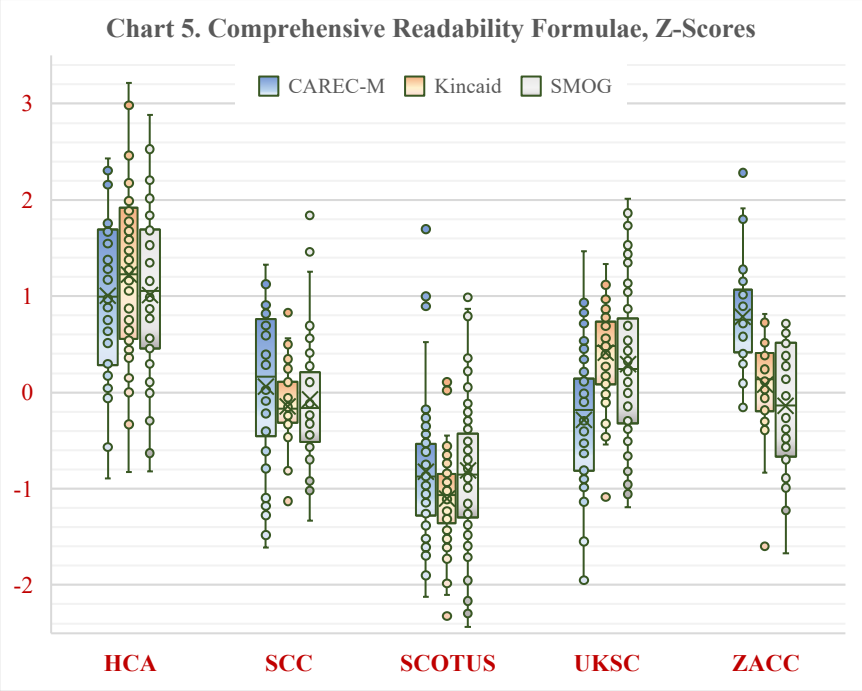
[126] *See NLP Tools for the Social Sciences*, *supra* note 98.

greatest familiarity effect, and SCOTUS decisions will have the least effect. Again, the difference between the courts with the highest and lowest proportions of *Top 20K* bigrams is rather large; to enhance this element of readability, both the SCC and SCOTUS could benefit from efforts to construe phrases in more common and familiar ways, as both the HCA and UKSC have shown is possible. Notably, the standard deviations are small across the studied apex courts, which suggests that each of the courts is internally consistent in the extent of its uses of more common bigrams from one decision to the next.

5. Readability Scores: CAREC-M, Flesch-Kincaid, and SMOG

Boxplots of comprehensive readability scores, reflecting each individual decision within the present study, are shown in Chart 5, below. Unlike the previous charts, however, the scores in Chart 5 are standardized z-scores,[127] allowing for easy visual comparisons between the three different readability measures—even though the scores do not originally use a common scale of measurement. Thus, the zero-line that runs horizontally across Chart 5 represents the location of the average *CAREC-M*, *Flesch-Kincaid*, and *SMOG* scores within the study. The boxplots show how far away each decision is from the study's averages (either above or below), in addition to showing the interquartile range for z-scores (the boxes), the median z-score (the central horizontal line within each box), and the average z-score (the "x" within each box) for each measure in each jurisdiction.

[127] Z-scores measure how far a particular raw score is from the average score for the entire sample in terms of standard deviations.



Looking at Chart 5, normalized readability results seem similar regardless of which formula is used—particularly at the HCA, the SCC, and SCOTUS. This point is reinforced through the use of statistical tests. Specifically, the correlations between readability results for each case within the study, based on the different readability formulae used, signify that the formulae are strongly correlated with one another. The Pearson's correlation coefficient for *SMOG* and *Flesch-Kincaid* is 0.749; the correlation for *CAREC-M* and *Flesch-Kincaid* is 0.571; and, the correlation for *CAREC-M* and *SMOG* is 0.396; $p < 0.01$ for all correlations.[128] These results

[128] Pearson coefficients range in value from -1 to +1. A score of 1 signifies a perfect correlation, and a score of 0 signifies that the variables are not correlated at all. Where all these readability correlation values are positive, the correlation is positive: an increase in one measure of readability would correspond with an increase in each other measure of readability. Where each of the correlations is close to or greater than 0.5, they can be classified in this context as strong correlations. Where the p-value is less than 0.05 in all of the above cases, it can also be said that the correlations are statistically significant.

indicate strong (or moderate strength, in the case of *CAREC-M* and *SMOG*) and statistically significant correlations that provide a degree of mutual reinforcement for the readability results. Thus, while *CAREC-M* scores are used in this Article as the relevant readability measure for the ensuing comparative and statistical analyses examining jurisdiction- and court-specific variables within this study, one can be confident that these *CAREC-M* scores are valid measures of readability in part due to their strong correlations with other readability measures that have longer histories within the field of applied linguistics.

With respect to individual measurements from Chart 5, the least readable decisions were from the HCA, with non-standardized (actual) scores of 0.472 (*CAREC-M*),[129] 20.2 (*Flesch-Kincaid*),[130] and 18.3 (*SMOG*).[131] The most readable decisions were from SCOTUS, with non-standardized scores of 0.171 (*CAREC-M*),[132] 7.2 (*Flesch-Kincaid*),[133] and 9.3 (*SMOG*).[134] Interestingly, each readability formula pointed to different decisions as the most- and least-readable decisions, but the single most readable decisions were from the United States regardless of what formula was used to measure readability, and the single least readable decisions were from Australia—again, regardless of the formula that was used.

The jurisdiction-specific average results for each of the three comprehensive readability formulae are reported in Table 5, below. For both *Flesch-Kincaid* and *SMOG* scores, the results are reported as average grade level scores. For *CAREC-M*, the results are reported as raw scores. The average measure (i.e., the sum of every individual decision's score within the jurisdiction, divided by the number of decisions in that jurisdiction) is reported for all of the

---

[129] *Northern Land Council v. Quall* [2020] HCA 33 (Austl.) (deciding an administrative law case involving questions of delegation in the context of Aboriginal land rights legislation).

