Getting Personal: Individualized and User-Searchable Readability Results for a Large Corpus of Canadian Adjudicative Decisions

*Mike Madden**

This article reports on an original empirical study of the quantitative readability of a large corpus of Canadian court and tribunal decisions. Specifically, the article uses a new law-specific readability formula that is designed to predict how well readers of different education levels will understand an adjudicative decision to compute readability scores for 1,621 discrete decision and opinion files, containing over 9.3 million words of text, from cases that were decided in 2022. The results are coded based on author name, author gender, jurisdiction, legal subject area, opinion type, word count, and other variables—and are made publicly available. Although the article highlights and discusses some of the more noteworthy results, its true value lies in its potential as a resource for others. Future researchers can query the results reported within this article for their own purposes to answer their own research questions, based on the variables that are of most interest to them. Similarly, authors of adjudicative decisions whose scores are reported in this study can find and contemplate their scores to assess whether they are satisfied with the readability levels of their decisions. And, other judges or tribunal members can search the results to identify highly readable decisions that might serve as inspiration to these judges in their future decisions.

* Mike Madden, CD, BA, MA, LLB, LLM, PhD. This research was generously funded by a Bombardier Canada Graduate Scholarship – Doctoral, awarded by the Social Sciences and Humanities Research Council. I thank Professors Elizabeth Judge, Suzanne Bouclin, and Wolfgang Alschner, all from the University of Ottawa's Faculty of Law, for their guidance throughout my doctoral studies. Their input significantly improved this article.

Introduction

In a previous article,[1] I reported on the creation of Madden's Adjudicative Decision Readability Score (MADRS)—a law-specific readability formula that predicts the extent to which Canadian audience members of different education levels are likely to understand court and tribunal decisions based on the language used within these decisions. In the same study, I also applied the MADRS formula to the full set of Supreme Court of Canada decisions from 2022 to demonstrate and discuss average readability scores for the different judges. Although the Supreme Court of Canada is a fascinating site of study for legal scholars, our tendency to focus on the things that happen at this court often means that we fail to look at how the law is applied and how justice is dispensed at other court levels and at administrative tribunals in Canada in the same meaningful detail. This article therefore attempts to call attention to the readability levels of English adjudicative decisions from across Canada, at all court levels, and from a wide range of administrative tribunals.

Unlike similar readability studies, however, this article also places a more personal spin on applications of the MADRS formula to adjudicative decisions in two ways. The first way is researcher-centric. This article describes and provides readers with access to readability results from a large corpus of decisions that have been individually coded for nine different variables. A

1. Mike Madden, "How Understandable are Adjudicative Decisions? Introducing and Applying Law's Own Readability Formula" (forthcoming 2026), *Legal Writing: The Journal of the Legal Writing Institute*, online: <papers.ssrn.com/sol3/papers.cfm?abstract_id=5110351> [Madden, "How Understandable are Adjudicative Decisions?"].

goal of the article is to give readers the ability to search and make meaning of these MADRS results in different ways that are personally relevant to different readers. The second way in which this article gets personal is author-centric: the readability results contained in this article identify the authoring judge or tribunal member for each decision within the study. As I explain in Part III, by creating a large-scale, personalizable set of readability results, these results can be leveraged by different people for different purposes. A researcher who is primarily interested in case law from Canadian territories, or decisions that are collaboratively written, or decisions involving immigration and refugee law, for instance, can quickly and easily extract readability results contained within this article and consider these results for the researcher's narrower purpose—without needing to undertake a separate readability study. Moreover, any of the 840 different authors of individually written decisions, who are identified by name within the results that follow, can find and contemplate the readability scores for their own decisions within the database of results. Are these results what the authoring judge or tribunal member were hoping to achieve, in terms of a decision's likely ability to be understood by its audience? Are there certain judges or tribunal members whose decisions are written more readably, and who might therefore serve as examples that other authors could emulate? Answers to these questions, and to many others that future researchers might develop, can potentially be found within the results of the study that is described in this article.

In a somewhat unusual sense, at least from the perspective of legal scholarship, this article does not strongly advance an argument—other than the argument that countless insights can be gained from studying the readability levels of adjudicative decisions using the data this article contains. This article, at its core, presents a novel resource for others, and illustrates through several examples how the resource can be manipulated by researchers to identify or answer interesting new research questions.

I. Background—Understanding Readability and MADRS

Before proceeding further, it would be helpful to first briefly describe the MADRS formula and its origins. MADRS was created from human expert assessments about how well people of different education levels would understand text extracts from actual Canadian court and tribunal decisions.[2] Through regression analysis, it was possible to develop this predictive formula that assigns readability scores to adjudicative decisions based on the linguistic properties of the decisions. In other words, the MADRS formula represents

2. *Ibid.*

a computer-based tool that can quantitatively measure the readability levels of Canadian court and tribunal decisions on a massive scale, with relatively little effort. And this formula outperforms other well-known general-purpose readability formulas in predicting human experts' readability scores for adjudicative decision texts.[3]

The formula relies on six linguistic variables to produce a global readability score. Four of the variables measure word-related properties in the text, as described below.

1. ***Kuperman Age of Acquisition (Content Words) / AoA_CW.*** This variable measures the age at which individuals would have understood a particular content word if someone used it in front of them, even if the subject individual did not use, read, or write the word themselves at that age.[4] Content words are words that tend to carry lexical meaning, and include nouns, verbs, adjectives, and some adverbs.[5] The content words "momma", "mama", "mom", and "potty" have the four lowest age of acquisition scores in the relevant database, while the content words "calceolaria", "architrave", "penury", and "thrombocytopenia" are all among the ten highest scoring words.[6]

2. ***Medical Research Council: Imageability (All Words) / MRC_Imag_AW.*** This variable measures the imageability of words used in a text. Imageability refers to the tendency for a word to evoke both a lexical and a visual representation of the word within the mind of a reader or

3. *Ibid*.

4. Victor Kuperman, Hans Stadthagen-Gonzalez & Marc Brysbaert, "Age-of-Acquisition Ratings for 30,000 English Words" (2012) 44 Behavior Research Methods 978 at 980–81. The authors of the study that created Kuperman's age of acquisition scores relied upon ratings from 1,960 American participants, who provided a total of 842,438 age of acquisition ratings for 30,121 words (i.e., each word was rated by approximately 28 different participants in the study).

5. Helen Bird, Sue Franklin & David Howard, "'Little Words' – Not Really: Function and Content Words in Normal and Aphasic Speech" (2002) 15:3–5 J Neurolinguistics 209 at 210.

6. Kuperman, Stadthagen-Gonzalez & Brysbaert's full database was erroneously not included within the supplemental material accompanying their article when it was first published. The database was subsequently made available in a notice of erratum, online: <link.springer.com> [perma.cc/2Z4U-FXT7] (as an ".xlsx" file called "ESM 1") (*supra* note 4).

listener.[7] For instance, the word "beach" (for which most of us have a strong mental image) has the highest imageability score in the relevant database, while the word "plenipotentiary" (the meaning of which I struggle to visualize) has the lowest score.[8]

3. ***Medical Research Council: Concreteness (Function Words) / MRC_ Concr_FW.*** This variable measures the concreteness of function words used in a text. Function words include pronouns, articles, conjunctions, and prepositions that fill grammatical functions, but that do not tend to carry much lexical meaning.[9] Concreteness means the extent to which something can be experienced by the senses.[10] So, "[a]ny word that refers to objects, materials, or persons should receive a high concreteness" score.[11] Concreteness is generally contrasted with abstractness, where highly concrete and highly abstract words fall at opposite ends of the rating spectrum.[12] For instance, the content word "milk" (describing a liquid that can be distinctively experienced by most people through taste, touch, sight, and smell) has the highest concreteness score in the relevant database.[13] In terms of function words, "me", "I", and "mine" have the three highest concreteness scores, while "as", "of", and "however" have the three lowest scores.[14]

---

7. Maya M Khanna & Michael J Cortese, "How Well Imageability, Concreteness, Perceptual Strength, and Action Strength Predict Recognition Memory, Lexical Decision, and Reading Aloud Performance" (2021) 29:5 Memory 622 at 622. See also Max Coltheart, "The MRC Psycholinguistic Database" (1981) 33:4 QJ Experimental Psychol 497. Imageability scores are drawn from the scores that were assigned to 9,240 different words found in Coltheart's Medical Research Council Psycholinguistic database.

8. The University of Western Australia, School of Psychology, "MRC Machine Usable Dictionary. Version 2.00" (1 April 1987), online: <websites.psychology.uwa.edu.au/school/MRCDatabase/mrc2.html> [web.archive.org/web/20260228030551/https://websites.psychology.uwa.edu.au/school/MRCDatabase/mrc2.html].

9. Sidney J Segalowitz & Korri C Lane, "Lexical Access of Function versus Content Words" (2000) 75 Brain & Language 376 at 376; Khanna & Cortese, *supra* note 7 at 623–24; Coltheart, *supra* note 7.

10. Khanna & Cortese, *supra* note 7 at 623–24. Concreteness scores are drawn from the scores that were assigned to 8,228 different words found in Coltheart's Medical Research Council Psycholinguistic database. See Coltheart, *supra* note 7.

11. Khanna & Cortese, *supra* note 7 at 623.

12. *Ibid* at 623–24.

13. The University of Western Australia, *supra* note 8.

14. *Ibid.*

4. ***General Inquirer Database: Legal (Nouns) / GI_Leg_Nouns.*** This variable measures the proportion of words in a text that are found on a defined list of 192 words associated with legal, judicial, or police matters.[15] Examples of nouns contained within the General Inquirer's legal word list (GI_Leg_Nouns) include the following: advocate, auditor, deposition, fugitive, junta, and legislator.

Another of the MADRS variables measures how commonly three-word phrases within a text can be found in a database of frequently used phrases.

5. ***COCA—Spoken (Proportion of Trigrams in the top 30K) / COCA_Spoken_TriTop30K.*** This variable measures how often trigrams (three-word phrases) within a text can be found within a list of the 30,000 most frequently used trigrams contained inside of the Corpus of Contemporary American English—Spoken (COCA–Spoken) sub-corpus. Individuals process frequently used multi-word phrases more easily than uncommon multi-word phrases.[16] And the frequency with which one encounters words or phrases in *spoken* English is likely a better marker of one's ability to understand a word or phrase than the frequency with which one encounters words or phrases in written English—since people tend to have greater exposure to spoken forms of language than to written forms.[17]

The final MADRS variable measures an element of the grammatical complexity of phrases within a text.

