Putting Numbers to Words:
Measuring the Readability of Court and Administrative Tribunal Decisions in Canada

Michael Madden

A thesis submitted to the University of Ottawa in partial fulfillment of the requirements for the Doctorate in Philosophy degree in Law

Faculty of Law
University of Ottawa

© Michael Madden, Ottawa, Canada, 2025

Table of Contents

List of Charts

List of Tables

List of Figures

List of Abbreviations

AI – Artificial Intelligence
ARI – Automated Readability Index
BCCRT – British Columbia Civil Resolution Tribunal
CanLII – Canadian Legal Information Institute
CML2RI – Coh-Metrix Second Language Reading Index
CHRT – Canadian Human Rights Tribunal
FC – Federal Court
FCA – Federal Court of Appeal
FKGL – Flesch-Kincaid Grade Level Formula
FRE – Flesch Reading Ease Formula
MADRS – Madden's Adjudicative Decision Readability Score
ML – Machine Learning
NLP – Natural Language Processing
ONCJ – Ontario Court of Justice
SCC – Supreme Court of Canada
SMOG – Simple Measure of Gobbledygook
SRL – Self-represented litigant
SST – Social Security Tribunal of Canada

Abstract

This inter-disciplinary law and linguistics dissertation empirically examines the quantitative readability levels of Canadian court and administrative tribunal decisions. It provides current, detailed, and relevant analyses about the extents to which readers of different education levels are likely to understand these adjudicative decisions.

The project makes two original contributions to our knowledge about readability levels of adjudicative decisions. First, it uses linguistic information drawn from recent Canadian decision texts that were scored for readability by human expert raters to create a new law-specific readability formula. Second, it applies the new readability formula to thousands of Canadian decisions from different court levels, authors, legal subject areas, and specialized contexts, in order to assess, compare, and analyze the readability levels of these decisions.

The analyses within this dissertation demonstrate that, with very few exceptions, Canadian judges and administrative tribunal members are writing decisions that cannot easily be understood by large segments of the Canadian population – likely numbering in the millions of people. This research suggests that these authors need to do more, or different, work in order to make their decisions more understandable to the people who may be interested in, or who are affected by, their decisions.

Cette thèse interdisciplinaire en droit et en linguistique examine empiriquement les niveaux quantitatifs de lisibilité des décisions des tribunaux judiciaires et administratifs canadiens. Elle fournit des analyses actuelles, détaillées et pertinentes qui nous permettent de savoir si les lecteurs de différents niveaux d'éducation sont susceptibles de comprendre ces décisions judiciaires.

La thèse apporte deux contributions originales à nos connaissances sur les niveaux de lisibilité des décisions. Premièrement, la thèse utilise des informations linguistiques tirées de textes de décisions canadiennes récentes dont la lisibilité a été évaluée par des évaluateurs experts humains, afin de créer une nouvelle formule de lisibilité spécifique au droit. Deuxièmement, elle applique la nouvelle formule de lisibilité à des milliers de décisions canadiennes provenant de différents niveaux de tribunaux, d'auteurs, de domaines juridiques et de contextes spécialisés, afin d'évaluer, de comparer et d'analyser les niveaux de lisibilité de ces décisions.

Les analyses effectuées dans le cadre de cette thèse démontrent qu'à quelques exceptions près, les juges et les membres des tribunaux administratifs canadiens rédigent des décisions qui ne sont pas facilement compréhensibles pour de nombreux Canadiens - probablement des millions de personnes. Cette recherche suggère que ces auteurs doivent travailler davantage, ou différemment, afin de rendre leurs décisions plus compréhensibles pour les personnes qui peuvent être intéressées ou affectées par leurs décisions.

Acknowledgements

This research was generously funded by the Social Sciences and Humanities Research Council through the award of a *Joseph-Armand Bombardier Graduate Scholarship* (Doctoral).

To carry out this research, I obtained approval from the University of Ottawa's Office of Research Ethics (see Appendix 1 – Office of Research Ethics Certificate of Approval).

When offering acknowledgements, it is common for a person to list and thank the many professional colleagues and mentors who helped them to achieve their goals, before concluding with a "last, but not least" message of thanks to their family members.

I can't follow this convention. My family members will never fall into a "last, but not least" category in my world. They are, and will always be, the most important people in my life. They are also the first people who I must thank, because they each encouraged me, motivated me, supported me, and – unlike with my previous 3 graduate theses – showed genuine interest in the contents of this dissertation.

Dylan is 12. He often sat in my office as I was writing this dissertation and asked me thoughtful questions about what I was doing, and why I was doing these things. He has a scientific mind, and used it to probe me on method-related choices that I had made in my experimental design. Talking through the dissertation with Dylan always helped me to better explain, in writing, what I was doing within the document. Thanks for this, Dylan.

Hallie is 20. She is on the cusp of entering into law school. She is passionate about justice. And she is passionate about reading and writing effectively. She is not a big fan of certain entrenched power structures, or of any situations wherein language is used to make powerful people feel smart at the expense of less powerful people. So, she gets why I wanted to write this thesis. And she was often the first person with whom I chatted when I reached different breakthrough stages of the project – offering suggestions and insights that I would never have considered on my own. She also helped me in my task of finding and communicating effectively with my student volunteers. Thank you for all of these things, Hallie.

But my wife, Sandra, is the one who really helped me to make this project happen. For the last several years, she took care of running the house, handling the minutiae, and making time for everyone else who I had neglected while I was writing, so that I could move forward with my "paper." More importantly, she used her own expertise (from her training as a physician) on reading, understanding, and presenting statistical information to help me with critical parts of the dissertation when I was working with statistics and data analysis. Her suggestions on organizing, sorting, and working through the hundreds of thousands of linguistic data points that were collected in the process of writing this thesis were absolutely invaluable, and probably saved me months of work. Sandra is also far more practical and diplomatic than I am, so she helped me to message

some of the more sensitive parts of this thesis with tact that I don't naturally possess. Thank you for everything, Sandra.

As someone who is fascinated by both the English language and by law, I have wanted to undertake this research project for a long time. After a military career that involved some moving around from place to place, and a workload that did not really give me the time that I needed to complete this project, I eventually found myself retired in Ottawa. Ottawa – the city where Professor Elizabeth Judge teaches and conducts research at the University of Ottawa's Faculty of Law. Professor Judge is probably the only person in Canada who could have supervised me in writing this thesis. She knows the subject area in a way that is literally awesome. But that's a small part of what a supervisor needs to know. She also knows her institution and its people: who to approach for different purposes, and how to frame a message in a way that makes things happen when they need to happen. And, perhaps most importantly, she came to know me almost immediately. She knew exactly what kind of talk I needed to hear from her (and there were so many different kinds of talks that I needed) well before I reached any crisis points. She was direct, encouraging, kind, knowledgeable, and understanding with me throughout the journey. Thank you, Elizabeth.

I am also indebted to my two other committee members, Professors Suzanne Bouclin and Wolfgang Alschner. Together, they steered me in all the right directions – particularly at the point when I was drafting my dissertation proposal. Their critical input at this stage made the process of writing the dissertation relatively easy, because the roadmap that we developed in the proposal was so robust. Thanks to each of you for all the time and energy that you spent on me.

I remain tremendously grateful to my two mentors from Dalhousie University's Schulich School of Law: Professors Steve Coughlan and Rob Currie. Both of you convinced me that I was capable of completing a PhD in law – even if I didn't have the stamina to do it with you in 2014. Steve's encouragement of my early empirical research efforts in law played a major role in convincing me that I could succeed in this empirical dissertation, and his advice on so many other topics has guided me well over the last 15 years. Rob's friendship and support of my military career, research projects, and university teaching experiences has also been instrumental in making me into the scholar who wrote this dissertation. It gave me the confidence to see myself as he saw me: as someone with worthwhile things to say about law. Rob and Steve, thanks for having my back through it all.

Finally, I owe a huge debt of gratitude to my 6 expert volunteers, who selflessly dedicated days of their free time to the task of reading and scoring legal texts. Five of these volunteers agreed to be identified by name, and copies of their short biographies are included in Appendix 2. They are an impressive group, and this project could not have happened without their participation. They contributed with tremendous focus and the combined wisdom of decades of legal practice experience dealing with people who often struggle to understand what the law means. Thank you Johanna, Ryan, Sarah, Seamus, James, and Volunteer 6.