[130] *Chetcuti v. Commonwealth of Australia* [2020] HCA 42 (Austl.) (deciding an immigration law case).

[131] *ABT17 v. Minister for Immigration and Border Protection* [2020] HCA 34 (Austl.) (deciding an administrative law case in the context of immigration and border protection legislation).

[132] Rodriguez v. FDIC, 140 S. Ct. 713 (2020).

[133] Davis v. United States, 140 S. Ct. 1060 (2020).

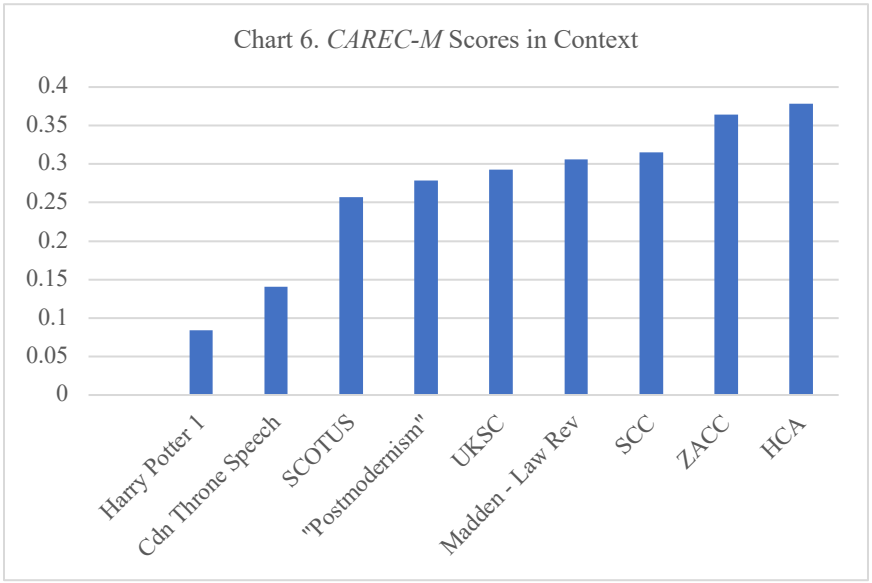
[134] Lomax v. Ortiz-Marquez, 140 S. Ct. 1721 (2020).

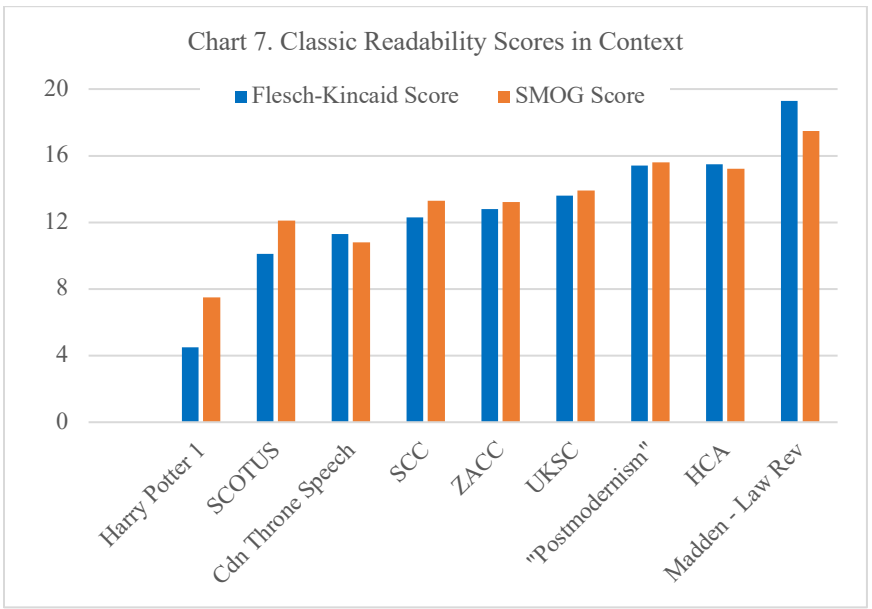
scores. Jurisdictions in Table 5 are ranked from lowest (most readable) *CAREC-M* score to highest (least readable), with associated *Flesch-Kincaid* and *SMOG* scores in subsequent columns.

Table 5. Comprehensive Readability Scores by Jurisdiction			
Apex Court	***CAREC-M* Score**	***Flesch-Kincai*d Score**	***SMOG* Score**
SCOTUS	0.257	10.1	12.1
UKSC	0.293	13.6	13.9
SCC	0.315	12.3	13.3
ZACC	0.364	12.8	13.2
HCA	0.378	15.5	15.2

To put these results into context, Charts 6 and 7, below, show how jurisdiction averages for readability scores, using *CAREC-M* (Chart 6) and *Flesch-Kincaid*/*SMOG* (Chart 7), compare against sample texts from different domains.[135]

[135] The sample texts consisted of the following: J.K. ROWLING, HARRY POTTER AND THE PHILOSOPHER'S STONE (1998); Gary Ayleworth, *Postmodernism*, *in* STANFORD ENCYCLOPEDIA OF PHILOSOPHY (Edward N. Zalta ed. 2015); Julie Payette, Governor General of Canada, Speech from the Throne (September 23, 2020) https://www.canada.ca/en/privy-council/campaigns/speech-throne/2020/speech-from-the-throne.html [https://perma.cc/S5WH-RPFQ]; and, Mike Madden, *Of Wolves and Sheep: A Purposive Analysis of Perfidy Prohibitions in International Humanitarian Law*, 17 J. CONFLICT & SEC. L. 439 (2012) (winner of the American Society of International Law's 2013 Baxter Prize for a paper that significantly enhances the understanding and implementation of the laws of war, but clearly not a contender for any readability awards).