6. ***Dependents / nominal (no pronouns, standard deviation) / Nominal_Deps_NN_StDev.*** The variable measures the number of

15. Harvard University General Inquirer, "Descriptions of Inquirer Categories and Use of Inquirer Dictionaries," online: <inquirer.sites.fas.harvard.edu/homecat.htm> [perma.cc/63B7-M8KG]. The "legal" word list originates from the Harvard IV-4 dictionary lists used by the General Inquirer, a 1966 database consisting of 11,000 words assigned to 117 different word lists (Harvard University General Inquirier, *ibid*). See also Scott A Crossley, Kristopher Kyle & Danielle S McNamara, "Sentiment Analysis and Social Cognition Engine (SEANCE): An Automatic Tool for Sentiment, Social Cognition, and Social Order Analysis" (2017) 49 Behavior Research Methods 803 at 806.

16. See generally Inbal Arnon & Neal Snider, "More Than Words: Frequency Effects for Multi-Word Phrases" (2010) 62 J Memory & Language 67.

17. Xiaobin Chen & Detmar Meurers, "Word Frequency and Readability: Predicting the Text-Level Readability with a Lexical-Level Attribute" (2018) 41:3 J Research in Reading 486 at 490.

dependent clauses associated with each noun-phrase in a text,[18] and computes the standard deviation for this number. A higher score for the variable *Nominal_Deps_NN_StDev* would likely be reflective of a text that has relatively more sentences with multiple dependents per nominal. In other words, a higher score would generally mean that a text uses longer sentences with deeper syntactical structures.

The final MADRS variable measures an element of the grammatical complexity of phrases within a text.

MADRS = 10.036
+ (0.471 * ***AoA_CW***)
- (0.023 * ***MRC_Imag_AW***)
+ (4.513 * ***GI_Leg_Nouns***)
- (0.650 * ***COCA_Spoken_TriTop30K***)
+ (1.802 * ***Nominal_Deps_NN_StDev***)
- (0.018 * ***MRC_Concr_FW***)

MADRS values that this equation produces fall somewhere on a scale from 1 to 7, based on the scoring criteria contained below in Table 1.

Table 1: MADRS Scoring Scale

This text could be easily understood by an adult who is fluent in English, who has some motivation to read the text, and:	
who has both a law degree and familiarity with the legal subject of the text.	7
who has a law degree.	6
who has an undergraduate degree, or higher level of non-legal education.	5
who has some post-secondary education.	4
who has completed secondary school.	3
who has some secondary school education.	2

18. Seymour Chatman, "Pre-adjectivals in the English Nominal Phrase" (1960) 35:2 American Speech 83 at 83–84.

who has no secondary school education, or who is less than fluent in English	1

As this scale suggests, texts that score between 2 and 7 would not be easily understandable to readers who lack fluency in English—regardless of how educated these readers might be. However, texts that score a 1 on the MADRS scale could be easily understood by readers who are not necessarily fluent in English, and by people who are fluent in English, but who have not completed any secondary schooling. Much more information about the MADRS formula, the way it was created and validated, and its demonstrated use in assessing (and potentially explaining) readability score variances between justices of the Supreme Court of Canada can be found in the study that initially reports on the readability formula.[19]

II. Motivation: Why Get Personal?

My research into court and tribunal decision readability levels has led me to conclude that there is not a single, universal readability target that all courts and tribunals should strive to achieve. Decision readability levels should not be driven blindly or arbitrarily by ever further efforts to reduce the linguistic complexity of these decisions. Rather, readability levels should be driven by considerations about audiences. Specifically, authors should ask themselves who is likely to have a want or need to read this decision? Once this question is answered, and the relevant audiences are identified, then authors should write in ways that maximize the chances of these audiences understanding the decisions.

However, the general public is very frequently going to be part of the intended audience for adjudicative decisions,[20] and decisions should probably

19. Madden, "How Understandable are Adjudicative Decisions?", *supra* note 1. Additionally, MADRS results can now be easily computed (alongside other well-known readability formulas) for single or multiple text files using the online calculator at www.readability.tools.com.

20. See e.g. Canada, Department of Justice, *The Honourable Sheilah Martin's Questionnaire (Questionnaire for the Supreme Court of Canada Judicial Appointment Process)* (Ottawa: DOJ, 21 December 2017) at Part 10, question 4, online: <fja.gc.ca> [perma.cc/P75P-MPWH]. ("[T]he public has a legitimate interest in knowing why a particular conclusion was reached and why a certain principle or position prevailed. The decisions should be clear and convincing" (*ibid* at Part 10, question 4). See also, Canada, Department of Justice, *The Honourable Jill Miriam Copeland's Questionnaire (Questionnaire for the Court of Appeal for Ontario's Judicial Appointment Process)* (Ottawa: DOJ, 30 May 2022), at Part II, question 4, online: <canada.ca> [web.archive.org/web/20260228020633/https://www.canada.ca/en/department-justice/news/2022/05/the-honourable-jill-miriam-copelands-questionnaire.html]. ("The public is also an important

be written to err on the side of inclusive, rather than exclusive readability levels. In other words, in cases of doubt about whether a decision is potentially of interest to a larger audience, it should be written so that the larger audience is likely to understand the decision—because this approach will help the author to achieve their (likely) intention of effectively communicating the decision's contents to the audience members.

As the above discussion suggests, there are at least some instances (e.g. narrow decisions on topics that involve settled law, where the parties are sophisticated and/or represented by counsel, and where there are no other public interest factors that might attract a wider audience) when the lowest common denominator of audience member reading competencies may still be quite high. In those circumstances (which become increasingly rare as one moves up the adjudicative hierarchy from tribunals to inferior, then superior, then appeal, and then apex courts), the authors of decisions may not need to take great pains to communicate their decisions in widely readable language.

I stress this point here to affirm that I did not begin the process of studying Canadian adjudicative decision readability levels with an *a priori* motive of creating social or political pressure to reduce the linguistic complexity levels (or to improve the readability levels) of decisions in all cases. It was impossible for me to know in advance whether decisions were already being written at levels that made them understandable to likely audiences, so such a motivation would have been premature, at best. As the results from other readability studies[21] and those within this article now show, we have strong reasons to believe that many—if not most—adjudicative decisions are being written in ways that

audience for Court of Appeal decisions. Judicial independence structures and limits the way that judges are accountable to the public. Written decisions (or oral decisions with adequate reasons) are one of the most important ways that judges are accountable to the public for how they carry out their judicial role" at *ibid* (Part 11, question 4).

21. See e.g. Brian M DeFriez, *Toward a Clearer Democracy: The Readability of Idaho Supreme Court Opinions as a Measure of the Court's Democratic Legitimacy* (Ph.D. Dissertation, University of Idaho, 2017) [unpublished]. The authors undertook a qualitative and quantitative study of Idaho court decisions, and noted among other things that the decisions were scoring higher on both the Flesch Reading Ease and Flesch-Kincaid Grade Level indices over time. Decisions published since 2010 required, on average, a grade 14.7 reading level (*ibid* at 100–02). See also Stephen M Johnson, "The Changing Discourse of the Supreme Court" (2014) 12:1 U New Hampshire L Rev 29. The authors measured both the length and readability levels Supreme Court of the United States (SCOTUS) opinions during the periods between 1931-1933 and 2009-2011, and found that the more recent decisions were longer and less readable (*ibid* at 57). Decisions during the earlier period required, on average, a grade 12.2 reading level, but a grade 13.3 reading level for the later period (*ibid* at 58, citing Stephen M Johnson, *1931-1933 Opinions by Opinion Type*, online (pdf): <http://www2.law.mercer.edu/elaw/readability/1931-1933 Opinions By Opinioin Type.pdf>; Stephen M Johnson, *2009-2011 Opinions by Opinion Type*, online (pdf): <www2.law.mercer.edu/elaw/readability/2009-2011

preclude large segments of their potential audiences from understanding the decisions. But this *a posteriori* conclusion was not one that I set out to prove.

My motive for undertaking this research project was rather more modest: I wanted to encourage judges and tribunal members to be more reflective and self-critical, by asking themselves whether they are succeeding in reaching their audiences. I wanted to give these authors a tool, and some example data, to show them how they might incorporate readability considerations into their decision-drafting workflows for the future. If I could achieve this ambition, then this article would do much more than simply point to a single snapshot of how readable decisions are in a particular year. It might instead nudge judges and tribunal members to think and write from a more audience-centric perspective, while simultaneously equipping them with a formula that they could use to verify their work—to see if the language of their decisions matches up with the expected reading competency levels of their audiences.

This author-focused strategy strives to communicate the key content of this article to what I feel would be the most appropriate group of people: the judges and tribunal members who write decisions, so they may consider how readability levels might or might not align with the expected reading competency levels of their intended audience members. As this statement implies, I see the MADRS formula as being perhaps most useful to authors, rather than to administrators or scholars or executive-level members of government who are responsible for appointing judges and tribunal members. The judge or tribunal member who presides over a particular proceeding—who hears all the evidence and/or reads all the material within the record—is probably in the best practical position to determine who will fall within the audience for the decision, and what the likely reading abilities are of those audience members.[22] These authors are therefore the most important people to reach with information about the MADRS formula, and how it can be used in different ways.

Opinions By Opinion Type.pdf>). See also Ryan Whalen, "Judicial Gobbledygook: The Readability of Supreme Court Writing" (2015) 125 Yale LJ Forum 200. See also Mike Madden, "Failure to Adapt: Readability Deficits in Canadian Court Decisions Involving Parties with Unique Reading Needs" (forthcoming 2026) 71 McGill LJ, online: <ssrn.com/abstract=5172897>. See also Mike Madden, "Message Undeliverable: An Empirical Assessment of the (Un)Readability of Canadian Court Decisions" (10 March 2025), online: <ssrn.com/abstract=5172476> [Madden, "Message Undeliverable"].

22. Of course, there are many others who might have better training and knowledge for the purposes of assessing cognitive and reading competency levels than judges (e.g. cognitive or educational psychologists). But these people do not have any standing responsibility for, or involvement in, the administration of justice. Therefore, when I say that judges are in the best practical position to decide what reading competencies are likely to be possessed by the audience members for their court decisions, I mean that they are in better positions than anyone else who has standing responsibilities for, or involvement in, the administration of justice.