1. Introduction

1.1 Overview

Canada is a country that "recognize[s] the supremacy of […] the rule of law,"[1] and a primary requirement of the rule of law is that law "must be capable of guiding the behaviour of its subjects."[2] Legal theorists and commentators seem to widely accept the idea that people should be able to figure out what the law requires of them.[3] The law, however, consists of many different sources. For instance, Canada's common law legal system includes a combination of rules, powers, and obligations, some of which exist only within statutes,[4] or case law,[5] or the Crown prerogative,[6] or even customary international law.[7] There is no principled reason why (and no legal scholar of whom I am aware has suggested that) only **some** sources of law need to be understandable to the people who are governed by the law's rules. In fact, the opposite suggestion has been advanced, at least with respect to case law: "[w]riting is the conduit through which courts engage with the

---

[1] *Canadian Charter of Rights and Freedoms*, Part I of the *Constitution Act, 1982*, being Schedule B to the *Canada Act 1982 (UK)*, 1982, c 11, preamble.

[2] Joseph Raz, *The Authority of Law* (Oxford: Oxford UP, 1979) at 213-14.

[3] This principle is reflected in the opening sentence of the issue paper published by New Zealand Law Commission, *Presentation of New Zealand Statute Law* (Wellington: Office of Parliamentary Counsel, 2007) at 8, online: <https://www.lawcom.govt.nz/sites/default/files/projectAvailableFormats/NZLC%20IP2.pdf?>: "Citizens should be able to know and understand the law that affects them. It is unfair to require them to obey it otherwise. This is an aspect of the rule of law". The same principle is seen in canonical legal and rule of law texts, such as Lon L Fuller, *The Morality of Law*. Revised ed (New Haven: Yale UP, 1971) at 39; and, John Finnis, *Natural Law and Natural Rights* (Oxford: Oxford UP, 2011) at 270-71. The principle continues to form the basis of arguments about access to law: Michael Curtotti & Eric McCreath, "A Right to Access Implies a Right to Know: An Open Online Platform for Research on the Readability of Law" (2013) 1:1 Journal of Open Access Law 1 at 6-7 ("the axiomatic position taken by this paper is that all individuals subject to law are entitled to know its content and therefore to have it written in a way which is reasonably accessible to them").

[4] See, e.g., *National Defence Act*, RSC 1985 c N-5, s 155(1) (empowering any officer of the Canadian Forces to arrest certain other individuals without a warrant).

[5] See, e.g., *R v Fearon*, 2014 SCC 77 (articulating a common law power that does not exist within any other source of law for police to search a person incident to that person's arrest).

[6] See, e.g., Patrick F Baud, "The Crown's Prerogatives and the Constitution of Canada" (2021) 3 Journal of Commonwealth Law 219, at 228 (describing the Canadian Crown Prerogative to send armed forces to war – a power that does not exist elsewhere within any statute).

[7] See, e.g., *R v Hape*, 2007 SCC 26 at para 39 (wherein the Supreme Court of Canada acknowledged that prohibitive rules of customary international law are automatically received into the common law of Canada unless legislation contradicts these rules of customary international law).

public. As such, the quality of judicial writing is an important element of the legal system – it determines the clarity of the rules that we live by."[8] In other words, case law should be communicated in a way that facilitates understanding by the people who might be affected by, or interested in, the law that it creates, expounds, or clarifies. And yet, as Whalen notes, we have very little empirical information about the readability levels of judicial decisions.[9]

This dissertation attacks existing knowledge gaps about the extent to which Canadian court and administrative tribunal decisions are readable. It fully accepts the idea that law (including explanations of the law by adjudicative bodies) must be communicated in ways that citizens[10] can reasonably understand and explores the next logical questions that follow from this idea. Specifically, this dissertation identifies how we might measure the readability levels of judicial and administrative tribunal decisions, and then uses the identified approach for measuring readability to report on actual readability levels for a wide variety of these decisions. The results therefore contribute to legal and linguistic knowledge about Canadian courts and tribunals in two distinct and novel ways. First, this dissertation identifies and validates the most appropriate mechanism – a new law-specific readability formula derived from expert human scores of adjudicative texts – for assessing readability in the unique context of Canadian legal/adjudicative English-language decisions. And, second, the dissertation provides a current snapshot, and analysis, of actual decision readability levels for different Canadian courts and administrative tribunals.

---

[8] Ryan Whalen, "Judicial Gobbledygook: The Readability of Supreme Court Writing" (2016) 125 Yale Law Journal Forum 200 at 200 ["Judicial Gobbledygoook"].
[9] Whalen, "Judicial Gobbledygook", *supra* note 8 at 200.
[10] I use the terms "citizens", "the people", "the governed", and similar expressions interchangeably within this proposal in order to avoid jarring repetition of any one expression. I acknowledge that in certain legal contexts, "citizens" are a sub-category of people to whom the law applies (in that the law often applies to all individuals within a state's territory, regardless of their citizenship or immigration status), but I will equate these different terms for the sake of aesthetic convenience in this proposal, and take them all to mean "the people to whom the law applies".

1.2 Background – An Emerging Interest in Decision Readability

Before going further, it would be useful here to clarify exactly what is meant by the term "readability" in the context of this dissertation. Readability here generally refers to a quantitative assessment of a text that is designed to predict how well the text might ultimately be understood by a generic reader, based solely on the linguistic features that the text contains.[11] In other words, readability is not concerned with assessments about how much a reader likes a text, or finds a text to be memorable, eloquent, or engaging; readability is, instead, only concerned with the comprehensibility of a text. To be clear, characteristics of a text that make it engaging or memorable for a reader may, under some circumstances, have a positive impact on the comprehensibility of a text – but this will not necessarily be the case. I make this point at the outset to highlight the fact that the "quality" of a text – however one chooses to define this term – is not necessarily the same as the "readability" of a text. A boring and ineloquent text may nonetheless be highly understandable to most readers, while a sublime and aesthetically pleasing text may be very difficult for most readers to understand.

There are two related (but distinct) quantitative dimensions to the readability assessments that are discussed within this dissertation: first, the assessments depend on quantitative measurements of different language properties that are contained within a text; and, second, the measurements are ultimately expressed in quantitative form on a scale of some sort, indicating whether more or less people are likely to understand the text.[12] To be clear, the term readability

---

[11] This understanding of the term readability is consistent with the way in which the term is used throughout William H Dubay, *The Principles of Readability* (Impact Information: California, 2004), and accords with definitions of the terms that have been used by other scholars whose work Dubay reviews (at 3) [*Principles of Readability*].

[12] *See* Alan Bailin & Ann Grafstein, *Readability: Text and Context* (New York: Palgrave Macmillan, 2016) at 15-45 (wherein the authors describe many of existing readability formulas, the quantitative variables upon which each formula depends, and the numerical scoring scale that each formula uses to express the formula's output) [*Readability*].

can (and does) take on other non-quantitative meanings in other contexts,[13] so I do not suggest that the way in which the term is used in this dissertation is uniquely correct.

Readability in the context of this dissertation should therefore be understood as a predictive measurement of ultimate comprehension, but as one that does not take into account the unique characteristics of any individual reader who might come across the text. Specifically, readability does not consider how factors like a person's background knowledge or interest might affect that person's ability to actually understand a text.[14] Instead, readability assessments focus only on the text itself, and on quantitative aspects of how the text uses language to express meaning. So, readability measurements can tell us how likely – based on relevant weights that are assigned to linguistic variables that act as proxies for ultimate comprehension (in ways that have been found to be statistically valid) – audiences in general will be to understand a text. But these readability measurements cannot tell us how likely a particular reader (with complex and unique lived experiences) will be to understand a particular text.

Additionally, because readability measurements only use linguistic features (that are correlated with, but that do not necessarily cause better levels of reading comprehension) to predict comprehension levels,[15] these measurements can only signal to a writer in a limited way how that writer might change a particular text in order to make it more understandable. Changing one's

---

[13] *See, e.g.*, Jeanne S Chall, "Varying Approaches to Readability Measurement" (1996) 25:1 Revue québécoise de linguistique 23, at 33-35 (discussing qualitative, judgement-based approaches toward assessing text readability and quality).