Several points about the comprehensive readability data in Charts 5, 6, and 7, and in Table 5, above, are worth noting. To begin with, although some fluctuation exists in terms of where the UKSC,

the SCC, and the ZACC rank as the middle three apex courts (depending on which readability formula is used), the results are consistent in ranking SCOTUS as the court with the most readable decisions, and the HCA as the court with the least readable decisions. The highest and lowest ranking courts occupy these places by clear margins regardless of which of the three comprehensive readability measures is used.

Accordingly, the general consistency of readability results can be contrasted with the much more variable results that are seen using any of the unidimensional linguistic indicators reported in Tables 1 through 4. The orders in which jurisdictions appear in these tables are all different. None of Tables 1 through 4 produce the same order as Table 5. This lack of correspondence between any single linguistic indicator and any comprehensive readability formula perhaps reflects the inherent weakness in the use of any one criterion to assess and predict the likely comprehensibility of a text. Reading theory suggests that comprehension is affected by many factors,[136] so a unidimensional linguistic indicator is probably incapable of accurately generating relative readability results in the same way that multidimensional formulae can generate such results. For this reason, among others, the present study accepts the new and sophisticated *CAREC-M* measure as the most useful readability measure, as well as accepts the *CAREC-M* measure as the dependent variable at the center of the ensuing comparative and statistical analyses.

The visual presentation of the data in Chart 5 also indicates that—for the most part—standard deviations for all measures and for all jurisdictions are relatively small. This finding suggests that each apex court tends to produce decisions with readability scores that are somewhat narrowly clustered around the court's average readability score. In other words, each apex court seems to have a "readability comfort zone" from which the court does not substantially depart in most decisions.

Some parts of the results in Table 5 can be compared with results reported in previous studies. For instance, Professor Johnson found that the mean *Flesch-Kincaid* grade level for SCOTUS opinions

[136] McNamara & Graesser, *supra* note 34, at 197.

during the 1931 to 1933 terms was 12.19 and was 13.30 during the 2009 to 2011 terms.[137] Comparing the results in Table 5 with Johnson's results, SCOTUS produced noticeably more readable decisions in 2020 than in both previously reported periods. Similarly, Whalen's study of annual average *SMOG* scores for SCOTUS decisions, discussed in Part II.B. above, shows that these scores ranged in value from approximately 13.5 to 14.5 during the most recent ten-year period of the study, conducted in the early 2000s.[138] The differences between these values and the score reported in Table 5 might signal either a downward trend in scores during more recent years or the existence of an outlier year in 2020 for SCOTUS. However, another possibility is that different methods contributed to the score differences: the present study removed all footnoted text from decisions prior to running the decisions through NLP software. If Whalen's study included footnotes, that inclusion may have driven *SMOG* scores higher (if, for instance, one speculates that footnotes are not as carefully constructed by authors to be as readable as the main body of a decision).

Perhaps the most important point to draw from Table 5 and Chart 5 is that apex courts that appear to perform the same functions within substantially similar common law legal systems issue decisions that are widely different in terms of their readability levels. This Article explores possible comparative explanations for this phenomenon in more detail below.

B. Discussion: Comparative Analysis of Readability Results

Before one can begin to assess whether readability variances across apex courts from the different jurisdictions can be explained by institution- or jurisdiction-specific factors, one must first identify and—to the extent possible—quantify each of these factors or variables for each jurisdiction. The following subparts present and discuss the potential readability impact of the comparative variables included in this study relating to the HCA, the SCC, the ZACC, the UKSC, and SCOTUS, and each court's broader operating environment.

---

[137] Johnson, *supra* note 115, at 58.
[138] Whalen, *supra* note 59, at 202–04.

1. Adult Secondary School Completion Rate

The percentage of adults in 2018 between the ages of 25 and 64 who did not complete upper secondary school in each of the studied jurisdictions is as follows: Canada, 8.38%; United States, 9.18%; Australia, 18.11%; United Kingdom, 20.71%; and, South Africa, 25.83%.[139] If judges tailored readability levels of their decisions to the overall adult education levels within their respective jurisdictions, then one might expect decisions to be the most readable in South Africa, and the least readable in Canada. However, comparing national averages for *adult secondary school completion rate* with apex court averages for *CAREC-M* readability scores from Table 5 shows that little correspondence exists between the two variables. For example, the HCA is ranked the lowest on readability, but Australia's adult education level is ranked in the middle; SCOTUS is ranked the highest on readability, but the United States' adult education level is ranked second. From this measure, *adult secondary school completion rate* does not sufficiently explain readability variances across apex courts.

The present study used *adult secondary school completion rate* as a variable to explore connections between the complexity of courts' decision language on the one hand, and the general reading abilities of the population in the jurisdiction on the other hand. While a more direct literacy measure might have been preferable to *adult secondary school completion rate*—itself a proxy measure of literacy—no such direct and common literacy measure exists for all five of the studied jurisdictions.

That being said, common and direct adult literacy measures are available for four of the selected jurisdictions. In a 2012 OECD adult literacy study, the following raw scores were reported (from highest-to-lowest literacy levels): Australia, 280.4; Canada, 273.5; United Kingdom, 272.5; and, United States, 269.8.[140] No data was available

---

[139] *Adult Education Level*, ORG. ECON. COOP. & DEV., https://data.oecd.org/eduatt/adult-education-level.htm [https://perma.cc/S8J8-FY6Y] (last visited Oct. 4, 2021).

[140] STATISTICS CANADA, SKILLS IN CANADA: FIRST RESULTS FROM THE PROGRAMME FOR THE INTERNATIONAL ASSESSMENT OF ADULT COMPETENCIES (PIAAC) 79, Tbl. B.1.1 (2013).

for South Africa. Reading these results along with the *CAREC-M* results from Table 5, the higher adult literacy rankings (in the four studied jurisdictions for which literacy data exists) correspond precisely with lower decision readability levels. The United States has the lowest literacy level and the most readable decisions. Australia has the highest literacy level and the least readable decisions. In other words, adult literacy levels may help to explain readability variances across jurisdictions—perhaps judges are aware of the general literacy needs of their populations and make efforts to tailor the readability levels of their decisions to match these needs. Further study of this hypothesis across a larger number of jurisdictions for which literacy data is available would be helpful.