However, creating the MADRS tool and pointing authors to largely anonymized data about how some of the many thousands of judges and tribunal members in Canada write their decisions is likely not enough to generate the readability nudge that I am striving for within my research. It is well-established in psychological literature that individuals tend to overestimate their own performance relative to their peers.[23] This phenomenon seems to apply (as predicted) to legal authors who self-assess their writing:[24] we tend to overestimate the quality of our own legal writing. It would therefore be easy—even expected—for any judge or tribunal member who encounters some of the more negative results within a decision readability study to tell themselves that the results probably reflect their peers' decisions, but not their own decisions.

It is partially for this reason that this article now gets personal. Below, I describe and provide access to readability results from a study of 1,621 Canadian court and tribunal decisions that were published in 2022. The author of the decision (or the fact that multiple authors co-wrote a decision) is identified for each decision within the Table of Results for this study. A total of 840 different authors wrote decisions whose results are included in this Table of Results. My hope is that, by providing the legal, judicial, and tribunal member communities with access to these personalized results, authors will find personal relevance to the contents of this article—and motivation to explore the results.

The current study also classifies each decision based on eight other variables: (i) court level, (ii) jurisdiction, (iii) legal subject, (iv) author's gender, (v) opinion type, (vi) number of judges agreeing with an opinion, (vii) percentage of judges agreeing with an opinion, and (viii) word count. The inclusion of these variables within the Table of Results allows others who read the results to manipulate the data in many different ways to shed light on questions that might be of interest to each of these different readers.

Ultimately, the study that is reported on within this article is really about providing outside users—ideally including judges and tribunal members who author decisions—with data that they can peruse and explore in their own self-directed ways.

---

23. Ethan Zell et al, "The Better-Than-Average Effect in Comparative Self-Evaluation: A Comprehensive Review and Meta-Analysis" (2020) 146:2 Psychological Bull 118.

24. Elizabeth Ruiz Frost, "Mentoring Legal Writers: Overcoming 'Illusory Superiority'" (2013) 73:9 Oregon State Bar Bull 13. This article discusses the quality of legal writing—and how to improve it—as a general matter and does not specifically refer to the readability of legal writing. However, the author makes the following general point about legal writing that could be said in equal measure about the *readability* of legal writing, specifically: "most of these new lawyers probably don't even know that they're not good writers. People aren't very good at knowing what they don't know" (*ibid* at 13).

III. Method: Getting MADRS Results for a Large Sample of Decisions

My research method for this study was similar in many respects to the methods used in similar readability studies. I first created a massive corpus of decisions that was designed to include files that differed from one another on each of the different variables that were being recorded. Then I proceeded to compute MADRS values for each decision within the corpus.

I initially contemplated using text-based data that others had previously compiled as the basis for my study. For instance, York University's Refugee Law Lab has already made a bulk, publicly available dataset consisting of approximately 185,000 decisions from the Supreme Court of Canada, the Federal Court of Appeal, the Federal Court, the Tax Court of Canada, and four federal tribunals, freely available.[25] Similarly, the Access to Algorithmic Justice project has created a massive database of Canadian case law from the above courts, tribunals, and the Court of Appeal for Ontario.[26]

However, my preliminary review of the data that I could access from these databases revealed that the text outputs included far more than just the raw opinion(s) from the authoring judge(s). The files also included significant front and end-matter (to varying degrees) such as the names of counsel who appeared, any headnotes, lists of cases cited within the decision, etc. In certain circumstances, decisions were printed in both English and French within the same entry. Additionally, text files from both data sets also included character-based formatting sequences (such as "\n" to signal line breaks, and "\u2019" to signal an apostrophe). I was unable to easily and computationally (i.e. other than manually) resolve these formatting issues with the bulk datasets. Since each issue had the potential to contribute to inaccurate MADRS measurements for the relevant decisions (by including text or characters that a reader would not encounter in any other typical print or online versions of the decision), I elected to create my own databases for this study as described below.

My goal for this study was *not* to build a representative corpus of Canadian adjudicative decisions by randomly sampling from the full population of decisions. Statistically, such a corpus would represent the full spectrum of different decision types from different places on different subjects, etc., in a proportional way. Thus, if the corpus consisted of decisions from 2022, we might expect it to include primarily decisions from Ontario, British Columbia, Alberta, and the Federal jurisdictions (since these jurisdictions produce far more decisions each year than other jurisdictions), and we might not expect to see any decisions from Nunavut or Prince Edward Island (since these jurisdictions

25. Refugee Law Lab, "Bulk Legal Datasets" online: <refugeelab.ca> [perma.cc/6E9F-93BG].

26. *Ibid*; Sean Rehaag & Simon Wallace, "A2AJ Canadian Legal Data" (2025), online: <a2aj.ca/canadian-legal-data/> [perma.cc/9GBJ-F65L]

produce approximately only 0.2% as many decisions as Ontario produces—that is, they produce only 2 decisions for every 1000 decisions from Ontario).[27]

Because a corpus like this would be representative (due to the random sampling technique that is used to create the corpus), it would allow researchers to draw certain high-level statistical inferences about the results. But the corpus would not necessarily include any cases involving different variables that occur within the whole population of cases at a very low frequency (like cases from Nunavut, or cases about environmental law), so the representative corpus would be limiting in terms of the different factors or variables that it would allow us to consider in relation to readability results.

Instead, my goal for this study was to build a corpus that aimed for a diversity of cases based on different variables, rather than on proportional representativeness. This approach means that we cannot statistically generalize about the full population of Canadian cases based on the results from within this non-representative corpus. However, the corpus likely contains a much wider breadth of different cases than it would otherwise. For instance, the corpus allows us to compare readability differences between decisions from Nunavut and British Columbia *within the corpus*, even if we cannot statistically extrapolate to extend our conclusion to all decisions from those jurisdictions.

A. File Selection

To ensure that the corpus contained a wide range of different decision types based on the different variables I planned to include within the study, I conducted a series of targeted searches using the CanLII databases. A brief description of the different search queries is included below, and I have kept a complete record of the search terms and history on file.

First, I included all opinions from the Supreme Court of Canada from 2022. I also included a random sample of twenty-five decisions from each provincial appeal court, and from the Federal Court of Appeal. In any cases where a provincial or territorial appeal court had released less than twenty-five decisions in 2022, I took all decisions from that court to meet the inclusion criteria. These searches ensured that a minimum number of appellate and apex court decisions would be included within the corpus.

Next, I included a random sample of twenty-five decisions from each provincial superior court, from the Federal Court, and from Courts Martial. In any cases where one of these courts had released less than twenty-five decisions in 2022, I took all decisions from that court to meet the inclusion criteria. I also included a random sample of twenty-five decisions from each province or

27. These statistics were calculated by searching for the total number of court and tribunal decisions within the CanLII database from 2022 and then comparing how many emerged from each jurisdiction.

territory's inferior court(s). In any cases where provincial or territorial inferior court(s) had released less than twenty-five decisions in 2022, I took all decisions from that level of court to meet the inclusion criteria. These searches ensured that a minimum number of trial court decisions would be included in the corpus.

I included a random sample of twenty-five decisions from the following administrative tribunals: British Columbia Human Rights Tribunal, British Columbia Civil Resolution Tribunal, Safe Roads Alberta, Saskatchewan Information and Privacy Commissioner, Manitoba Labour Board, Ontario Consent and Capacity Board, Ontario Condominium Authority Tribunal, New Brunswick Workers Compensation Appeals Tribunal, Nova Scotia Utility and Review Board, Canadian International Trade Tribunal, Social Security Tribunal of Canada, and, Immigration and Refugee Board of Canada. These searches ensured that at least some tribunal decisions from each geographic jurisdiction in Canada would be included in the corpus, from across a range of different legal subject areas.

I took an additional 200 court decisions and 200 tribunal decisions based on the following targeted searches:

- Courts:
    - search full-text for the term "court"; sort by newest, take the first 100 results; and
    - search full-text for the term "court"; sort by oldest, take the first 100 results;
- Tribunals:
    - search full-text for the terms "law and decision and evidence and fact and proof"; sort by relevance, take the first 50 results;
    - search full-text for the terms "law and credibility and standard"—sort by relevance, take results numbered 51–100;
    - search full-text for the terms "law and review and decide"; sort by newest, take the first 50; and
    - search full-text for the terms "law and review and decide"; sort by oldest, take the first 50 results.

Decisions from Quebec courts were excluded as these decisions were overwhelmingly written in French, and my readability formula is based on the English language. The variables that contribute to the formula all rely on measurements about uniquely English words, phrases, or grammatical structures. The *COCA_Spoken_TriTop30K* variable, for instance, compares the trigrams in a text against a list of the top 30,000 trigrams that appear in a corpus of spoken English texts. There is no identical corpus for the French language. Nor is there any obvious reason why three-word phrases that appear

frequently in spoken English would also appear with the same frequency (in directly translated form, if such a translated form were possible to create) in spoken French, and would have the same effect on comprehension in French.

The same kind of language-specific problems exist for other MADRS variables. This is perhaps why scholars have concluded that "[u]nfortunately, existing approaches cannot be easily extended to handle thousands of different languages".[28] Formulas from one language cannot necessarily predict readability levels of texts from another language. MADRS, because of the language-specific variables that the formula relies upon, cannot be assumed to operate effectively on French texts—so the ensuing studies within this article have not applied the formula to French texts.

For a similar reason, I made best efforts to exclude any cases where it was evident that the decision had originally been written in French but was also translated and published in English. Exclusions on this basis often arose for Federal Court decisions (where the text "[ENGLISH TRANSLATION]" would appear at the top of certain decisions that had originally been written in French). Because translated decisions might not express the authoring judge's intentions as accurately as decisions that were written first in English, I excluded these decisions from my corpora where they were identified.

Additionally, any opinion with less than 400 words was excluded. The MADRS formula was developed using texts of approximately 600 words,[29] so opinions with less than 400 words would not necessarily provide enough linguistic information for the MADRS formula to accurately score the text.

B. Text Extraction, File Preparation, and Variable Coding

Once a case was identified for inclusion within the corpus, I navigated to the web-based publication of the decision and manually copied from the first word to the last word of the body of the decision (or opinion, as appropriate). As this method choice suggests, all header and preliminary information that preceded the body of a decision (such as headnotes, counsel names, hearing location and date, docket number, etc.) and all end-matter that followed the last word of the body of a decision (such as footnotes, counsel names, court orders that did not form part of the body of the decision, etc.) were omitted from my selections of text. The selected text was then pasted into a new plain text file and saved within a file directory specific for this corpus. Each text file contains a single judicial opinion, not necessarily the entire decision. Where a

28. Jesse Saba Kirchner, Justin Nuger & Yi Zhang, "An Extensible Crosslinguistic Readability Framework" in Pascale Fung, Pierre Zweigenbaum & Reinhard Rapp, eds, *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora* (Singapore: ACL Anthology, 2009) 11 at 11.