[14] Dubay, *Principles of Readability*, *supra* note 11.

[15] See Bailin & Grafstein, *Readability, supra* note 12 at 23-24 (for a criticism of the assumption that correlations between linguistic features and reading comprehension successes are also proof of a causative relationship between these two variables): "[t]he major issue here is not whether in general simple sentences are or are not easier to read than more complex ones. The underlying question is whether or not a simple correlation is sufficient to attribute causality in relation to readability. Is the mere fact that one can find a correlation between reading levels and the number of simple sentences contained in a text enough to establish that simple sentences necessarily make a text easier to read?"

writing in order to score better on a particular readability formula's scale will not necessarily make a text more understandable to any specific reader (or set of readers), since the variables that contribute to every known readability formula do not actually measure reading comprehension: they just predict reading comprehension probabilities for a text, based on the language used within the text. Thus, as Klare – the creator of one established readability formula – has noted, readability assessments were not intended to be used in remedial ways: "[c]ertain principles of readability, based on research done on and with formulas, *can* be used in writing. But this does not mean that a *formula* should be considered a dictator of good style by the writer or a basis for restricting individuality by the editor. Primarily, a formula is a means of *rating* a piece of writing after it has been written."[16]

The above brief description of what the concept of readability entails, and the accompanying cautions about how readability measurements should be used, are probably sufficient for introductory purposes. I will return to a discussion of readability theory, and the ways in which readability formulas have evolved over the last two centuries, in more detail within Chapter 2. For now, having established a rudimentary understanding of how quantitative readability measurements function, it is possible to consider how these formulas are – and could be – used to assess the readability of Canadian adjudicative decisions.

---

[16] George R Klare, "The Measurement of Readability: Useful Information for Communicators" (2000) 24:3 ACM Journal of Computer Documentation 11 at 18 (emphasis in original). For a similar caution, see Edgar Dale & Jeanne S Chall, "A Formula for Predicting Readability" (1948) 27:1 Educational Research Bulletin 11 at 20:

> [W]e must be cautious about "writing for a readability formula." We must remember at all times that a formula is a statistical device. It means [in the context of The Dale-Chall formula] that, on the whole, longer sentences make comprehension more difficult. This does not mean that all long sentences are hard to read and understand. There are some very short sentences that may be harder to comprehend than longer ones. The same holds true for the use of familiar words. On the whole, the more unfamiliar the words used, the harder the material will be to understand. But sometimes familiar words are used in a symbolic or metaphoric sense. "To be or not to be" is not an easy idea although the sentence is short and the separate words used would usually be called simple and familiar ones. Readability formulas are not sensitive to such subtle variations in meaning.

Until recently, no Canadian courts or administrative tribunals made public efforts to establish readability targets for their written products, or to measure actual readability levels for their decisions. In 2018, however, Canada's Social Security Tribunal (SST) began training its administrative decision-makers to write in more plain language, as part of an effort "to make all aspects of the Tribunal's practices more accessible to Canadians."[17] The Chair of the Tribunal "**set a reading target of grade 9 for all decisions**"[18] based on the Flesch-Kincaid Grade Level readability formula.[19] A subsequent study of the Tribunal's decisions found "that 32% (n=60) met that target and a further 33% (n=62) followed closely at grade 10,"[20] which illustrates that more than two thirds of the tribunal's decisions were not achieving the Chair's readability target. The concluding sentences of this study noted both the infrequency of readability studies focused on judicial and quasi-judicial decisions, and the link that the Tribunal perceives between decision readability levels and access to justice: efforts "to systematically measure readability to reveal insights that drive continuous improvement are rare. This report is that attempt in the Social Security Tribunal's ongoing pursuit of enhanced access to justice for Canadians."[21]

The British Columbia Civil Resolution Tribunal (BCCRT), another administrative tribunal that functions much like a small claims court, also strives to ensure that its written products are readable to a large audience:

> Everything at the CRT, from the website, to the Solution Explorer (our expert system), to our emails and forms, to the final decisions, **is written at about a grade 6 reading level**. To do this, we train our staff and members on how to write in plain language, and we audit our writing using online accessibility assessment tools. We also test everything we develop

---

[17] Social Security Tribunal of Canada, "An Evaluation of How Easy it is to Read Decisions of the Social Security Tribunal" (last modified 10 May 2022), online: <https://sst-tss.gc.ca/en/our-work-our-people/evaluation-easy-it-read-decisions-social-security-tribunal> ["An Evaluation"].
[18] Social Security Tribunal of Canada, "An Evaluation", *supra* note 17.
[19] Social Security Tribunal of Canada, "An Evaluation", *supra* note 17.
[20] Social Security Tribunal of Canada, "An Evaluation", *supra* note 17.
[21] Social Security Tribunal of Canada, "An Evaluation", *supra* note 17.

> with community legal advocates who represent those with the highest barriers to accessing justice.[22]

The BCCRT's target grade level for its written products is not surprising; the Chairperson who articulated the target, Shannon Salter, has elsewhere referred to research showing "that public health information needs to be written at a **grade 6 reading level** to be widely understood,"[23] and has argued that, "[w]ith equally complex jargon and processes, public legal information must be written at a similar level."[24] Salter also appears to link readability with public access to justice and law: "public legal information and court and tribunal forms routinely use complex language and 'legal-eze' that assume a high degree of facility in English. Part of the CRT's commitment to putting the public first includes truly ensuring that CRT information and forms are as simple and clear as possible."[25]

No other Canadian court or administrative tribunal has published either target, or actual, readability levels of its decisions, but the SST and BCCRT examples may represent the beginning of a movement toward increased interest in quantitative readability studies of the written products that emerge from such institutions.[26] These two tribunals, with their testing processes and their drive for continuous improvement, seem to see the publication of written legal material as an

---

[22] Shannon Salter, "RDO Speak to Shannon Shalter [SIC] About All Things ODR" (November 2019), online: <https://resolvedisputes.online/blog_shannon_salter.html> (containing a transcript of an interview with Shannon Salter, who was the Chair of the BC CRT at the time of the interview) (emphasis added).

[23] Shannon Salter & Darin Thompson, "Public-Centered Civil Justice Redesign: A Case Study of the British Columbia Civil Resolution Tribunal" (2017) 3 McGill Journal of Dispute Resolution 113 at 124 ["Public-Centered Civil Justice Redesign"], citing Richard S Safeer & Jann Keenan, "Health Literacy: the gap between physicians and patients" (2005) 72(3) Am Fam Physician 463 at 468 (emphasis added).

[24] Salter & Thompson, "Public-Centered Civil Justice Redesign", *supra* note 23 at 124.

[25] Shannon Salter, "Online Dispute Resolution and Justice System Integration: British Columbia's Civil Resolution Tribunal" (2017) 34 Windsor Yearbook of Access to Justice 112 at 124.

[26] I would distinguish this movement from the related "plain language" movement in law, which began several decades ago (see, e.g.: Ruth Sullivan, "The Promise of Plain Language Drafting" (2001) 47:1 McGill Law Journal 97 [The Promise of Plain Language Drafting"]; and see David St L Kelly, "Legislative Drafting and Plain English" (1986) 10 Adelaide Law Review 409), because this earlier movement was largely focused on plain language in legislation, contracts, forms, and other legal documents – rather than in judicial decisions.

important element of ongoing communication with, rather than simply a broadcast to, their audiences.

The readability concerns with which these tribunals are occupied raise at least two important questions: (1) how readable should adjudicative decisions be; and, (2) how can we measure the actual readability levels of adjudicative decisions? The first question tries to establish a target, while the second question tries to establish a method for measuring text readability that could, among other things, verify the extent to which texts are attaining the intended standard. My dissertation assumes that it is possible to develop an answer to the first question that is consistent with the overarching principle that people who are governed and affected by the law should be able to understand it, likely through political- or policy-based reasoning (that considers matters such as the intended audience for decisions, the breadth and depth of understanding that an audience would need to have in order for a decision to be considered readable, and whether there could ever be universal, or only context-specific, readability targets for adjudicative decisions). However, my dissertation does not attempt to categorically answer this first question. There probably are, and should be, large variances in the different audience-centric readability standards that might be desired for adjudicative decisions under different circumstances – as the general readability targets set by the BCCRT and SST (i.e.: Grade Levels of 6 and 9, respectively) suggest – so my dissertation largely sets this first question aside as one that needs to be determined on a case-by-case basis by the authors of adjudicative decisions, and by those to whom these authors are accountable.