2. Clerk Involvement

It is difficult to precisely quantify the extent to which law clerks participate in drafting judicial decisions; the question has perhaps been most carefully considered in the American context. For instance, Professors Rosenthal and Yoon used NLP techniques to study the use of function words by SCOTUS Justices within their opinions and, from the high (and increasing) variability in the use of function words across decisions by individual Justices, found that clerks were likely responsible for much of the authorship of SCOTUS decisions.[141] More generally, in the United States, law clerks regularly draft opinions for judges,[142] and, as a result, "the judge has [essentially] been transformed from a craftsman to an editor."[143] SCOTUS Justices employ four clerks each (except for the Chief Justice, who employs five clerks)[144] and can therefore draw from their law clerks to greater extents than judges of other apex courts with fewer clerks.

In contrast, HCA associates (the Australian equivalent of United States law clerks) do not directly participate in drafting decisions.

---

[141] Jeffrey S. Rosenthal & Albert H. Yoon, *Judicial Ghostwriting: Authorship on the Supreme Court*, 96 CORNELL L. REV. 1307, 1337–39 (2011).

[142] John Leubsdorf, *The Structure of Judicial Opinions*, 86 MINN. L. REV. 447, 487 (2001).

[143] J. Daniel Mahoney, *Foreword: Law Clerks: For Better or For Worse*, 54 BROOK. L. REV. 321, 339 (1988).

[144] Todd C. Peppers, *Of Leakers and Legal Briefers: The Modern Supreme Court Law Clerk*, 7 CHARLESTON L. REV. 95, 107 (2012).

However, associates may be involved in revising and proofreading drafts.[145] Moreover, each Justice of the High Court employs two associates,[146] as opposed to the four clerks hired by each SCOTUS Justice.

Law clerks at the ZACC are involved in decision-writing in a manner similar to HCA associates. ZACC law clerks do not typically draft judicial opinions, but instead provide research assistance and cite-check opinions.[147] Commentators note that this cite-checking also involves reading and making suggestions related to "[s]pelling, grammar, format and style" as part of an extremely thorough revision process.[148] Each Constitutional Court Judge has two South African law clerks and may also have one foreign law clerk.[149]

Judicial assistants in the United Kingdom (roughly equivalent to United States law clerks) were only introduced in 2001 at the UKSC (or its precursor court), and Judges of the UKSC are still experimenting with ways to best use their assistants.[150] Judicial assistants do not draft decisions,[151] and a recent statement of their duties included on the UKSC's recruiting website did not mention work reviewing, revising, or cite-checking decisions.[152]

---

[145] Katharine G. Young, *Open Chambers: High Court Associates and Supreme Court Clerks Compared*, 31 MELB. U. L. REV. 646, 660 (2007).

[146] *Id.* at 658.

[147] See *About Law Clerks*, CONST. CT. OF S. AFR. https://www.concourt.org.za/index.php/law-researchers/about-law-clerks [https://perma.cc/5WVT-6XAB] (last visited Oct. 4, 2021), for a description of the roles and responsibilities of law clerks at this court.

[148] Hugh Corder & Jason Brickhill, *The Constitutional Court of South Africa*, *in* THE JUDICIARY IN SOUTH AFRICA 355, 372 (Cora Hoexter & Morné Olivier eds. 2014).

[149] *About Law Clerks*, *supra* note 147; *see also* Corder & Brickhill, *supra* note 148, at 370 (noting that six of the Court's judges will typically have a foreign clerk, usually from the United States).

[150] Nina Holvast, *The Power of the Judicial Assistant/Law Clerk: Looking Behind the Scenes at Courts in the United States, England and Wales, and the Netherlands*, 7 INT'L J. FOR CT. ADMIN. 10, 20 (2016).

[151] *Id.* at 22.

[152] Hays Recruiting Experts, *Person Specification – UKSC Judicial Assistants 2021/22*, https://microcontrib.hays.com/documents/4856148/0/JAPersonSpecandJD2021.pdf [https://perma.cc/XKR5-A9DS] (last visited Oct. 16, 2021).

Additionally, there are only eight judicial assistants for the entire Court; thus, some of the judges do not use judicial assistants.[153]

At the SCC, law clerks have been institutionalized in somewhat the same manner as SCOTUS. Each Canadian judge now hires four law clerks, as is typical in the United States.[154] An empirical study of variability of judges' writing styles from year-to-year suggests that most SCC judges likely rely on their clerks to draft opinions at least some of the time.[155] This study is consistent with Professor Sossin's prior descriptive account of the work performed by law clerks at the SCC, wherein Sossin notes that clerks regularly work on, or write, draft decisions.[156]

Based on the above information about how many clerks may be involved in drafting judicial decisions, and how actively clerks might participate in writing decisions, *clerk involvement* is apparently lowest in (1) the United Kingdom, then (2) Australia, followed by (3) South Africa, then (4) Canada, and is highest in (5) the United States. From these rankings, alternative inferences can be drawn. If one believes that judicial decisions are likely to be more readable when judges use a highly collaborative drafting process involving one or more clerks, then one might expect the United States to produce the most readable decisions. In contrast, if one suspects that less-experienced law clerks would be apt to use more complicated language and communication styles (perhaps to prove their worth or demonstrate their intelligence),[157] then one might expect the United Kingdom to produce the most readable decisions.

---

[153] Holvast, *supra* note 150, at 22–24.

[154] *Law Clerk Program*, SUP. CT. OF CAN. (Feb. 4, 2021), https://www.scc-csc.ca/empl/lc-aj-eng.aspx [https://perma.cc/7CUA-4R83].