29. Madden, "How Understandable are Adjudicative Decisions?", *supra* note 1.

selected case included multiple opinions, each opinion was saved and processed separately. This is why the corpus contains a non-rounded number (1,621) of files.

For each file, I manually reviewed the decision and opinion to input metadata for the file into a separate spreadsheet. Specifically, I coded each decision for the following variables:

- court level (apex, appeal, superior, inferior, or tribunal);
- jurisdiction (that is, the specific province, territory, or federal jurisdiction);
- legal subject (based on my own assessment of the dominant subject of the decision);
- judge name;
- judge gender (determined in all but seven cases by searching publicly available information like court or tribunal websites, or announcements of appointment, to ascertain how the different judges or tribunal members identify themselves);
- opinion type (that is, a "solo" opinion in cases where a decision-maker presided alone, or a "unanimous", "majority", "concurring", or "dissenting" opinion where a panel of decision-makers was involved);
- number of judges agreeing (a raw count of how many decision-makers signed on to an opinion); and
- percentage of judges agreeing (calculated by dividing the number of decision-makers agreeing with an opinion by the total number of decision-makers on the panel).

C. Linguistic Processing of Files / MADRS Calculations

I processed all text files containing court decisions and opinions through three different linguistic software applications to measure the relevant variables within the files that contribute to a text's MADRS value.

Specifically, I processed the files through the Tool for the Automatic Analysis of Lexical Sophistication (TAALES) program.[30] TAALES generates results for different lexical sophistication and complexity variables, including four of the variables that contribute to the MADRS formula: *MRC_Imag_AW*, *MRC_Conc_FW*, *AoA_CW*, and *COCA_Spoken_TriTop30K*. I then processed the files

30. Kristopher Kyle, Scott Crossley & Cynthia Berger, "The Tool for the Automatic Analysis of Lexical Sophistication (TAALES): Version 2.0" (2018) 50:3 Behavior Research Methods 1030, DOI: <doi.org/10.3758/s13428-017-0924-4> [perma.cc/RP5E-B7BR].

through the Sentiment Analysis and Cognition Engine (SEANCE) program.[31] SEANCE results contain data on different sentiment, cognition, and topic-related variables, including the variable *GI_Leg_Nouns*, which contributes to the MADRS formula. Finally, I processed the files through the Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC) program.[32] TAASSC results contain data on different syntax-related variables, including the variable *Nominal_Deps_NN_StDev*, which contributes to the MADRS formula.

Once values had been calculated by these three different linguistic programs for each of the six variables that contribute to the MADRS formula, for each file within a corpus being studied, it was possible to compute MADRS values for each file using the regression equation described in the previous section. Following these steps allowed me to compute MADRS values for all decisions and opinions across the corpora.[33]

IV. Results: Some Examples and a Resource for Further Study

MADRS results were calculated for 1,621 decision or opinion files, written by 840 individually identified authors (alongside many different groupings of collective authors), containing approximately 9.3 million words of text. Some of the other characteristics of this corpora of opinion and decision files are shown graphically, below.

31. Scott A Crossley, Kristopher Kyle & Danielle S McNamara, "Sentiment Analysis and Social Cognition Engine (SEANCE): An automatical tool for sentiment, social cognition, and social order analysis" (2016) 49:3 Behavior Research Methods 803, DOI: <doi.org/10.3758/s13428-016-0743-z> [perma.cc/2AJU-HB7L].

32. Kristopher Kyle, Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication (Doctoral Dissertation, Georgia State University, 9 May 2016) [unpublished], DOI: <doi.org/10.57709/8501051> [web.archive.org/web/20260315110023/https://scholarworks.gsu.edu/items/8046a5d0-2d60-44c8-9ab7-207f263ce8b5].

33. To confirm that no data handling errors occurred, two independent Microsoft Excel files were created to compute MADRS results: the linguistic results were separately transferred into each Excel file, the MADRS equation was entered separately into each file, and the results were compared between the two files. The results were identical.

Chart 1: Breakdown of Corpus Files by Court Level

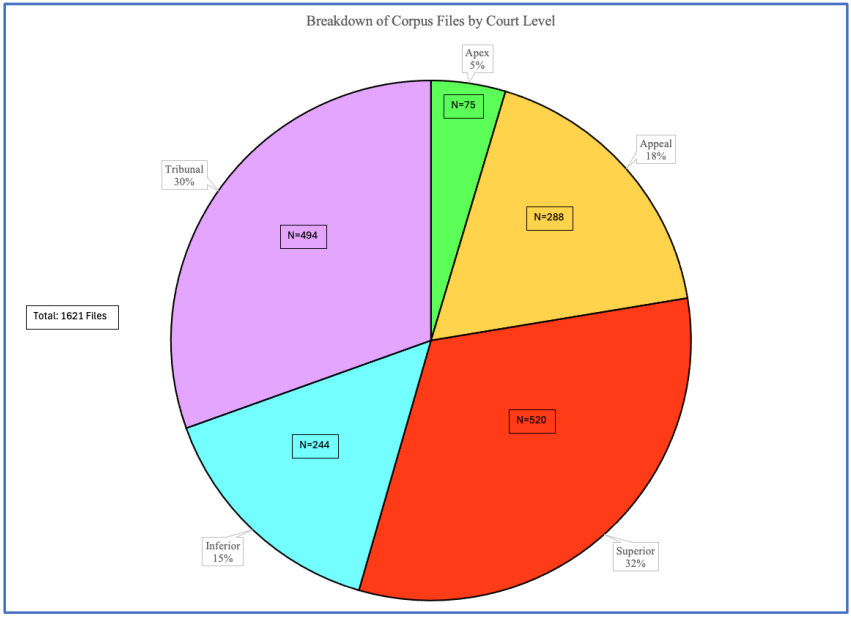
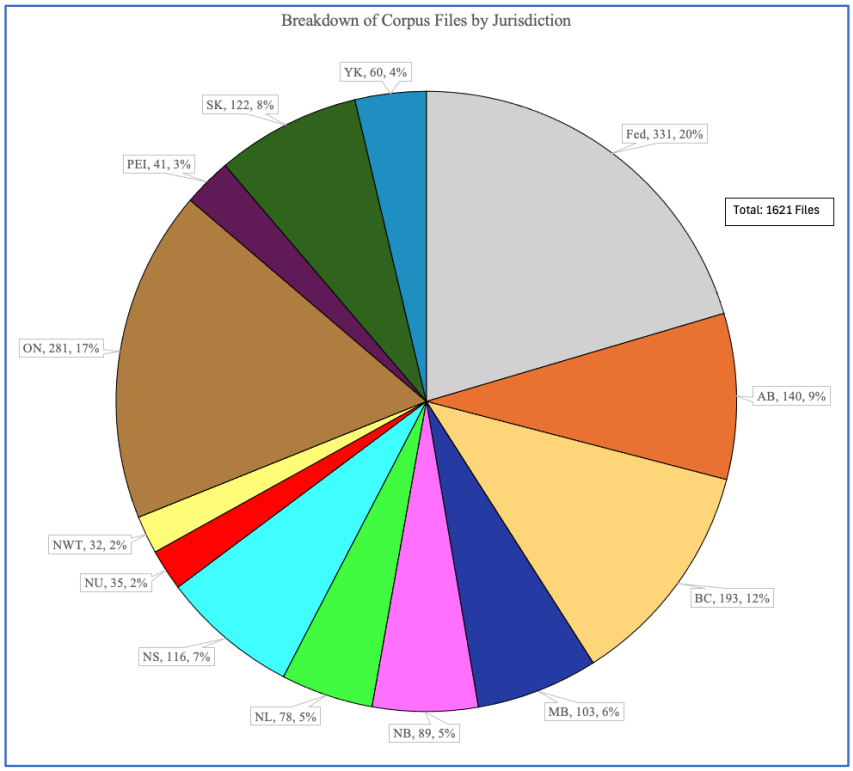


Chart 2: Breakdown of Corpus Files by Jurisdiction



A. Random Sampling and Statistical Inferences

Before moving on to a discussion of some of the more salient results from this study, I reiterate that the results from this non-representative study are not (necessarily) generalizable to the broader population of Canadian cases as a whole. So, conclusions drawn about the MADRS results from this sample of 1,621 cases are only certain to be valid in respect of these cases, but not necessarily to other Canadian cases.

I have stressed the non-representative nature of my data set out of an abundance of caution to discourage those who do not have strong backgrounds in statistical analysis and quantitative research methods from misinterpreting the results. That being said, my message of caution is admittedly an oversimplification.

As Wu has noted, quantitative studies that sample from a population—as opposed to studying the complete population—are always fallible because

of the likelihood that the sample does not accurately represent the full population.[34] This is potentially problematic because the foundation of any statistical inference is an assumption that the studied sample is representative of the population as a whole.[35] That assumption will almost never hold perfectly true. Even purportedly random survey samples of human subjects are really only "random contact [surveys]".[36] These are surveys in which people are randomly contacted despite their being non-random due to unknown and unknowable respondent biases in their decisisons regarding whether to participate.[37] In the case of non-random sampling from a population, other selection biases (that manifest in any decisions a researcher makes about which elements of the population to include and exclude from the sample) will also create a likelihood that the sample does not accurately represent the full population.[38]

However, population samples that are collected at random at least allow researchers to determine the probability of any individual element of the population being included in the sample.[39] This, in turn, allows researchers to estimate confidence levels and margins of error associated with the sample relative to the full population.[40] It is not statistically possible to compute confidence levels and margins of error in the same way when a sample has been collected non-randomly. This is perhaps why random sampling has historically represented a "gold standard" in the design practice of quantitative research methods.[41]

Nonetheless, the key issue when making statistical inferences is the assumption that a studied sample is representative of the broader population from which it is drawn, and it is possible to make this assumption even in cases where a sample has been non-randomly collected—although the assumption

34. Changbao Wu, "Statistical inference with non-probability survey samples" (2022) 48:2 Survey Methodology 283 at 284.

35. Statistics Canada, "Probability Sampling" in *Statistics: Power from Data!* (Government of Canada, 2021) at section 3.2.2, online: <www150.statcan.gc.ca/n1/edu/power-pouvoir/ch13/prob/5214899-eng.htm> [web.archive.org/web/20260116100023/www150.statcan.gc.ca/n1/edu/power-pouvoir/ch13/prob/5214899-eng.htm] [Statistics Canada, *Power from Data*].