My dissertation instead focuses primarily on the second question: how can we verify or measure the actual readability levels of adjudicative decisions? In other words, I assume that a decision's author can determine who forms the audience for each decision, and what reading

abilities the members of this audience are likely to have. I therefore try within this dissertation to give these authors a means of measuring, and a better understanding of, whether they might actually reach their audiences with their language. And, after calling attention to actual decision readability levels in select contexts, I offer some initial observations about whether authors in these contexts are likely to be reaching their audiences in the ways that they intend.

Although it might seem, at first glance, that it should be relatively straightforward to develop an answer to this second question about how to measure the readability levels of adjudicative decisions, certain features of legal English complicate the answer.[27] On the one hand, it is clear that several English readability formulas already exist. Two of these formulas – the Flesch Reading Ease (FRE) and the Flesch-Kincaid Grade Level (FKGL) formulas – are incorporated directly into Microsoft Word's "Editor" feature, making them widely available to any users of this software program.[28] And both formulas have already been used to measure the readability of court opinions in at least some past studies.[29] On the other hand, however, these formulas were not created for use on uniquely legal texts, and research from other technical domains has indicated that general-purpose readability formulas do not necessarily perform well when applied in more specialized contexts, where arrangements and uses of the English language may deviate from patterns that one typically encounters in more popular forms of spoken or written English.[30]

---

[27] See, generally, Sol Azuelos-Atias, "On the Incoherence of Legal Language to the General Public" (2011) 24 International Journal for the Semiotics of Law 41 (for an overview of the different ways in which legal language is distinct from, and confusing to lay users of, ordinary language) ["On the Incoherence of Legal Language"].

[28] Richard Johnson, "How to Apply the Flesch Kincaid Readability Formula to Your Content," February 23, 2021, online: <https://www.optimonk.com/how-to-apply-the-flesch-kincaid-readability-formula-to-your-content/> [https://perma.cc/6ZDM-DWND], explains how to calculate one's FKGL score for a document using the embedded calculator within Microsoft Word.

[29] *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* (2017) (Unpublished Ph.D. Dissertation, University of Idaho's College of Graduate Studies); and *see* Stephen M Johnson, "The Changing Discourse of the Supreme Court" (2014) 12:1 University of New Hampshire Law Review 29 ["Changing Discourse"].

[30] Jiaping Zheng & Hong Yu, "Readability Formulas and User Perceptions of Electronic Health Records Difficulty: A Corpus Study" (2017) 19:3 Journal of Medical Internet Research ["Health Records Difficulty"].

For this reason, it cannot be assumed that an existing general-purpose readability formula will accurately predict readability levels of adjudicative opinions and decisions that are written in law's unique style of English.[31] I do not make any such assumption within this dissertation. Instead, in Chapter 3, I create a law-specific readability formula, and test this formula's performance against other general-purpose readability formulas. After this test is complete (and the results of the test show that my law-specific formula outperforms other general-purpose formulas), I apply the formula to thousands of Canadian adjudicative decisions in order to measure their readability levels using the tool that has been demonstrated to be the most accurate predictor of readability scores that human experts would assign to these decisions. But it was not a forgone conclusion that a law-specific formula would outperform other general-purpose formulas. The work that demonstrates how I arrived at this conclusion is contained within, and represents the bulk of, Chapter 3 of this dissertation.

As this brief introduction to readability suggests, my dissertation is centred on the topic of the quantitative readability of Canadian administrative tribunal and court decisions. It attempts to identify a contemporary, precise, and legal domain-specific method for assessing the readability of quasi-judicial and judicial decisions, since no such method is currently in use. It also illustrates how this new method can be used to measure the readability levels of different types of decisions in order to do a number of other things, such as: testing our intuitions about decision readability levels; providing benchmarks for assessing current readability performance and future readability improvements; comparing readability levels across multiple variables; and, providing readability data for large-scale samples of adjudicative decisions in a way that is simply not practicable using

---

[31] *See* Brenda Danet, "Language in the Legal Process" (1980) 14 Law & Society Review 445, at 469, and 476-77 (for descriptions of some of the ways in which legal English is different from more conventional English) ["Language in the Legal Process"].

other qualitative approaches. A more complete picture of the research questions that motivate this dissertation is provided in the following section.

1.3 Research Questions

1.3.1 Core Question: How Readable are the Decisions of Canadian Courts and Administrative Tribunals

The overarching research question that this dissertation asks is simple: how readable are the decisions of Canadian courts and administrative tribunals? These decisions represent the most significant, and typically the only, communicative output from the judges or tribunal members who issue decisions that affect the rights, obligations, privileges, and duties of ordinary people within their jurisdictions.[32] The reasons contained in decisions

> foster better decision making by ensuring that issues and reasoning are well articulated and, therefore, more carefully thought out. The process of writing reasons for decision by itself may be a guarantee of a better decision. Reasons also allow parties to see that the applicable issues have been carefully considered, and are invaluable if a decision is to be appealed, questioned, or considered on judicial review. Those affected may be more likely to feel they were treated fairly and appropriately if reasons are given.[33]

And, although we do not know the extent to which parties will actually read the decisions that dispose of their cases across all courts and tribunals, there is at least some evidence to suggest that parties will read their entire decisions. In a survey of claimants whose cases were decided by the SST, that tribunal found that 87% of claimants read the whole decision; 7% read most or some of the decision; and, 6% read very little or none of the decision.[34]

---

[32] See generally, Barry Sullivan & Ramon Feldbrin, "The Supreme Court and the People: Communicating Decisions to the Public" (2022) 24:1 University of Pennsylvania Journal of Constitutional Law 1.
[33] *Baker v Canada (Minister of Citizenship and Immigration)*, [1999] 2 SCR 817 at para 39 (citations omitted).
[34] Social Security Tribunal of Canada, "Evolution of Plain Language Decision Writing" (February 28, 2023), at s 5.3, online: <https://www.sst-tss.gc.ca/en/our-work-our-people/evaluation-plain-language-decision-writing> ["Evolution of Plain Language Decision Writing"].

However, one should not infer from these statistics that the people who are 'reading' the entire decisions are actually 'understanding' them (and I explain the distinctions between reading, understanding, and using a decision below, in section 2.8). The same survey revealed that 42% of all respondents found the wording in their decisions to be somewhat or very complicated, and 19% of respondents (rising to 42% of respondents who had been **unsuccessful** at the SST) found that the decisions were somewhat or very unclear.[35] These statistics present a compelling case for decision-makers to consider the reading competency levels of the parties before them when writing their decisions. In this sense, it is perhaps important to distinguish between the somewhat mechanical act of reading a text, on the one hand, and the related, but more involved cognitive task of processing a text to the point of understanding it, on the other hand.

Although this SST survey offers us some insight about how parties receive and understand adjudicative decisions, there are reasons for decision-makers to consider more than just the parties when writing decisions. With respect to courts, Goldstein suggests that they have a somewhat unique "obligation to maintain the Constitution, in opinions of the Court and also in concurring and dissenting opinions, as something intelligible. […] The Court's goal is to render opinions […] that contemporary society can understand."[36] According to Goldstein, these elements of intelligibility and accessibility in the decisions from the courts are "requisite to a government by consent."[37]

Ash et al have also seized on this idea of court decisions as a means of demonstrating accountability to the population: "[o]ne way that judges strengthen the legitimacy of their policy

---

[35] Social Security Tribunal of Canada, "Evolution of Plain Language Decision Writing", *supra* note 34.
[36] Joseph Goldstein, The Intelligible Constitution: The Supreme Court's Obligation to Maintain the Constitution as Something We the People Can Understand (New York: Oxford UP, 1992) at 19 [*The Intelligible Constitution*].
[37] Goldstein, *The Intelligible Constitution*, *supra* note 36 at 19.

choices given low accountability is by providing written justifications based on shared principles, which are then published as judicial opinions."[38] These authors go on to infer that, "[p]resumably, this legitimizing function is best served when the general population can understand the written justifications."[39]

In other words, as both Ash et al and Goldstein suggest, judges benefit from a high degree of security in their roles, which shields these decision-makers from many of the more conventional forms of accountability (like performance appraisals, employment sanctions, and an ability to terminate a person without cause if sufficient notice or severance is provided) that other types of employees and workers face in their workplaces. Ash et al and Goldstein imply that courts therefore have related obligations to produce decisions that people can understand, perhaps as a way of compensating for the high-security environment in which judges operate.