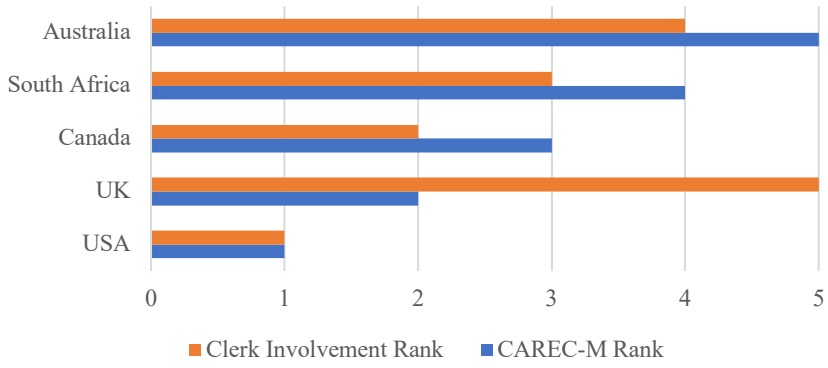
[155] Kelly Bodwin, Jeffrey S. Rosenthal & Albert H. Yoon, *Opinion Writing and Authorship on the Supreme Court of Canada*, 63 U. TORONTO L.J. 159, 186 (2013).

[156] Lorne Sossin, *The Sounds of Silence: Law Clerks, Policy Making and the Supreme Court of Canada*, 30 U. BRIT. COLUM. L. REV. 279, 296–98 (1996).

[157] This theory seems to be espoused in a similar context by RICHARD A. POSNER, THE FEDERAL COURTS: CHALLENGE AND REFORM 156 (rev. ed., 1999) (suggesting that law clerks are the "the proximate cause of the increasing prolixity of federal judicial opinions. The law clerks have time to write at length and a fondness for the apparatus of scholarship – footnotes and citations – that is natural in those who have just emerged from their academic chrysalis").

When these results for *clerk involvement* are compared with *CAREC-M* results from Table 5, some correspondence exists between the level of involvement of law clerks in decision drafting processes and readability, as shown in Chart 8, below.

Chart 8. Comparing Ranks: *CAREC-M* & Clerk Involvement



The United States has the highest level of clerk involvement and the most readable decisions. Canada is lower than the United States, but higher than South Africa and Australia on both measures, and Australia is lower than the United States, Canada, and South Africa on both measures. The problem with the above findings is the United Kingdom's results: the UKSC has the lowest level of *clerk involvement* but the second most-readable level of *CAREC-M* scores. If the United Kingdom is disregarded as an outlier (for instance, because the court has developed other highly effective means of producing readable decisions despite its low reliance on law clerks in the decision-drafting process), then *clerk involvement* seems to explain some of the variance in readability scores across apex courts.

The relationship between *clerk involvement* and *Flesch-Kincaid* scores from Table 5 is somewhat easier to assess. Comparing these two sets of measures shows that *clerk involvement* and *Flesch-Kincaid* ranks correspond almost exactly, with only Australia and the United Kingdom each "off" by one rank across the two measures. In other words, *clerk involvement* seems to explain

Flesch-Kincaid variances even more effectively than *CAREC-M* variances.

Based on the above discussion regarding *clerk involvement*, it seems that a more collaborative decision-drafting process using law clerks tends to at least correspond with an increase in the kinds of surface-level readability measures (words per sentence, and syllables per word) that contribute to the *Flesch-Kincaid* grade level score but corresponds somewhat less with increases in readability levels based on more sophisticated (*CAREC-M*) criteria. One might hypothesize that involving clerks in the drafting process helps courts to identify and reduce the use of long words and sentences but is less helpful in reducing the use of other, more nuanced textual features that more directly reflect levels of text complexity (such as words with a high age of acquisition, or trigrams that are extremely uncommon). Apex courts that rely on clerks to assist judges in increasing the readability levels of their decisions may wish to seek linguistic training opportunities for these clerks wherein the latest advances in readability theory and NLP techniques could be briefly introduced to the clerks.

3. Court Politicization

As with *clerk involvement*, quantifying the extent to which an apex court or its judges are politicized or aligned ideologically with a political party is difficult. Although no study has produced a master index of politicization to describe apex courts worldwide, Professor Weiden has comparatively studied politicization levels of the Canadian, American, and Australian apex courts.[158] In his study, Weiden looked first at the extent of partisan and non-partisan appointments of judges (by comparing an ideology score for each appointed judge with the ideology of the government party that appointed the judge)[159] to these apex courts between 1990 and 1999, finding that the United States had the highest proportion of partisan appointments (0.917), then Australia (0.8), and then Canada (0.5).[160] Weiden noted that these politicization results "comport with the

[158] David L. Weiden, *Judicial Politicization, Ideology, and Activism at the High Courts of the United States, Canada, and Australia*, 64 POL. RSCH. Q. 335 (2011).
[159] *Id.* at 337–38.
[160] *Id.* at 338.

scholarly consensus regarding the supreme courts of the United States, Canada, and Australia."[161] Weiden's study then looked at the extent to which each of the apex courts tended to decide cases along ideological lines and found that the tendency was most apparent at SCOTUS, then at the HCA, and was least apparent at the SCC.[162]

Unfortunately, no equivalent study has considered the politicization levels of the UKSC or the ZACC, so reliance on other commentary (leading to subjective relative assessments of politicization) is needed in this study to complete the categorization of the selected apex courts in terms of *court politicization*. The UKSC is generally accepted as being at the lowest end of the politicization spectrum.[163] The ZACC, in contrast, is considered to be somewhat similar to the SCC in terms of how progressive the two courts are in advancing different political agendas.[164] However, judges of the ZACC notably tend to have strong ties with the ruling party (which was advancing a progressive agenda as of 2018),[165] so the ZACC would appear to be slightly more politicized than the SCC.