36. See generally, Michael A Bailey, "Comments on 'Statistical inference with non-probability survey samples' − Non-probability samples: An assessment and way forward" (2022) 48:2 Survey Methodology 313 at 316 [emphasis added].

37. *Ibid*.

38. Statistics Canada, *Power from Data*, *supra* note 35 at section 3.2.3.

39. C Wu, *supra* note 34 at 284.

40. *Ibid*.

41. Loan R van Hoeven et al, "Aiming for a representative sample: Simulating random versus purposive strategies for hospital selection" (2015) 15:90 BMC Medical Research Methodology 2 at 2, DOI: <doi.org/10.1186/s12874-015-0089-8> [perma.cc/ZP3D-6ZEA].

may be riskier in such cases.[42] But this risk will fall somewhere along a sliding scale, often depending on the extent to which selection biases are likely to have an effect on the variable being studied within the sample.

To put this idea into more concrete and relevant terms, my case samples within the present study would likely have been at greater risk of not representing the full populations from which they were drawn if, for example, I had only selected decisions written by judges who attended a plain-language decision-writing workshop within the twelve-month period before decisions were written, or who were members of a (fictional) Canadian Society for the Preservation of Technocratic English. These kinds of selection biases could reasonably have been expected to produce results that differed in readability measurements (skewing them toward better readability in the former case, and worse readability in the latter case) when compared to decisions written by all judges or tribunal members.

In reality, however, the cases that I selected for inclusion within the present study were identified using quota sampling—which is a recognized approach for ensuring that enough cases from each subpopulation of interest are included within a larger sample.[43] Furthermore, within each subpopulation, cases were identified for inclusion using either random (for the twenty-five decisions that I took from each Superior and Appeal court from each federal, provincial, and territorial jurisdictions), or close-to-random (in cases where, for instance, I selected sequential tribunal decisions that used the words "law and review and decide" after the results had been sorted by relevance) approaches. I suggest that these latter approaches were close-to-random because I had no theoretical reason for thinking that the selection approaches in these situations would impact readability results.

As the above discussion suggests, there are no obvious reasons for suspecting that cases within the present study are not representative of the larger populations from which they have been drawn. A careful approach to considering the results that follow would simply drawing statistical inferences and generalizing from the results, because we do not know—and cannot calculate—a confidence level or margin of error for results from the samples relative to the full populations of cases. But if we accept that the samples are still likely to be largely representative of the populations from which they have been collected (because we have no

42. Statistics Canada, *Power from Data*, *supra* note 35 at section 3.2.3. "[I]n order to make inferences about the population, it requires strong assumptions about the similarity between the sample and the population even though the respondents are self-selected. Due to the selection bias presented in all non-probability samples, these are often dangerous assumptions to make" (*ibid* at section 3.2.3).

43. *Ibid* at section 3.2.3. "Sampling is done until a specific number of units (quotas) for various subpopulations have been selected. Quota sampling is a means for satisfying sample size objectives for the subpopulations" (*ibid* at section 3.2.3). In the present study, quotas were established for each of the categories described above (under the heading "File Selection").

reasons for thinking that my selection methods would lead to any positively or negatively skewed readability results), then it may still be appropriate to draw inferences and make generalizations from the data. There is no bright-line rule of statistical practice that precludes drawing such inferences. And the approach of using non-random data collection techniques even for the purposes of inferential statistical analyses is gaining popularity.[44] Wu notes that quota surveys (employing the same kind of quota-based approach that the present study uses), are particulary "widely used and can be successful under certain conditions".[45]

I have elected to describe the results that follow within this article in a more cautious way (that is, as if these results cannot be used to draw inferences and generalize about the larger populations from which they emerge), to reduce the possibility of these results being misunderstood. Nonetheless, the analyses that follow may illuminate research questions or paths that would be fruitful for further inquiry in the future. This discussion may essentially suggest that, if a particular phenomenon was observed in this study of a broad (but not representative) sample of decisions, then perhaps it would be worthwhile to explore in other targeted studies whether the same phenomenon can be said to exist in the full population of Canadian cases.

A spreadsheet file in Microsoft Excel's "xlsx" file format containing all of the results and associated metadata from the present chapter's study can be found within the Harvard Dataverse at the following link: https://doi.org/10.7910/DVN/NIQ2KJ. Judges, tribunal members, researchers, and any other interested individuals are encouraged to download and view this data for their own purposes.

B. Putting MADRS Values into Context

Before presenting some of the readability results from the present study, it might be helpful to put MADRS scores into a more familiar context. I therefore calculated MADRS values for several other judicial texts that might be known to most common law lawyers. These texts, and their corresponding MADRS results, are discussed briefly below.

One of these texts is drawn from a famous English case written by Lord Denning, about the delights of cricket in the village, and the nuisance it creates for some homeowners.[46] From a literary and legal writing perspective, this is perhaps one of the most memorable cases that law students will study. I calculated MADRS values for the opening paragraph of Lord Denning's

44. C Wu, *supra* note 34 at 283.

45. *Ibid* at 283–84.

46. *Miller v Jackson*, [1977] EWCA Civ 6 (UK).

opinion in this decision[47] (in isolation), and for Lord Denning's entire opinion. The first paragraph scores a 2.7, while the full text of Lord Denning's opinion scores a 3.6. So, while a person with incomplete high school education could easily understand the opening paragraph, this person would need more than a completed secondary school diploma to understand the opinion as a whole. The difference in these results is not surprising: as the opinion moves from general matters of everyday life (at least, for villagers in England in 1977) to more technical legal matters of nuisance, pleadings, easements, prescriptions, and discretionary remedies, one might expect the reading difficulty of the text to increase.

Two other famous tort cases that most law students would likely have studied include the cases of *Palsgraf v Long Island Railroad Co* (involving a woman who was injured when a man dropped a package that exploded while she was boarding a train),[48] and *Donoghue v Stevenson* (involving a woman who felt ill after drinking ginger beer from a bottle that contained a decomposed snail).[49] I computed MADRS values for Cardozo J's majority opinion in the former case, and for the complete decision (consisting of five separate opinions) from the latter case. *Palsgraf* scored 4.7, and *Donoghue* scored 5.0. As these scores suggest, the cases may be memorable for many reasons, but this does not necessarily mean that they are highly readable, since something close to a completed undergraduate degree would be needed for a reader to understand both of these decisions.

Finally, I have examined a Canadian example of a judicial author—Watt JA, formerly of the Court of Appeal for Ontario, now retired—who "is seen by some as part of a welcome trend away from impenetrable, legal jargon".[50] I calculated MADRS values for one of Watt JA's notable decisions,[51] and have

47. See *ibid*. The first few sentences of the opening paragraph are reproduced here:

> In summertime village cricket is the delight of everyone. Nearly every village has its own cricket field where the young men play and the old men watch. In the village of Lintz in County Durham they have their own ground, where they have played these last seventy years. They tend it well. The wicket area is well rolled and mown. The outfield is kept short. It has a good club-house for the players and seats for the onlookers.

Ibid at para 1.

48. 298 NY 339 (App Ct 1928).

49. [1932] UKHL 100 (BAILII), online: <bailii.org> [perma.cc/JPD8-VZRL].

50. Kirk Makin, "The judge who writes like a paperback novelist", *The Globe and Mail* (10 March 2011), online: <theglobeandmail.com> [perma.cc/HHN7-NX64].

51. *R v Shafia*, 2016 ONCA 812 (CanLII).

again produced scores for the opening section of the decision in isolation,[52] and for the decision as a whole. The opening section scored 2.5, while the entire decision scored 5.1. Again, the scores are not surprising when one reads the decision. The opening section is written in a distinctly different (more conversational, more informal) style than the remainder of the decision. And the vocabulary in the opening section (discussing boats, a car, death, family members, etc.) is simple and concrete in comparison with the vocabulary in other parts of the decision (particularly from paragraph 167 onward, wherein Watt JA discusses things like the four threshold requirement for expert evidence, the balance of probative value and prejudicial effect, and *ante-mortem* statements of the deceased that were not tendered under the state of mind exception to the hearsay rule).

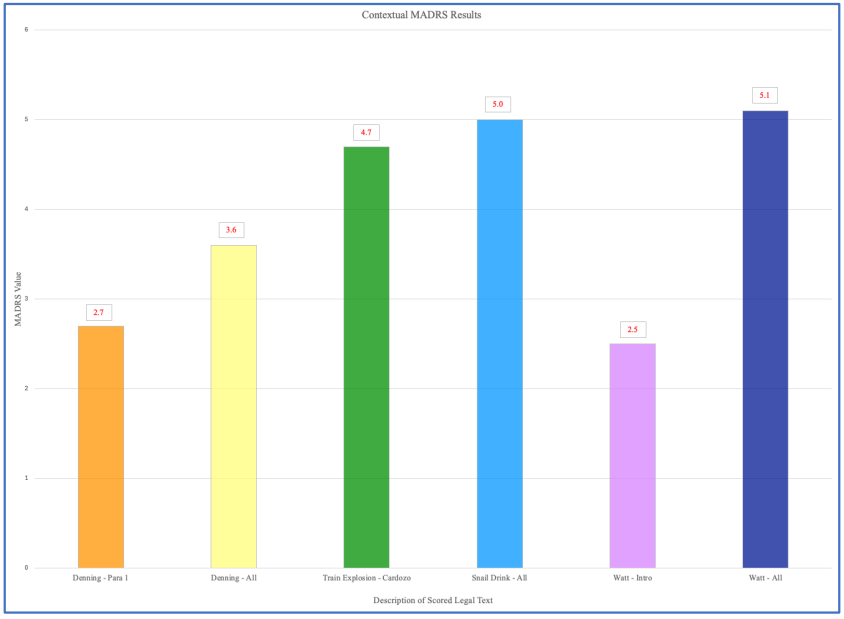
The MADRS results for each of the above contextual documents are shown below, in Chart 3.

52. See *ibid*. The first four paragraphs are reproduced here to demonstrate how Watt JA wrote the opening section:

> Boaters who travel the Rideau Canal system between Colonel By Lake and Lake Ontario pass through a series of locks at Kingston Mills. Lock gates open. Boats enter. Lock gates close. Boats leave.
>
> But not always.
>
> One morning – June 30, 2009 – boaters could not enter Kingston Mills Locks. Oil on the water of the upper lock. A closer look revealed its source. Something in the water.
>
> In the water was a Nissan Sentra. Driver's window open. Ignition off but not locked. Headlights off. Seatbelts unfastened. Front seats reclined. Rear name plate damaged.