This sentiment is echoed in remarks made by a then-judge of the Nova Scotia Court of Appeal who was discussing how to write decisions. In this context, Justice Saunders described his "thesis that we who sit as judges are permitted to do so because of an unwritten pact we have with the community to conduct ourselves competently, respectfully and fairly - an arrangement I call - the 'morality of judicial reasoning.'"[40] The implication that underpins these statements from scholars and judges is essentially the following: if judges want to reassure the population that they are performing their functions legitimately in spite of the relative absence of accountability

---

[38] Elliott Ash, et al, "Translating Legalese: Enhancing Public Understanding of Court Opinions with Legal Summarizers" in CSLAW '24: Proceedings of the Symposium on Computer Science and Law (New York: Association for Computing Machinery, 2024) 136 at 136 ["Translating Legalese"].
[39] Ash, et al, "Translating Legalese", *supra* note 38 at 136.
[40] Jamie WS Saunders, "Reflections on the Art and Science of Decision-Making and Decision-Writing" (Nov 2017), *Professional Development Conference: Maritime Provinces' Law Societies* (Halifax, NS), online: <https://www.courts.ns.ca/sites/default/files/editor-uploads/From%20the%20Bench/Reflections_on_the_Art_and_Science_of_Decision_Making_11_09_17.pdf>, at 71-72 ["Reflections"].

mechanisms, then they must produce understandable decisions that effectively show people how they arrived at their conclusions.

This idea that civic engagement with judicial decisions can occur and contribute to the public good, while perhaps more aspirational than real in many (particularly more mundane) cases, seems to be gaining traction. The Israeli Supreme Court, for instance, recently started televising broadcasts of hearings in selected constitutional cases, and found from the broadcast ratings that citizens were clearly trying to access the work of the court directly, rather than through media summaries or other intermediaries.[41] The Supreme Court of Canada also recently reported that its Case in Brief summaries of decisions had been viewed "almost a million times on the Court's website"[42] over a 30-month period, which suggests that ordinary people are seeking to understand the Court's decisions in fairly high numbers.

This phenomenon of making a court's processes and conclusions available and understandable to the public helps to strengthen the legitimacy of the courts themselves: "courts lack their own coercive powers, however, they depend on public opinion to encourage the political branches to comply with their decisions."[43] In this sense, the rule of law is strengthened when the public is capable of understanding court decisions, because broad public (i.e.: voter) understanding of decisions helps to promote respect for and compliance with the decisions, particularly by elected government officials who rely on public support for re-election. Sullivan & Feldbrin pragmatically observe that "[w]ritten opinions may sometimes seem like an extravagance to busy judges, but

---

[41] Sullivan & Feldbrin, "The Supreme Court and the People", *supra* note 32 at 70-71.

[42] Supreme Court of Canada. "News Release" (8 Oct 2020), online: <https://decisions.scc-csc.ca/scc-csc/news/en/item/6964/index.do>.

[43] Sullivan & Feldbrin, "The Supreme Court and the People", *supra* note 32 at 85. The authors (at FN 352) cite US Supreme Court Justice Frankfurter's dissenting opinion in *Baker v Carr*, (1962) 369 US 186 at 267 in support and explanation of their claim that courts lack coercive powers: "The Court's authority – possessed of neither the purse nor the sword – ultimately rests on sustained public confidence in its moral sanction."

they are central to the public's understanding of how the judicial system is supposed to work."[44] These authors suggest that intelligible written decisions are the means through which judges "demonstrate that they have heard and considered the parties' evidence and legal arguments, and that they have taken seriously what they have heard. It is also through opinions that the reasons supporting judicial rulings are opened to intellectual critique."[45] Sullivan & Feldbrin even suggest that judges have a "democratic duty"[46] to produce understandable decisions: "[w]hen opinions are not written in a way that makes them accessible to the general public, the Court fails in its fundamental obligation, not only to 'say what the law is,' but to do so in a way that is intelligible to citizens."[47] If the people must be capable of understanding court decisions in order for Canada's system of government to operate as it should, then the question of whether, or to what extent, adjudicative decisions in Canada are readable is rather significant.

Beyond just these third-party arguments about how judges should speak within their decisions, judges themselves also repeatedly profess to be communicating with everyday members

[44] Sullivan & Feldbrin, "The Supreme Court and the People", *supra* note 32 at 21.
[45] Sullivan & Feldbrin, "The Supreme Court and the People", *supra* note 32 at 22.
[46] Sullivan & Feldbrin, "The Supreme Court and the People", *supra* note 32 at 28.
[47] Sullivan & Feldbrin, "The Supreme Court and the People", *supra* note 32 at 28.

of the public,[48] at all levels of courts from the Supreme Court of Canada (SCC),[49] to courts of appeal,[50] to superior (trial-level) courts.[51]

Before going further on this point, however, it is perhaps helpful here to briefly discuss the potential sources of information that are available to us about how judges see their audiences. Readers of the 4 previous footnotes will have noticed that the sources in these footnotes were all similar: they were responses to questionnaires that candidates for federal judicial appointments must complete as part of the judicial application process, and that are now generally disclosed to the public once that judge is appointed. More specifically, the quoted text in each of these footnotes

---

[48] See, e.g., Canada, Department of Justice. *The Honourable Mahmud Jamal's Questionnaire (Questionnaire for the Supreme Court of Canada Judicial Appointment Process)* (17 Jun 2021), online: <https://fja-cmf.gc.ca/scc-csc/2021/nominee-candidat-eng.html>:

> [T]he Court's decisions speak directly to the Canadian public. The public has a right to understand how their highest court is grappling with the most pressing legal issues of the day. As Chief Justice Dickson underscored: 'Our judgments touch the lives of all Canadians. They should convey meaning to all who read them, whether or not they are learned in the law'[…]. This is especially important at a time when many justice system participants are self-represented. Clear and accessible reasons concretely improve access to justice.

[49] Canada, Department of Justice. *The Honourable Michelle O'Bonsawin's Questionnaire (Questionnaire for the Supreme Court of Canada Judicial Appointment Process)* (22 Aug 2022) online: <https://fja-cmf.gc.ca/scc-csc/2022/nominee-candidat-eng.html> [*O'Bonsawin's Questionnaire*]: "Another audience [for SCC decisions] includes lower courts, the legal profession, legislatures, academics, media and the public."

[50] Canada, Department of Justice. *The Honourable Jill Miriam Copeland's Questionnaire (Questionnaire for the Court of Appeal for Ontario's Judicial Appointment Process)* (30 May 2022), online: <https://www.canada.ca/en/department-justice/news/2022/05/the-honourable-jill-miriam-copelands-questionnaire.html>:

> The public is also an important 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. This is not a process of pandering to the public or trying to write a decision that the public will like. Judicial independence and impartiality prevent a judge from proceeding in that manner. But a judge has a duty to explain what decision he or she has made and why. This is an aspect of justice being done and being seen to be done. The public has a right to review the reasons for decision, and comment on and criticize them if they feel that is warranted.