Based on the above studies and descriptions, *court politicization* amongst the apex courts can logically be characterized as being lowest in (1) the United Kingdom, then (2) Canada, (3) South Africa, (4) Australia, and, highest in (5) the United States. If one believes that more highly politicized courts are likely to produce

---

[161] *Id.* That the SCOTUS is the most highly politicized court of the three is likely not a surprise to most readers. As between the SCC and HCA, Weiden's study seems to affirm what has been observed by others. *See* Brice Dickson, *Comparing Supreme Courts*, *in* JUDICIAL ACTIVISM IN COMMON LAW SUPREME COURTS 1, 3 (Brice Dickson ed., 2007) (noting with citations to three other studies that the HCA is "often described as a very 'political' court").

[162] Weiden, *supra* note 158, at 340.

[163] *See, e.g.*, Dickson, *supra* note 161, at 12 ("British judges are notorious for doing what parliament tells them to do."); *id.* at 17 ("In the United Kingdom the top court has been relatively free from political criticism since it was established in its modern form in the 1870s . . . . [T]he Lords of Appeal have generally speaking enjoyed a reputation as careful analysts and loyal implementers of the doctrine of the sovereignty of Parliament.").

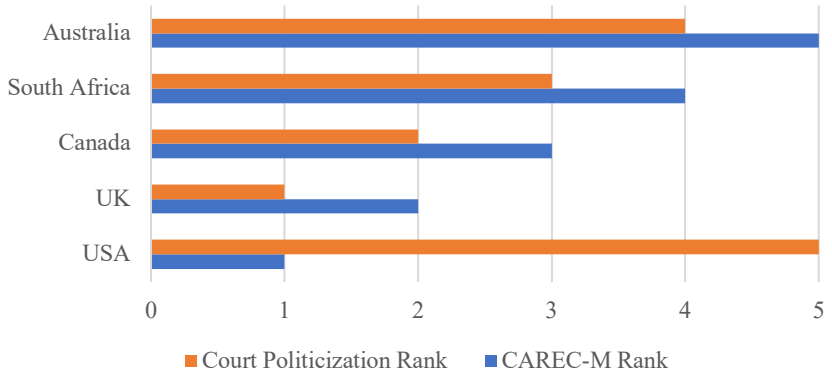
[164] *Id.* at 11–12.

[165] David Landau, *Courts and Support Structures: Beyond the Classic Narrative*, *in* COMPARATIVE JUDICIAL REVIEW 226, 231–32 (Erin F. Delaney & Rosalind Dixon, eds., 2018).

more readable decisions (perhaps because the judges of these courts are particularly concerned about how their rulings may be understood and received by members of the polity with whom the judges are ideologically aligned—including fewer literate members), then one would expect that SCOTUS decisions would be the most readable, and that UKSC decisions would be the least readable.

When these results for *court politicization* are compared with *CAREC-M* results from Table 5, there appears to be almost no correspondence between the level of politicization of an apex court and the readability level of that apex court's decisions, as shown in Chart 9, below.

Chart 9. Comparing Ranks: *CAREC-M* & Court Politicization



Apart from the United States, which has the highest level of both politicization and readability, no other jurisdiction's readability scores seem capable of being explained in terms of court politicization. The present study suggests that no strong correlation exists between politicization levels and readability; however, since the measures used to gauge politicization levels in the present study were, admittedly, somewhat *ad hoc*, a more comprehensive and objective assessment of politicization levels in a future study would be helpful in confirming or rejecting the inference of non-

correspondence drawn from the current study in relation to those two variables.

4. Panel Size

The results for *panel sizes* at the selected courts are shown in Table 6, below, ranked from smallest average panel size to largest average panel size.

Table 6. *Panel Sizes* at Apex Courts			
Apex Court	***panel sizes* Used**	**Most Common *panel size***	**Average *panel size***
HCA	1; 3; 5; 7	5	5
UKSC	3; 5; 7	5	5.1
SCC	5; 7; 9	9	8.6
SCOTUS	8; 9	9	8.9
ZACC	7; 8; 9; 10; 11	10	9.4

If one believes that decisions may be more readable when emerging from larger panels (perhaps because judges strive to write more readably to sway or persuade their peers in order to secure a majority), then one would expect the ZACC to produce the most readable decisions and the UKSC and the HCA to produce the least readable decisions.

From a visual inspection of the average panel size results in Table 6, alongside the *CAREC-M* scores in Table 5, it is apparent that there is no direct correspondence between these two variables. However, each case within the study provided discrete measures for *panel size* and *CAREC-M* readability score, which makes conducting a statistical correlation test possible to determine the relationship between the two variables. The Pearson's correlation coefficient for *panel size* and *CAREC-M* is -0.177, $p < 0.005$. The negative correlation signifies an inverse relationship between *panel size* and *CAREC-M* (i.e., as panels grow in size, *CAREC-M* scores decrease—reflecting improved

readability). The correlation size indicates a moderate effect size,[166] and the low p-value suggests that this low-moderate correlation is statistically significant.

The above results suggest that larger panels are associated with more readable decisions. For a court like the HCA, there appears to be little room to increase panel size (since the court only consists of seven judges), although the Chief Justice could likely rely upon full panels more often, instead of five-judge panels. For the UKSC, however, there is ample scope to increase panel sizes. The UKSC currently consists of eleven judges and typically consists of twelve judges, but most frequently has used panels of only five judges. The President of the UKSC could experiment with larger panels more often in an attempt to improve the Court's readability scores.

5. Former Law Professors per Judge

The results for the average number of *former law professors per judge* on each case heard by the selected apex courts are shown from highest to lowest, in Table 7, below.

Table 7. *Former Law Professors per Judge* at Apex Courts	
Apex Court	**Average *former law professors per judge* on a Panel**
SCOTUS*	0.33
SCC*	0.33
UKSC	0.19
HCA	0.14
ZACC	0.00

An asterisk * – denotes a tie.

One might expect that having former law professors on a panel would elevate the linguistic level (the complexity of language used) in the decision—perhaps because academic language tends to be more complicated and less understandable to the general population

[166] JACOB COHEN, STATISTICAL POWER ANALYSIS FOR THE BEHAVIORAL SCIENCES 80 (2d ed., 1988) (explaining a correlation with a strength of between 0.1 and 0.5 to be one with a moderate effect size).

than other more-common forms of language. As such, one would expect the SCC and SCOTUS to produce the least readable decisions and the HCA and the ZACC to produce the most readable decisions.