Ibid at paras 1–4.

Chart 3: Contextual MADRS Results

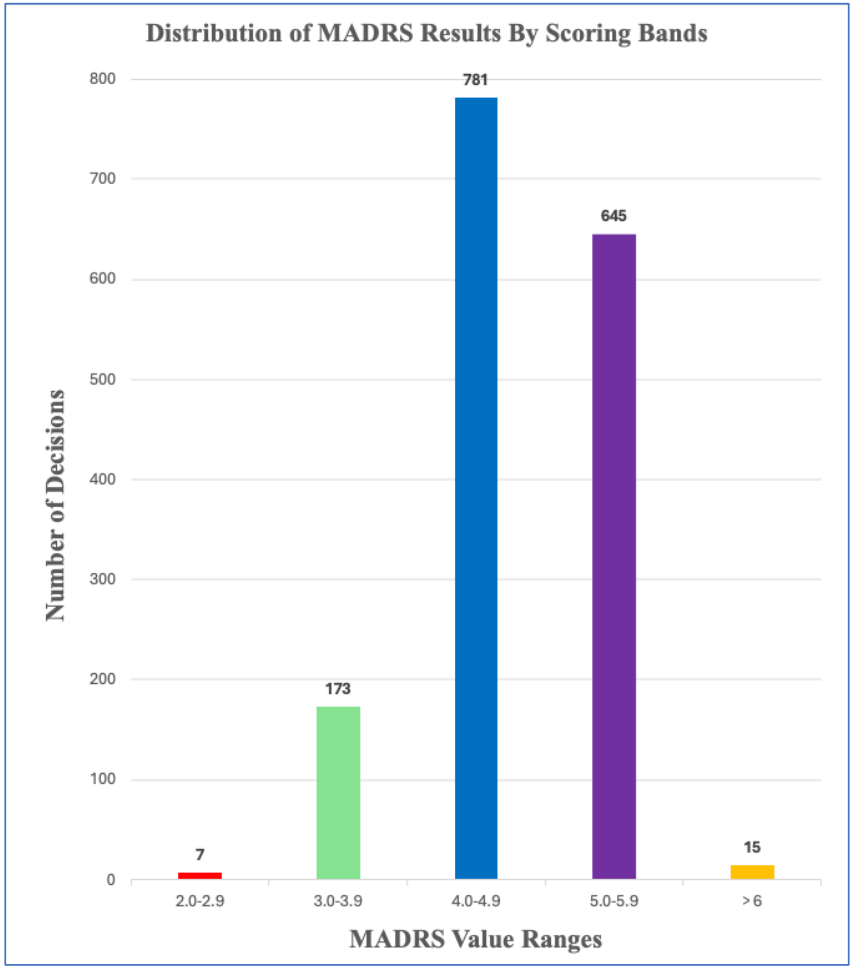


These examples should illustrate through familiar texts how the MADRS formula applies to the writings of different judges, whose readability scores cover a spectrum of over 2.5 points across the middle of the 7-point MADRS scale.

C. Overall Average and Scoring Distribution

The average MADRS value for all files within this dataset was 4.8 (very close to a level that would require a completed undergraduate degree from a reader for them to understand the average decision). Scores ranged from 2.2 to 6.6. A graph illustrating the distribution of MADRS values for this corpus, including how many decisions fell within each scoring range, is shown below in Chart 4.

Chart 4: Distribution of MADRS Results by Scoring Bands



The interquartile[53] MADRS range for these decisions (that is, the scoring range for the middle 50% of the decisions) was between 4.41 and 5.25. With respect to the seven most readable decisions, four were from inferior courts, and three were from superior courts. Six of these most readable decisions involved

53. Michael O Finkelstein & Bruce Levin, *Statistics for Lawyers*, 3rd ed (New York: Springer, 2015) at 25. "Interquartile range" refers to the range between the twenty-fifth and seventy-fifth percentiles; it offers a view of the middle fifty per cent of the data points, and is less sensitive to outlier points than other dispersion measures (*ibid* at 25).

criminal cases, and one was a family law decision. With respect to the fifteen least readable decisions, none were from inferior courts, but multiple decisions were from every other level (tribunals: two; superior courts: four; appeal courts: six; and, Supreme Court of Canada: three). They covered many subjects, but eight of the least readable decisions were about criminal law.

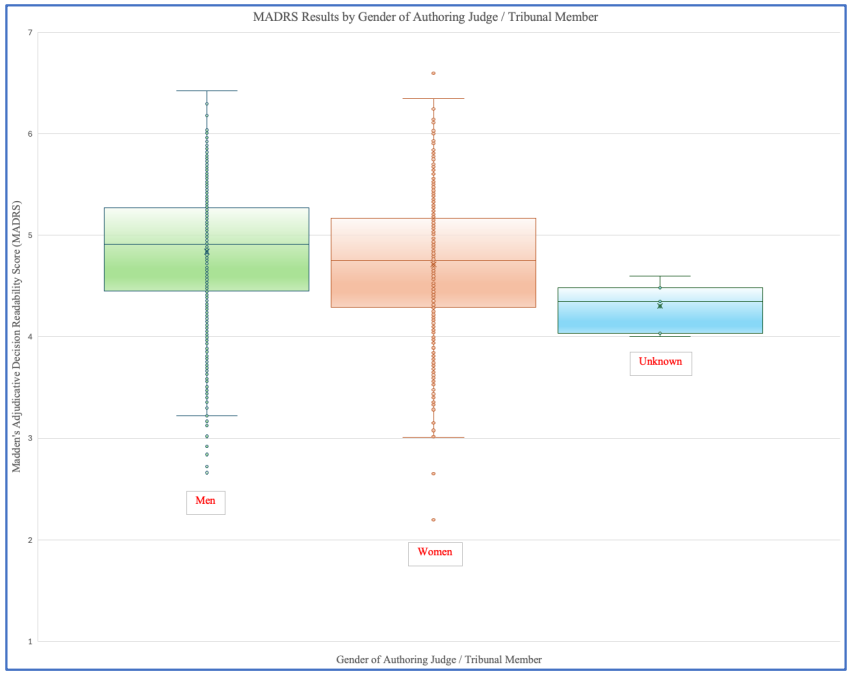
D. Gender and Individual / Collaborative Authorship

Decisions written by men (N=805) were slightly less readable than decisions written by women (N=698): the average for men was 4.84, and the average for women was 4.71. Only seven decisions were written by authors whose gender I was unable to identify, and these decisions had an average MADRS value of 4.30; all these authors were tribunal members. None of the 833 authors whose gender information was available to me identified as anything other than a man or a woman. The 2.5% readability difference between men and women authors is small: gender does not appear to be a major factor in determining how readable an adjudicative decision will be. Since I have elected to not draw statistical inferences about the larger population of cases from the samples that are included in the present study, the above discussion includes only descriptive statistics.[54]

A visual representation of MADRS results by gender of the authoring judge is shown below, in the boxplots contained in Chart 5. For each group, the "box" represents the interquartile range for MADRS values in the group. The middle horizontal line represents the median, or middle, score (the score that has an equal number of higher and lower scores within the sample). The "x" represents the average score. The "whiskers" extending above and below each box extend to show the full range of the scores in the group, or to 1.5 times the size of the interquartile range—whichever is lesser. Individual data points extending above or below the whiskers represent outlier values that are noteworthy for their distance away from the average for that group. Similar boxplots are shown for elsewhere in this article, and should be read in the same manner as the current boxplot.

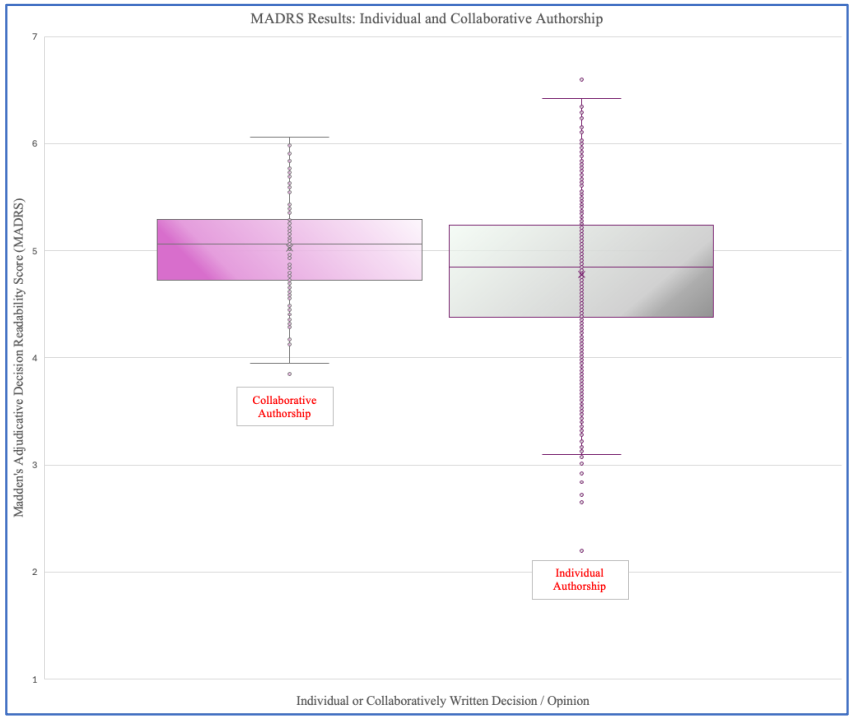
54. For this reason, no tests of statistical significance are included here. The averages represent the true averages for the cases within the study, and no claims are made about what the averages might say about other cases that do not form part of the current study.

Chart 5: MADRS Results by Gender of Authoring Judge or Tribunal Member



There were also 117 collaboratively written decisions and opinions (that is, decisions written by two or more identified authors, "the Court", or some other collective group of decision-makers) within the corpora, and the average MADRS value for these files was 5.03. The 1,504 individually authored decisions had an average MADRS value of 4.78. The full spectrum of results for these two types of decisions/opinions is shown in the boxplots within Chart 6.