[51] Canada, Department of Justice. *The Honourable Renu Mandhane's Questionnaire (Questionnaire for the Ontario Superior Court of Justice Judicial Appointment Process)* (28 Mar 2022), online: <https://www.canada.ca/en/department-justice/news/2022/03/the-honourable-renu-mandhanes-questionnaire.html>: "Reasons must provide guidance to the public, including the media. This requires reasons that are written in plain language, efficient, and available in accessible formats. Judgments should be fulsome enough that the reader does not need to consult pleadings, case law or secondary sources to understand them. On the other hand, the reasons cannot be overwhelming or require specialized knowledge."

was sourced from each candidate's reply to question 4 in the questionnaire, which asks, "who is the audience" for decisions rendered by the court to which you are applying?[52]

On the one hand, these questionnaires offer us valuable insight into the actual or asserted[53] beliefs of the candidates before they took office as judges in their new courts. On the other hand, however, it is possible that these were uninformed or unrealistic beliefs, since they were expressed by candidates who had not yet started their work of judging at their new courts. Perhaps these beliefs merely represent aspirations from the candidates that do not necessarily map onto their actual practices in writing decisions at their new courts, once they find that they cannot achieve, or they otherwise adjust, their aspirations at these new courts for any number of reasons.

I am inclined to believe that the latter proposition is more accurate than the former, and that candidates' responses on judicial application questionnaires are generally reliable sources of evidence about how judges think and behave. The judges responding to these questionnaires – particularly for applications to courts of appeal and the SCC – have generally been judging already for some time: in 3 of the 4 questionnaire responses that were cited above, the applicants were already occupying positions as federally-appointed judges (but were applying for positions on higher courts),[54] and this path to a position as a federally-appointed judge is not unusual (as seen in the biographies of the many other judges whose application questionnaires are subsequently cited in this dissertation). Furthermore, any applicant for a federally-appointed judicial position must, statutorily, have a minimum of 10 years' experience as a lawyer in order to be appointed,[55]

---

[52] *Supra*, notes 48-51.
[53] I distinguish between actual and asserted beliefs here because it is possible that candidates are expressing sentiments that they think will lead to their appointment as judges instead of sentiments in which they truly believe. This might occur in any situation where the candidates see these two sentiments as being different from one another.
[54] *Supra*, notes 48-50. Justice Mandhane (*supra* note 51), was appointed to the bench directly from the practice of law.
[55] See, e.g., *Judges Act*, RSC, 1985, c J-1, s 3; *Federal Courts Act*, RSC, 1985, c F-7, s 5.3; and, *Supreme Court Act*, RSC, 1985, c S-26, s 5.

so each applicant would likely have a strong understanding of the functions and audiences for judicial decisions from the court to which they are applying, that was acquired over their years of experience as a member of the bar. Thus, while responses on questionnaires from applicants are not necessarily true predictors of how judges conceive of their audiences once they are appointed to their new courts, they are perhaps one of the strongest sources of evidence that we can rely upon under the circumstances.

One of the circumstances that I refer to here is the culture or conception of judicial independence in Canada that often seems to make judges reluctant to speak publicly about their deliberative and decision-writing processes. As Khan observed when conducting research into these facets of judging, many courts and judges refuse, on grounds relating to judicial independence, to answer questions about the content or existence of style and structure guidance documents that the court may use, tools that the courts might use to improve their decisions, and other similar questions pertaining to timelines and processes for completing decisions.[56] This reality makes it difficult to acquire more diverse and accurate sources of information about how sitting judges conceive of the audiences for their decisions – which is another reason why I rely heavily on statements drawn from judicial application questionnaires to determine who falls within the intended audiences for court decisions.

That being said, some sitting judges have spoken publicly and extra-judicially about the audiences for court decisions. For instance, former Chief Justice of Canada McLachlin delivered the following remarks, while she still occupied that role, as part of a joint National Judicial Institute

---

[56] Jon Khan, *'The Life of a Reserve': How Might We Improve the Structure, Content, Accessibility, Length & Timeliness of Judicial Decisions?* (unpublished LL.M. Thesis, University of Toronto, 2019), online: <https://tspace.library.utoronto.ca/bitstream/1807/98120/1/Khan_Jon_%20_201911_LLM_thesis.pdf> at 57-63 [*Life of a Reserve*].

/ Canadian Institute for the Administration of Justice seminar on "Advanced Judgement Writing" (presumably to judges, who are mentioned 14 times in the speech, which does not mention administrative tribunals at all): "Often in provincial courts of appeal and always at the Supreme Court of Canada, the audiences include the parties, the parties' lawyers, interveners, legal practitioners, legal academics, the press **and the general public**. Plain, clear, readily understood language is the only way to address these diverse audiences."[57] Former Chief Justice McLachlin also stated, within the same speech, that "[j]udgment writing at its simplest, is an attempt to communicate. A good judgment is one that bridges the understanding gap between the judge's mind and the recipients' minds and communicates. A judgment that does not communicate is a failed judgment, no matter how learned or gracefully phrased. To communicate, you must know your audience and how they think."[58] From these remarks, it seems that Canada's pre-eminent sitting judge believed that the general public was an audience for her (and other courts') decisions, and that one of her obligations was to understand how this audience thinks so that she might better communicate with this audience. The fact that these remarks were likely delivered to other judges at a seminar that was focused on writing better decisions also suggests that the former Chief Justice's views on this question are accepted by, or are familiar to, the profession of judges in Canada.

---

[57] Beverly McLachlin, "Remarks of the Rt Hon Beverley McLachlin, PC, Chief Justice of Canada" (May 16, 2010), *NJI and CIAJ Advanced Judgment Writing Seminar* (Perth, Ontario), online: <https://ciaj-icaj.ca/wp-content/uploads/podcasts/2020/11/podcast1b_16may10-nji-ciaj-judgment-writing-course-copy.pdf>, at 11 (emphasis added) ["Remarks"].
[58] McLachlin, "Remarks", *supra* note 57 at 10.

A slightly different perspective was offered by a sitting judge of the Nova Scotia Court of Appeal in a presentation that he originally crafted for a "Boot Camp for Decision-Makers" event.[59] This judge, Justice Saunders, told decision-makers,

> [r]emember your audience. First and foremost, you are writing for the parties so that the litigants, win or lose, will understand the basis of the decision. To a lesser extent, your audience is the public, being those members of the community who may be interested in the outcome. And the Bar, in the sense that your decision may have important precedential value to future cases. But do not write for the Court of Appeal.[60]

While the above two (and presumably most other) judges may weigh the importance of their audiences in different ways at different times, it is clear that both recognize the general public as an audience for a decision, and encourage writers of decisions to consider the audience when writing decisions.

This evidence from sitting judges accords with the types of answers that judges offer in response to application questions about audiences for court decisions. It does not seem to be seriously disputed by either sitting or prospective judges that members of the general public are part of the audience for their decisions. We now have extensive evidence from publicly-available application questionnaires submitted by people who have gone on to become judges, but who were not necessarily judges – and who were categorically not judges of the courts to which they were applying – at the time that they provided their questionnaire responses. Although I tend to rely most heavily on these application questionnaire responses throughout this dissertation when describing how judges conceive of their audiences, the above discussion demonstrates why these questionnaires are a valid, if not the best, source for this type of information that we can reasonably procure under the circumstances.

---

[59] Saunders, "Reflections", *supra* note 40 at 2 (FN 1).
[60] Saunders, "Reflections", *supra* note 40 at 71.

But it is not just courts in Canada that render adjudicative decisions: many are made by administrative tribunals. These tribunals also seem to have a lay audience in mind for their decisions: "[t]here is a larger audience for your decisions as well, the most important of which are people who may participate in your process at a future point. […] You should be writing for the average person appearing in front of you. Assume that your reader is generally well-informed but without specialized training in the area being written about."[61] In a recent case involving a judicial review of an administrative tribunal's decision, Canada's Federal Court also endorsed the idea that tribunal members should be writing for non-lawyers: Justice Zinn complimented the tribunal member whose decision he was reviewing for the "well-written and plain language decision."[62] The judge went on to say that the tribunal member had produced "an easily understood decision, summarizing the correct legal procedures and tests while citing no judicial authority, and addressing the decision to the appellant, even though he was represented at the appeal. This approach promotes access to justice and is commendable."[63] As Justice Zinn's words suggest, administrative tribunals should be writing for ordinary people, even if the parties before them are represented by lawyers.