The average *former law professors per judge* results in Table 7 alongside the *CAREC-M* scores in Table 5 indicate some correspondence between these two variables—but not in the way previously suggested. The courts with the most law professors on panels produced more readable decisions, rather than less readable decisions. Again, each case within the study provided discrete measures for both variables, allowing for statistical correlation tests to be performed to determine the relationship between these two variables. The Pearson's correlation coefficient for *former law professors per judge* and *CAREC-M* is -0.364, $p < 0.005$. The negative correlation tells us that as the proportion of former law professors on a panel grows, *CAREC-M* scores decrease—reflecting improved readability. The correlation size indicates a correlation of moderate strength,[167] and the low p-value suggests that the correlation is statistically significant.

As these statistics show (perhaps counterintuitively), having former law professors on an apex court panel seems to make a positive difference in terms of readability. Accordingly, rather than contributing to an elevation in language complexity, former law professors actually appear to reduce complexity. Perhaps law professors' experiences of distilling complex legal concepts into easily understood cognitive packages for the benefit of law students carries through to the bench, such that the presence of former law professors on panels helps the authoring judges for the panels to write more readable decisions.

While a court or a Chief Justice likely cannot manipulate this *former law professors per judge* variable to any significant extent without creating burnout for the judges who have experience as law professors, or an inequitable assignment of judicial duties, the executive branch of government in each jurisdiction can likely capitalize on the link between law professor experience and higher readability scores. In particular, a government that is serious about

[167] *Id.*

improving the readability of judicial decisions produced by its apex court may wish to give some consideration to the idea of privileging, during the judicial selection and appointment processes, those candidates who have past professorial experience. This strategy might be especially helpful in South Africa, where there are currently no former law professors appointed to the ZACC.

6. Degrees per Judge

The results for average number of *degrees per judge* on each case at the selected courts are shown from lowest to highest, in Table 8, below.

Table 8. *Degrees per Judge* at Apex Courts	
Apex Court	**Average *degrees per judge* on a Panel**
UKSC	1.62
ZACC	2.11
HCA	2.28
SCOTUS	2.33
SCC	2.55

This variable was initially included within the study, in much the same way that *former law professors per judge* was included within the study, based on a hypothesis that a greater number of more-educated or academically-experienced judges on panels would be associated with higher levels of language complexity within the decisions. From the above discussion relating to *former law professors per judge*, however, one might now hypothesize that having more educated judges on a panel would enhance, rather than reduce, readability.

From a visual inspection of the average *degrees per judge* results in Table 8 alongside the *CAREC-M* scores in Table 5, there does not appear to be any meaningful correspondence between the two variables. The discrete measures for these variables were tested for correlation. The Pearson's correlation coefficient for *degrees per*

judge and *CAREC-M* is 0.09, p = 0.175. This correlation is so weak as to be practically non-existent.[168] Furthermore, the high p-value suggests that the results of the correlation test are statistically insignificant. In other words, the data relating to these two variables do not have a meaningful association that can provide insight into readability variances across apex courts.

7. Women per Judge

The results for average number of *women per judge* on each case at the selected courts are shown from highest to lowest, in Table 9, below.

Table 9. *Women per Judge* at Apex Courts	
Apex Court	**Average *women per judge* on a Panel**
ZACC	0.49
SCC	0.44
HCA	0.38
SCOTUS	0.32
UKSC	0.21

This variable was included within the study based on research suggesting that women tend to write more readably than men[169] and, therefore, having more women on a panel would be associated with higher readability levels of decisions produced by the panel. If this hypothesis were true, then one would expect the ZACC and the SCC to produce the most readable decisions, with SCOTUS and the UKSC producing the least readable decisions.

The average *women per judge* results in Table 9 alongside the *CAREC-M* scores in Table 5 show some correspondence between

[168] *Id.* at 79–80 (noting that the effect size of correlations that are lower in strength than 0.1 cannot even be characterized as small).

[169] Hengel, *supra* note 89, at 80–82. Hengel's work uses five readability measures to conclude that articles written by women in key economic journals are more readable than articles written by men—despite many gender biases that exist within the academic publishing world. *See id.*

these two variables—but not in the way that might be expected. The UKSC and SCOTUS had the lowest proportion of women on their panels, but the highest decision readability levels. Since each case had discrete measures for both variables, correlation was assessed between these two variables. The Pearson's correlation coefficient for *women per judge* and *CAREC-M* is 0.266, $p < 0.005$. The positive correlation tells us that, as the proportion of women on a panel grows, *CAREC-M* scores also increase, reflecting a lower readability level. The correlation size is of moderate strength[170]—although this correlation is stronger than that which exists between *panel size* and *CAREC-M*. The low p-value suggests that the correlation is statistically significant.

This correlation is difficult to explain; however, as acknowledged, these results do not reflect the demographic characteristics of the individual authors of judicial decisions—only of the entire group of judges who comprised the panel. In that respect, one can reconcile research showing that women produce more readable texts than men with the above results (showing that higher numbers of women on a panel correlate with less readable decisions). The present study did not consider whether the authors of decisions identify as men or women. However, no obvious explanation for the correlation between higher numbers of *women per judge* on a panel and lower readability levels has been presented.

Because of this correlation's counterintuitive and inadequately explained nature, one should be cautious to propose legal or policy interventions intended to improve readability by referencing this correlation. Further research about how different gender balances on judicial panels can affect decision readability levels (and judicial decision-making more generally) would be helpful.

8. Multivariable Modeling to Explain Readability Variances

The above discussion attempts to explain how each comparative variable, in isolation, associates with readability levels for the studied apex courts. By focusing more specifically on relevant variables for which there were statistically significant correlations, this study, through regression analysis, developed a multivariable

[170] COHEN, *supra* note 166, at 80.

model that addresses readability variances across jurisdictions in a more comprehensive manner.