Chart 6: MADRS Results: Individual and Collaborative Authorship



As this chart shows, only 2% of collaboratively written texts (2 of 117) scored below 4 on the MADRS scale. In contrast, approximately 12% of individually written texts (178 of 1504) scored below 4 on this scale. Furthermore, 54% of collaboratively written texts (63 of 117) scored higher than 5, while only 40% of individually written texts (597 of 1504) exceeded this score. In other words, individually authored decisions and opinions within the present study are generally more readable than collaboratively authored opinions—based on several different frames of observation. Again, the above discussion includes only descriptive, not inferential, statistical information.[55]

This comparison of results for collectively and individually authored decisions highlights an interesting and unanswered question about causation. Courts of appeal and the Supreme Court of Canada tend to decide cases as panels—thereby creating the possibility of collective authorship. Judges of

55. For this reason, no tests of statistical significance are included here. The averages represent the true averages for the cases within the study, and no claims are made about what the averages might say about other cases that do not form part of the current study.

superior courts (with the exception of Ontario's Divisional Court) and inferior courts, in contrast, only ever decide cases alone, so collective authorship is not possible for these judges. We also know from a related readability study that MADRS results are lower for inferior and superior courts than they are for appeal courts and the Supreme Court of Canada.[56]

This data therefore raises the following causation question: are average MADRS values higher for collectively authored decisions because these decisions are only possible from appeal and apex courts that generate higher MADRS results for other reasons? Or are appeal and apex courts' average MADRS values higher than inferior and superior courts' values because the former courts tend to have more collectively authored decisions, and collectively authored decisions cause higher MADRS results? Or, perhaps the court level variable and the collective or single authorship variable are both independent contributors to higher MADRS values. Or maybe neither value causes higher scores, and the relatively higher MADRS results are actually attributable to some other unidentified variable that is coincidentally reflected in both the authorship and court level variables. In reality, the current quantitative analysis cannot answer these causation questions—it can only highlight the relationships between scores and variables, which might signal where a closer look (through a study that is designed to expose causative factors) would be warranted.

E. Scoring Across Jurisdictions

The breakdown of average MADRS result for opinions and decisions based on jurisdiction, and the number from each jurisdiction, is shown below in Table 2.

Table 2: Average MADRS Results by Jurisdiction

Jurisdiction (Total Number in Parentheses)	MADRS Average
British Columbia (N=193)	4.57
Yukon (N=60)	4.59
Alberta (N=140)	4.70
Ontario (N=281)	4.73
Nova Scotia (N=116)	4.78
New Brunswick (N=89)	4.82
Saskatchewan (N=123)	4.84

56. Madden, "Message Undeliverable", *supra* note 21.

Prince Edward Island (N=41)	4.87
Newfoundland and Labrador (N=78)	4.88
Nunavut (N=35)	4.92
Manitoba (N=104)	4.92
Canada – Federal (N=331)	4.94
Northwest Territories (N=32)	4.96

While these results suggest that there are not massive variations in average scores across jurisdictions, the differences are noticeable. Perhaps one might explain the higher scores in the Canadian federal jurisdiction by noting that these courts and tribunals generally deal with legal subject areas that are different from those dealt with by courts and tribunals in provinces and territories, because of how which legislative competence is divided between levels of government in Canada.[57] However, this explanation does nothing to account for the substantial readability differences between courts and tribunals in British Columbia, on the one hand, and the Northwest Territories or Manitoba, on the other hand—since the jurisdictional scopes of the courts of each province and territory are substantially similar.

If nothing else, the above results might suggest where opportunities for dialogue across courts and tribunals from the different jurisdictions exist. Exchanges of ideas, techniques, professional development experiences, or other related elements of information among the authors from more- and less-readable jurisdictions might help identify strategies to improve readability scores across the board, or to pull less readable jurisdictions towards the readability levels achieved by British Columbia's courts and tribunals.

F. Scoring Across Legal Subject Areas

The breakdown of average MADRS results for opinions and decisions based on legal subject area, and the number from each subject, is shown below in Table 3.

Table 3: Average MADRS Results by Legal Subject Area

Legal Subject (Total Number in Parentheses)	MADRS Average
Motor Vehicle (N=36)	4.30
Social Security (N=59)	4.30

57. *The Constitution Act*, 1867, 30 & 31 Vict, c 3, ss 91 and 92.

Family (N=128)	4.34
Human Rights (N=36)	4.36
Health (N=37)	4.50
Negligence (N=26)	4.59
Regulatory (N=16)	4.62
Workers Compensation (N=25)	4.63
Contracts (N=84)	4.63
Utilities (N=25)	4.66
Information and Privacy (N=29)	4.75
Immigration and Refugee (N=76)	4.76
Insurance (N=49)	4.77
Wills and Estates (N=21)	4.79
Criminal Law and Procedure, including Sentencing (N=379)	4.81
Real Property (N=94)	4.82
Constitutional (N=49)	4.91
Labour and Employment (N=81)	4.97
Law of Professions (N=29)	5.00
Tax (N=47)	5.01
Land Use (N=22)	5.09
Administrative Law (N=46)	5.15
Evidence (N=10)	5.21
Trade (N=25)	5.24
Civil and Federal Procedure (N=117)	5.27
Class Actions (N=15)	5.32
Intellectual Property (N=11)	5.45

The relative ranks of many of these results are not particularly surprising or problematic. In each of trade, class action, and intellectual property (IP) law, one would expect the parties to be represented by lawyers. And, in both trade and IP law, one might also expect the parties to be relatively sophisticated, such that they could understand a more complex adjudicative decision even without the assistance of a lawyer, since a certain degree of business sophistication would be generally expected of someone working in trade or IP fields.

However, there is more cause for concern in other areas. Administrative justice is intended to be simpler and more accessible than traditional forms of court-based justice.[58] It is disappointing to see that decisions in this area of law are markedly less readable than both the average decision (across all subjects) and decisions in many other areas of law. Where only one-third of Canada's adult population has a completed undergraduate degree,[59] the average administrative law decision is likely to be beyond the comprehension of a large majority of Canadians.

Similarly, it is unfortunate that procedural law decisions are so difficult to read. These decisions often involve unrepresented litigants whose cases are brought to an end because of a procedural rule, without ever being adjudicated on the merits of the case.[60] And, as research conducted by the Social Security Tribunal of Canada suggests, one of the single most important things that ordinary people are looking for in their decisions is an understandable explanation of what they must do next (procedurally) to move their cases forward.[61] Where the average procedural law decision is written at a level that would require somewhere between a completed undergraduate degree and a law degree for a reader to understand the content of the decision, it seems likely that many litigants will find these decisions to be unreadable.

G. Individual Author Scores: Highs and Lows

Since each decision file that was scored within the current dataset was coded to include metadata consisting of the name of the authoring judge or tribunal member, it is possible to explore how different writers performed from a readability perspective. The idea of identifying decision-related trends or information in respect of individual judges is not new and often yields fascinating insights. Rehaag, for instance, has undertaken a large-scale study of the rates at which individual Federal Court judges grant leave to seek judicial review in refugee matters, and has found astonishingly large variances (with leave being granted 77.97% of the time for the judge with the highest grant

58. Lorne Sossin, "Designing Administrative Justice" (2017) 24:1 Windsor YB Access Just 87 at 94.

59. Statistics Canada, "Table 98-10-0384-01: Highest level of education by census year: Canada, provinces and territories, census metropolitan areas and census agglomerations" (9 December, 2022), online: <150.statcan.gc.ca> [perma.cc/NZ5W-E8UY].

60. See e.g. *Chisan v Klooster*, 2022 ABQB 502.

61. Social Security Tribunal of Canada, "Evolution of Plain Language Decision Writing" (28 February 2023), at s 5.4, online: <sst-tss.gc.ca> [perma.cc/F79R-6EUG]. In this study, the tribunal surveyed and sought qualitative input from claimants who had received decisions from the tribunal—among several other data sources used within the study. More than anything else, claimants expressed a desire for more clear information about what they needed to do next.

rate, and 1.36% of the time for the judge with the lowest grant rate).[62] This research caused the Court to consult with Rehaag, and to make efforts to reduce variances rates.[63] In another context more related to the language of court decisions, Craig undertook a comprehensive qualitative study of the language in Watt JA's more notorious decisions, to argue that his focus on stylistic quirks in his decisions was disrespectful to the people affected by the crimes that were at issue in the decisions.[64] As these examples show, there is often value in speaking bluntly (but not disrespectfully) about the ways in which particular judges make or write their decisions.

However, before discussing some of the potentially more interesting individual results from the present study, a reiteration of my previous caution about how these results must be understood is needed here. The decisions that contributed to the results described in this article were not randomly sampled from the larger population of Canadian decisions. This means that the current results cannot reliably be used to extrapolate toward conclusions about all Canadian cases.

Additionally, there were 1,504 individually written decisions within the dataset from 840 discrete authors—for an average of approximately 1.8 decisions from each author. If most judges and tribunal members write between 20–50 decisions per year, then an actual sample size of somewhere between 19–40 decisions from each judge would be needed in order to have reliable data about each individual decision-maker's average readability score. With one exception,[65] there were no decision-makers within the current dataset for whom there were more than thirteen decisions. A small minority of authors had more than three decisions. Consequently, the individual results described below are unlikely to be representative of any decision-maker's readability performance across all of their cases.

That said, several decision-makers with at least three results from the current study demonstrated a consistent ability to write highly readable decisions. The most readable average for an author with at least three decisions was Jessica Derynck of the British Columbia Human Rights Tribunal (BCHRT). Her average MADRS value was 3.48, based on three decision scores. Nothing in Derynck's biography obviously suggests that she would produce more readable decisions than her peers:

62. Sean Rehaag, "Judicial Review of Refugee Determinations: The Luck of the Draw?" (2012) 38:1 Queen's LJ 1 at 25.

63. Sean Rehaag, "Judicial Review of Refugee Determinations (II): Revisiting the Luck of the Draw" (2019) 45:1 Queen's LJ 1 at 3–5.

64. Elaine Craig, "Judicial Audiences: A Case Study of Justice David Watt's Literary Judgments" (2018) 64:2 McGill LJ 309.

65. Ronald J Kruzeniski, who is Saskatchewan's Information and Privacy Commissioner, wrote 23 decisions that appear within the current set of results.