Beyond ordinary members of the public, judges also attempt to communicate with other non-legal professionals within their decisions, from audiences that could include "a journalist, a researcher or an individual seeking to make the law more accessible,"[64] or "others working in the

---

[61] Ombudsman Saskatchewan, *Practice Essentials for Administrative Tribunals* (Saskatchewan: Ministry of Justice and Attorney General, 2020) at 68-69.
[62] *Bashir v Canada (Citizenship and Immigration)*, 2025 FC 19, at para 36.
[63] *Bashir v Canada (Citizenship and Immigration)*, 2025 FC 19, at para 37.
[64] Canada, Department of Justice. *The Honourable Sebastian Grammond's Questionnaire (Questionnaire for the Federal Court Judicial Appointment Process)* (8 Jun 2021), online: <https://www.canada.ca/en/department-justice/news/2021/06/the-honourable-sebastien-grammond-questionnaire.html>.

industry or business of the defendant, and possibly class members [in class action cases],"[65] or "complainants or victims"[66] in criminal cases.

Although some judges seem to perceive of their audiences as only including legal and judicial professionals,[67] the common view expressed by the Canadian judges cited above, and corroborated by information from other sources,[68] suggests that non-legal professionals form part of both the intended and the actual audiences for adjudicative decisions. To date, however, limited (or no) data is available about how well judges' and tribunal members' decisions are capable of being understood by the public. If we accept that non-lawyers are audiences for adjudicative decisions at least some of the time, then scholars, judges, government executives and administrators, and the Canadian public might all want to know the extent to which our court and administrative tribunal decisions are readable.

---

[65] Canada, Department of Justice. *The Honourable Barbara J Norell's Questionnaire (Questionnaire for the Supreme Court of British Columbia Judicial Appointment Process)* (13 Apr 2018), online: <https://www.canada.ca/en/department-justice/news/2018/04/the-honourable-barbara-j-norells-questionnaire.html>.

[66] Canada, Department of Justice. *The Honourable Shaun S. Nakatsuru's Questionnaire (Questionnaire for the Ontario Superior Court of Justice Judicial Appointment Process)* (11 May 2017), online: <https://www.canada.ca/en/department-justice/news/2017/05/the_honourable_justiceshaunsnakatsurusquestionnaire.html>. Justice Nakatsuru appears to have achieved some measure of success in reaching a wide and non-legal audience: see Marlene Leung, "'Transformative' decision: How a judge's decision is turning legal heads," CTV News (9 Mar 2015), online: <https://www.ctvnews.ca/canada/transformative-decision-how-a-judge-s-decision-is-turning-legal-heads-1.2270358>.

[67] For instance, Richard A Posner, "Judges' Writing Styles (And Do They Matter?)" (1995) 62 University of Chicago Law Review 1420, at 1431, notes that, "[f]or many judges the implied audience consists primarily of the judge or judges of the lower court whose decision is being reviewed and the lawyers for the parties. Anyone else is an (authorized) eavesdropper."

[68] See, e.g., Canada, Department of Justice. *The Honourable April D Grosse's Questionnaire (Questionnaire for the Alberta Court of Queen's Bench Judicial Appointment Process)* (11 July 2018), online: <https://www.canada.ca/en/department-justice/news/2018/07/the-honourable-april-d-grosses-questionnaire.html>: In describing how the public is an audience for judicial decisions, Justice Grosse notes that, "with the increase in self-represented litigants, or at least people who take some personal role in assessing their own legal rights without a lawyer, it is now very common for members of the public to consult online databases such as CanLII to review judicial decisions." Additionally, in my preliminary discussions with CanLII's former Chief Executive Officer, for instance, I have discovered that the site generally has somewhere in the vicinity of 525,000 unique visitors who access judicial decisions from the database each month. Since there are less than 150,000 licensed lawyers, and about 2000 federally- and provincially-appointed judges in Canada, one could reasonably infer from CanLII user information that a large number of individuals who are not lawyers or judges are reading Canadian case law.

A small number of Canadian courts and administrative tribunals have recently acknowledged the importance of communicating decisions in ways that the public is more capable of understanding: in addition to the SST and BCCRT examples, the SCC also started issuing a "Case in Brief" summary for each of the court's decisions starting in 2018. According to the SCC, "Cases in Brief are short summaries of the Court's written decisions drafted in reader-friendly language, so that anyone interested can learn about the decisions that affect their lives. They are prepared by communications staff of the Supreme Court of Canada."[69]

Collectively, there seems to be a strong desire to produce readable decisions – as a recent appointee to the SCC, Justice O'Bonsawin, suggests in her *Questionnaire for the Supreme Court of Canada Judicial Appointment Process*: "[t]he Supreme Court of Canada must also write in clear, plain, concise and understandable language so that all its audiences can understand. […]. The Supreme Court of Canada must be mindful that the audience can be highly educated, partially educated and/or uneducated."[70] Justice Gomery makes a similar point: "decisions must avoid the unnecessary use of jargon that make judges and lawyers feel smart but alienate the ordinary reader. To the extent that legal terminology is necessary (as it sometimes is), judges should reduce it, to the extent possible, to language that a non-lawyer can understand."[71]

However, no Canadian court or administrative tribunal other than the BCCRT and the SST has publicly declared how they measure readability, or whether they are meeting their readability targets. To the extent that other courts and tribunals have readability goals, they may not yet be

---

[69] Supreme Court of Canada, "Cases in Brief" (last updated 29 Jul 2022), online: <https://www.scc-csc.ca/case-dossier/cb/index-eng.aspx>.
[70] *O'Bonsawin's Questionnaire*, *supra* note 49.
[71] Canada, Department of Justice. *The Honourable Sally A Gomery's Questionnaire (Questionnaire for the Ontario Superior Court of Justice Judicial Appointment Process)* (3 Nov 2017), online: <https://www.canada.ca/en/department-justice/news/2017/11/the_honourable_sallyagomerysquestionnaire.html>.

matched with strategies to achieve these goals. These institutions may simply be unaware of the potential mechanisms and paradigms that exist, or that could be developed, to accomplish such measurements. But this silence from most Canadian courts and administrative tribunals about measuring (and therefore potentially ensuring) decision readability levels may also be the result of deliberate choices not to measure readability – perhaps in recognition of Goodhart's Law, which warns of the dangers of allowing measurements to become performance targets.[72] Choices not to measure readability may also flow from a perceived lack of resources or time to devote to the task, in the face of competing priorities for the relevant courts and tribunals. Or these choices may flow from a sense that it is inappropriate, or a violation of judicial independence, to look behind (or to make suggestions to change) the ways that judges write their decisions.[73]

Regardless of the reasons for our current lack of information about how readably Canadian courts and administrative tribunals write their decisions, this dissertation identifies a fit-for-purpose quantitative readability formula, and some initial decision readability measurements that are calculated using this formula. Together, these two products of my dissertation expose how understandable different classes of adjudicative decisions are likely to be to their audiences, based

---

[72] Tamara Goriely, "The English Legal Aid White Paper and the LAG Conference" (1996) 3:3 International Journal of the Legal Profession 353 at 358-59:

> Quality systems need to capture inchoate qualitative concepts into quantitative data which can be collected and analysed. Thus they use proxies to measure underlying variables. For example, the measure of how long one is kept waiting in a casualty department is captured by the proxy of the proportion of people who are seen by a member of medical staff within five minutes. Before quality systems are introduced, the proxies may be excellent approximations of underlying quality. However, Goodhart's law suggests that the more one targets the proxy, by insisting that performance indicators are met, the weaker becomes the relationship between the proxy and the underlying variable. Thus hospitals may be employing 'hello' nurses to greet people on arrival in accident and emergency departments. The performance target is met, but only at the cost of withdrawing resources from staff involved in treatment: the queue to see the doctor may become longer.

[73] See Khan, *Life of a Reserve*, *supra* note 56 at 57-63 (for a discussion of how reluctant judges are to engage with questions about how they perform their work, perhaps because of their perceptions of what judicial independence requires of them) and especially at 63: "this [survey] data indicates that courts have expansive views about judicial independence: they believe that judicial independence precludes them from imposing structural or stylistic requirements upon judges."

on quantitative measurements of different linguistic features that the decisions contain. It will remain to be seen whether Canadian courts and administrative tribunals spontaneously and voluntarily choose to use this new readability formula. However, the products of my research are now available to those who study, critique, and have the power to legislate about, the practice of adjudicative decision-writing. In this sense, the dissertation has two things to offer to both the authors of adjudicative decisions, and to all others who have an interest in knowing the extent to which decisions are likely to be understood by different audiences. First, the dissertation identifies a (and perhaps the most) suitable tool for measuring the readability of Canadian decisions. And, second, the dissertation applies this tool to illustrate how understandable thousands of examples of decisions from the recent past are likely to be to audience members of different education levels.