Specifically, a multiple regression analysis was run to explain the variance of *CAREC-M* scores (the dependent variable) from the comparative (independent) variables *panel size*, *former law professors per judge*, and *women per judge*. The multiple regression model, with all three independent variables, produced a coefficient of determination, $R^2 = 0.238$, $F(3,229) = 23.80$, $p < .005$.[171] All three variables added statistically significantly to the model, $p < .05$. Regression coefficients and standard errors can be found in Table 10, below.

Table 10. Multiple Regression Results for *CAREC-M*

CAREC-M	***B***	**95% *CI* for *B***		***SE B***	***β***	**R^2**	**ΔR^2**
		LL	***UL***				
Model						0.237	0.227
Constant	0.343*	0.314	0.371	0.015			
Panel size	-0.009*	0.013	-0.005	0.002	-0.293*		
Former law professors per judge	-0.110*	-0.165	-0.560	0.028	-0.247*		
Women per judge	0.170*	0.107	0.232	0.032	0.361*		

Note. Model = "enter" method in SPSS Statistics. ***B*** = unstandardized regression coefficient. ***CI*** = confidence interval. **LL** = lower limit. **UL** = upper limit. ***SE B*** = standard error of the coefficient. ***β*** = standardized coefficient. **R^2** = coefficient of determination. **ΔR^2** = adjusted coefficient of determination. * $p < 0.001$

[171] There was linearity as assessed by partial regression plots and a plot of studentized residuals against the predicted values. There was independence of residuals, as assessed by a Durbin-Watson statistic of 1.579. There was homoscedasticity, as assessed by visual inspection of a plot of studentized residuals versus unstandardized predicted values. There was no evidence of multicollinearity, as assessed by tolerance values greater than 0.1 (all tolerances were, in fact, greater than 0.7). There were no studentized deleted residuals greater than ±3 standard deviations, no leverage values greater than 0.2, and no values for Cook's distance above 1 (all distances were < 0.5). The assumption of normality was met, as assessed by a Q-Q Plot.

For the present study, the most important information in Table 10 is the R^2 value for the regression model: 0.238. This figure signifies that the model—using *panel size*, *former law professors per judge*, and *women per judge* as predictors of *CAREC-M* scores—can explain 23.8% of the variance in these scores. This R^2 value can be classified as a moderate effect size.[172] In other words, these court-specific variables that are distinct within each jurisdiction can explain a moderate (but statistically significant) extent of readability variances across apex courts from different common law jurisdictions.

V. CONCLUSION

The study described in this Article presents several illuminating findings. First, the results offer a starting point for further research into readability levels of court decisions in several countries within which there are few, if any, studies of decision readability—and within which no such studies use the depth and breadth of NLP measurements, or the *CAREC-M* comprehensive readability formula, that the present study uses.

Second, the results show that there are substantially different ways for apex courts (that all perform similar or analogous functions) to communicate their decisions, in terms of decision length, word concreteness, the use of academic language, and the use of more or less common two-word phrases. Courts or governments that are concerned with increasing the readability levels of their decisions can look to comparable jurisdictions to see, in many cases, where there is relative room for gains to be made in specific categories of language usages. Additionally, to the extent that any efforts to increase readability levels within a particular jurisdiction might face resistance from judges or courts (who may feel that their decisions are already communicated as effectively as possible to their audiences), the comparative results of the present

[172] *See* COHEN, *supra* note 166, at 413–14 (suggesting that, as general guidance across behavioral and social science fields such as sociology, psychology, and economics, a 2% proportion of variance explained ("PV") would be a small effect size, a 13% PV would be a moderate effect size, and a 26% PV would be a large effect size).

study offer compelling evidence that the ways in which any one court communicates today are not necessarily the only, or the best, ways to communicate. In this sense, some comparative "peer pressure" could be a useful force for change in encouraging lower-performing apex courts to dedicate more attention to the readability levels of their decisions.

Third, from observational (i.e., non-statistical) perspectives, this study highlights where factors, such as clerk involvement and education or literacy levels within a population, may help to explain readability variances across apex courts. The results and discussion in this Article addressing these points will be important to government and court officials who care about readability levels. For instance, those who wish to increase the readability of judicial decisions will likely (based on this study) want to explore how law clerks could be used more actively in drafting decisions. Those who want to maintain the readability *status quo* in a particular jurisdiction can (based on this study) suggest that current readability levels are calibrated to match literacy levels within the jurisdiction, and that comparative measurements across apex courts do not provide appropriate readability targets because these measurements must also be understood in light of the different education and literacy levels in these foreign jurisdictions.

With all of that being said, it must be acknowledged that the extent to which readability variances across apex courts can be statistically explained by court- or jurisdiction-specific factors (like *panel size*, *former law professors per judge*, and *women per judge*) is only 24%, so there are likely a wide range of other variables associated with differing readability levels that future studies could (and should) consider to further explain readability variances. While the present study suggests that comparative factors play a moderate role in explaining readability score variances from one apex court to another, this study did not consider any variables related to the characteristics of the authors of individual judicial decisions, and the impact of these variables on readability scores. Intuitively, one might expect that variables relating to authorship of a decision, rather than to the jurisdiction or court that produced the decision, would be influential in explaining readability scores. Thus, the limited extent of the readability variances explained by the present

study may, as much as anything else, point toward future research questions asking about how author-specific factors could explain readability variances in court decisions.

In the meantime, however, the present study provides a much-needed snapshot of the current readability landscape across the five studied apex courts. The present study cannot explain whether a particular citizen will actually understand all of the salient points contained within, for example, a particular employment law decision (or any other kind of decision) released by a citizen's national apex court, but the study does reveal far more about the quantitative readability of apex court decisions in the five jurisdictions than previously known. With this new data and analysis, perhaps some action can now be taken to more critically assess whether societies are happy with the current judicial decision readability levels, or whether (and what) interventions are needed to enhance readability so that citizens can better understand the law in their respective jurisdictions.