> Ms. Derynck holds a Bachelor of Applied Arts degree from Ryerson University (2002) and a law degree from the University of Victoria (2009). She was called to the bar of British Columbia in 2010. [. . .] Before her appointment to the Tribunal, Ms. Derynck worked as in-house legal counsel at a union for six years. Previously she practiced human rights, labour and employment law at a law firm in Vancouver.[66]

If anything, Derynck's previous practice experience may have been expected to condition her to write more complicated legal documents. As in-house counsel, her clients would likely be more sophisticated than many users of the legal system, so her need to address them in simple terms would not be particularly strong. Maybe one's background is not a useful predictor of how readably one will write after assuming an adjudicative position.

The second-most readable average for an author again came from a member of the BCHRT: Emily Ohler, the Chairperson of the Tribunal. Ohler had four decisions within the current study, and a MADRS average of 3.67 for these decisions. It is encouraging to see readable decisions being produced by a leader within the tribunal, ideally as a model for other members to emulate.

While the above results do not necessarily show that everyone at the BCHRT, or even Ohler and Derynck, consistently write at highly readable levels, they do point to a potentially useful future comparative study. Does the BCHRT actually write decisions that are more readable than decisions from other human rights tribunals in Canada? And, if so, what factors are driving these better readability levels? Is the BCHRT doing something that other tribunals, like the CHRT, could also do in order to improve readability—especially considering that the core legal issues that human rights tribunals deal with are all very similar?

The third most-readable MADRS average for an author with at least three scores in the current study is from a judge of Nova Scotia's Provincial (inferior) Court: Judge van der Hoek. Judge van der Hoek had three results that averaged to 3.71—well below the average of 4.45 that I calculated for all 244 inferior court decisions within the current study.

It can be helpful to have an objective (although admittedly imperfect) basis for our assessments of how well particular authors write decisions, at least with respect to the readability dimension of the decisions. This information might serve as a useful starting point for further qualitative assessments that seek to more holistically identify who writes exemplary decisions, or to identify writing mentors for new decision-makers. The point here is not to state conclusively how decision-makers or society will benefit from our ability to identify the

66. BC Human Rights Tribunal, "Tribunal Members" (last modified 29 January 2026), online: <bchrt.bc.ca/tribunal/organization/members/> [perma.cc/5BBW-W92R].

decisions and authors whose scores fall at the most readable end of the MADRS spectrum, but simply to suggest how this knowledge might represent a first step in leading us toward the potentially useful goal of making many decisions more readable.

At the other end of the spectrum, the author with the least readable average MADRS value, who had at least three decisions in the study, was Charlesworth J from the Nunavut Court of Justice. Justice Charlesworth had four decisions within this study, that averaged a score of 5.65. Justice Charlesworth's appointment announcement explains that she had previously practiced criminal law, family law representing children, and worked within Queen's University's Legal Aid clinic.[67] In light of her location in Nunavut, where 57% of the population does not have a completed secondary school (or equivalent) level of education,[68] it would be interesting to see whether Charlesworth J's readability scores, and those of her judicial colleagues in Nunavut, score similarly high on the MADRS scale across a much wider sample of cases. The education data for this jurisdiction suggests that readability scores ought to be much closer to, or below, 3.0 on the MADRS scale in order to be understood by most members of the population there.

Justice Pfuetzner of the Court of Appeal of Manitoba had the second-least readable average decision score for three or more decisions within the current study. Her three decisions averaged a score of 5.56 on the MADRS scale. Justice Campbell of the Supreme Court of Yukon, and Brown J (formerly of the Supreme Court of Canada) were tied with the third-least readable average decision score, coming in at 5.55. In Brown J's case, the reported average includes all of his decisions from 2022, so it provides a clear and accurate picture of his readability performance that year without any need for statistical inferences.

Conclusion: Readability Data for the Masses?

I have highlighted above some of the results from this article's study of a diverse corpus of Canadian adjudicative decisions that struck me as being among the most interesting. But I appreciate as both a producer and consumer of scholarly research that it can often be unsatisfying to only have access to the

67. Department of Justice Canada, "Government of Canada Announces Judicial Appointments in Nunavut" (22 June 2018) online (news release): <canada.ca> [web.archive.org/web/20260228031135/https://www.canada.ca/en/department-justice/news/2018/06/government-of-canada-announces-judicial-appointments-in-nunavut.html].

68. Statistics Canada, "Table 98-10-0385-01: High school completion by census year: Canada, provinces and territories, census metropolitan areas and census agglomerations" (4 October 2023), online: <www150.statcan.gc.ca> [perma.cc/P7K2-FQP6].

analysis and data that an author has chosen to focus upon in their report of a particular study. In an attempt to free readers of this dissertation from the shackles of my own approach to considering and analyzing the results of this study, I have made the full spreadsheet of results publicly available to all those who might find value in different parts of the results.[69]

This results file, in the form of an ".xlsx" spreadsheet, will allow anyone who downloads a copy of the file to organize, search, and analyze the results in customizable ways. For instance, one might want to explore whether average readability levels are different between decisions from judges who are women, compared to tribunal members who are women. Alternately, one might look at readability differences between courts of appeal for each of the different jurisdictions, between criminal law decisions from the federal versus provincial and territorial jurisdictions, or between family law decisions from inferior versus superior courts. All of these search queries are possible by extracting the relevant subset of coded results from the publicly available results file.

It is also possible to search for results from a particular author (from among the 840 who have results within the current study), or a particular decision (from among the 1,621 decisions within the studied corpus). In other words, every author whose decisions were scored within this study can find their own scores, and the sub-scores (or component scores) for each linguistic variable that contributes to the MADRS formula, within the results file. This information might help authors realize, for instance, whether their syntax is driving up their MADRS scores—due to a higher proportion of sentences with many dependent clauses, that results in higher scores, for the variable Dependents per nominal (no pronouns, standard deviation) / *Nominal_Deps_NN_StDev*. Or it might help authors see if their use of legalese is higher within their decisions than in those of their peers, as reflected in the variable General Inquirer Database—Legal (Nouns) / *GI_Leg_Nouns*, among other things.

The data is not only useful to authors whose results are reported within the current study. Every author of future adjudicative decisions now has access to low-scoring MADRS files across a wide range of legal subjects. Ideally, this type of readability-scored precedent might help authors inductively identify, from decisions at the more readable end of the MADRS spectrum, the specific types of writing (on the topics that are relevant to these future authors) that make for more readable decisions. The low-scoring files might serve as models for future authors to emulate if they are striving to reach wider audiences with lower levels of reading competencies. Research in multiple other domains has shown that emulation, the use of exemplar texts to help demonstrate to writers how their own writing should look (to improve the quality, or readability, or conformity to genre expectations), and other critical observational learning approaches can

69. Mike Madden, "MADRS_MaddenDissertation_Chapter6Results" (30 May 2024), online (dataset): <doi.org/10.7910/DVN/NIQ2KJ>.

all be successful strategies that drive change in the subject's writing.[70] Writers can become better at their craft by observing and reflecting upon the models produced by others.

And, if nothing else, perhaps knowing that personalized readability scores about a large sample of adjudicative decision-makers are now in the public domain might motivate all such decision-makers to take the time needed to write future decisions in ways that will effectively communicate the contents of their decisions to their audiences. Scholars, peers, politicians, and members of the public may henceforward pay closer attention to readability metrics in their assessments of how well they think that judges and tribunal members are performing their functions—and this possibility may put pressure on decision-makers to better tailor their language within decisions to the education and reading competency levels of audience members for their decisions.

This would not necessarily be an unqualified positive development: there may be both predictable and hidden adverse consequences if individual judges' writings became a more prominent subject of public and political commentary. But a heightened focus on the readability of adjudicative decisions does not necessarily mean that decision-makers should expect pressure to simplify the language of their decision in all cases—since the important consideration is not whether everyone can understand a decision, but whether every *audience member* can understand the decision. Where the audience is sophisticated and narrow in scope, it may not be necessary to focus much on producing highly readable decisions.

In other cases, however, pressure to produce more readable decisions might nudge decision-makers to be more self-critical and reflective about their writing;

---

70. See e.g. Zhiwei Wu, "Understanding Students' Mimicry, Emulation and Imitation of Genre Exemplars: An Exploratory Study" (2019) 54 English for Specific Purposes 127 (in the context of academic writing); Tzipora Rakedzon & Ayelet Baram-Tsabari, "To Make a Long Story Short: A Rubric for Assessing Graduate Students' Academic and Popular Science Writing Skills" (2017) 32 Assessing Writing 28 (in the context of writing about scientific topics, including for the purpose of making the writing more readable); Martine AH Braaksma et al, "What Observational Learning Entails: A Multiple Case Study" (2006) 6:1 L1 – Educational Studies in Language and Literature 31 (in the context of secondary school students' writing). Braaksma et al observed the following:

> Results supported our assumptions about the effectiveness of observational learning. Observers were strongly engaged in metacognitive activities. They observed the models' writing, identified and conceptualized the writing strategies, evaluated the performance of the models and reflected explicitly on the observed performances. The performance of these activities suggests that observers internalized, applied, and developed criteria for effective writing.

Ibid at 53.

it might motivate them to attend (and/or more critically evaluate the usefulness of particular) education and training seminars about decision-writing; and it might encourage them to seek out models from among their peers for the purposes of emulation (to the extent that this strategy might be appropriate and helpful). It is possible that any or all of these steps would holistically improve the comprehensibility of a judge or tribunal member's decisions. And it is likely that any improvements in the comprehensibility of the author's decisions would be reflected in reduced MADRS scores for those new decisions, relative to the previous ones.

The preceding paragraph essentially proposes an ongoing feedback loop for authors that involves MADRS results based on the following broad steps:

1. a decision-maker assesses their current readability scores;[71]

2. if those scores are not consistent with the decision-maker's expectations in terms of the likelihood of the decisions being understood by the intended audiences, then the decision-maker initiates some sort of intervention (e.g., writing seminar, self-study, peer feedback, expert feedback, etc.) designed to improve the readability of their decisions;

3. the decision-maker implements the lessons learned from the intervention within future decisions; and

4. steps 1–3 are repeated as necessary until the decision-maker is satisfied with the readability of their decisions.

Time will tell whether judges and tribunal members will adopt this process or otherwise increase their efforts to write more readable decisions, but at least there is now a large body of public data that might shine some light on the areas where readability needs of audiences are probably not being met. Where the average decision or opinion within the current 1,621-file database scores at a level that is very close to requiring a completed university degree from readers in order for these readers to understand the decision, and where most Canadians do not have this level of education, one can reasonably argue that we are now overdue for the type of readability nudge that this article hopes to generate within the community of Canadian adjudicative decision-makers.

71. This can easily be done online using the web calculator availiable at www.readabilitytools.com.