This core research question that asks about the readability levels of Canadian court and administrative tribunal decisions cannot be answered in a single step. Answering this overarching question necessarily requires me to also explore and answer the following subsidiary research questions.

1.3.2 Subsidiary Question 1: How should readability be calculated?

As I explain in Chapter 2, many English readability formulas have been developed over the last 100 years, each with subtly different components, strengths, and weaknesses. Although the Flesch Reading Ease and Flesch-Kincaid Grade Level formulas are often relied upon in studies that measure the readability of legal texts,[74] and texts from other disciplines,[75] these formulas are now somewhat dated, do not necessarily reflect the current thinking in terms of our behavioural

[74] DeFriez, *Toward a Clearer Democracy*, *supra* note 29; Johnson, *supra* note 29, "Changing Discourse".
[75] George R Milne, Mary J Culnan & Henry Greene, "A Longitudinal Assessment of Online Privacy Notice Readability" (2006) 25:2 Journal of Public Policy & Marketing 238 (consumer privacy).

and theoretical understanding of the reading process, and do not perform as well as more contemporary formulas.[76]

However, as I have already suggested above, even more recent and sophisticated formulas that were derived from human-generated measurements of the readability of non-specialized texts may not be suitable for assessing the readability of texts from within a specialized domain: a recent study from the medical field illustrated that three general-purpose readability formulas (all producing readability measurements that were strongly correlated with one another) were not accurate in predicting the human-perceived reading difficulty of electronic medical record documents[77] that are "full of medical jargon, abbreviations, and other domain-specific usages and expressions that are ill-suited for the lay people (patients)."[78] As this research suggests, the unique language within specialized domains may present barriers to readability that are simply not well accounted for when one uses common readability formulas:

> This pattern of inconsistency [between general-purpose readability formulas and human-generated readability assessments] highlighted the inadequacy of these formulas' utility in measuring EHR [Electronic Health Record] readability. It also highlighted their weakness in testing readability of documents of complex topics such as medicine, as they were developed to help users in the education community to gauge text difficulty below 12 grade. All 3 formulas relied on word counts and sentence counts to estimate text readability. The implicit assumption that longer words were more difficult, however, could often be violated. For instance, abbreviations that were not normally used outside the medical domain, such as 'CHF' (Congestive Heart Failure) and 'EKG' (electrocardiogram), were prevalent in EHR notes, without full definitions. Because these short abbreviations often comprised very few, if any, syllables, they would have exactly the same impact on the readability score as did the common stop words such as 'the.' However, the abbreviations were obviously one of the barriers for a patient to understanding an EHR note.[79]

---

[76] Scott A Crossley, Stephen Skalicky & Mihai Dascalu, "Moving Beyond Classic Readability Formulas: New Methods and New Models" (2019) 42 Journal of Research in Reading 541 at 557 ["Moving Beyond Classic Readability Formulas"].
[77] *See generally*, Zheng & Yu, "Health Records Difficulty", *supra* note 30.
[78] Zheng & Yu, "Health Records Difficulty", *supra* note 30.
[79] Zheng & Yu, "Health Records Difficulty", *supra* note 30.

The weaknesses of general-purpose readability formulas in specialized domains like the healthcare field have prompted researchers to propose new approaches for deriving readability formulas in these fields. For instance, expert raters (such as "health literacy and clinical experts and a patient representative" in one study,[80] and raters with "an advanced degree, experience teaching literacy-based classes, expertise in medical discourse, and prior experience rating medical text samples" in another study)[81] have been used to generate human measurements of the readability of a body of specialized texts, so that the texts and their associated linguistic properties could subsequently be used to generate new quantitative readability formulas.

Legal English is somewhat similar to medical language, in that it is characterized by technical, archaic, Latin, and French terms,[82] and "frequently adopts ordinary words but gives them a specific legal meaning that laypersons might never guess."[83] And, Legal English tends to refer to past case names in the body of current documents as a short-hand way of importing legal concepts, modes of thinking, and conclusions from the past cases into the current documents – although not all lawyers are equally skilled in the art of deploying case names, and not all readers will derive the intended (or any uniform) meaning from a particular case name.[84] However, legal English also differs from both medical language and ordinary English in a number of other ways (e.g.: through formality, impersonality, the use of conditional phrases, parallel structures, etc.)[85]

---

[80] Sasikiran Kandula & Qing Zeng-Treitler, "Creating a Gold Standard for the Readability Measurement of Health Texts" [2008] American Medical Informatics Association Symposium Proceedings 353 at 354.
[81] 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" (2020) 13:4 Journal of Communication in Healthcare 344 at 347 ["Predicting the Readability of Physicians' Secure Messages"].
[82] Danet, "Language in the Legal Process", *supra* note 31 at 476-77.
[83] Danet, "Language in the Legal Process", *supra* note 31 at 469.
[84] Kris Franklin, "*See Erie*: Critical Study of Legal Authority" (2008) 31:1 University of Arkansas at Little Rock Law Review 109, especially at 113-116.
[85] David Griffin, "Considering Legal English: Consensus and Complications" (2024) 31:1 International Journal of Speech, Language and the Law 77, at 80-83; Danet, "Language in the Legal Process", *supra* note 31 at 477-81.

that might similarly make this form of language less comprehensible to lay readers.[86] For instance, as Azuelos-Atias has argued, legal language – including legal English – often resembles ordinary language at a superficial level.[87] In this sense, legal language will generally fail to provide readers with explicit cues that alert them of their needs to explore and decode law's specialized meanings for otherwise-ordinary terms.[88] All of these features of legal English make it frequently incomprehensible to the general public, in ways that might similarly confound general-purpose readability formulas that are designed to predict how well a member of the general public will understand a text.

As the above discussion indicates, a legal domain-specific approach would likely be the most appropriate way to assess the readability of Canadian court and administrative tribunal decisions. It is possible that a new formula, generated from a reference corpus of legal texts, might be needed. It is also possible, however, that one of the many existing general-purpose readability formulas might prove to be the most appropriate for measuring the readability of adjudicative decisions. Prior to this dissertation, no one had specifically tested how well general-purpose readability formulas perform on law-specific texts, or whether a new law-specific formula might outperform the existing general-purpose formulas. One of the major undertakings of this dissertation is therefore to identify the most appropriate tool for assessing the readability of judicial and administrative tribunal decisions (as explained in Chapter 3).

The above discussion has explicitly focused on readability tools that can be used on English texts, and on the peculiarities of legal English. As this English-specific discussion perhaps signals,

---

[86] Gerald Lebovits, Alifya V Curtin & Lisa Solomon, "Ethical Judicial Opinion Writing" (2008) 21 Geo J Leg Ethics 237 at 251.
[87] Azuelos-Atias, "On the Incoherence of Legal Language", *supra* note 27 at 46-47.
[88] Azuelos-Atias, "On the Incoherence of Legal Language", *supra* note 27 at 46-47.

readability formulas tend to be language-specific, and readability formulas have only been developed thus far for a small minority of the world's known languages.[89] Furthermore, to the extent that some readability formulas might conceivably be capable of being applied equally to different languages,[90] no research has established that any such readability formula actually performs equally across different languages. It seems improbable that such research could ever exist: languages are each unique, and the variables that might make one language easier or more difficult to understand will not necessarily have the same impact in another language, assuming that the variables exist across comparator languages.[91] It is probably for these reasons that scholars have largely abandoned the idea of developing universal or cross-language readability approaches.[92]

Canada is a largely bilingual country, in which courts and tribunals in different places can issue decisions in either, or both of, English and French. This dissertation only considers readability levels of English-language decisions. It would obviously be preferable to consider readability levels of all Canadian adjudicative decisions, regardless of whether they appear in English or French. After all, the legal force of the decisions is the same regardless of the official language in which the decisions are written, other factors being equal. My inability to consider readability levels of both Canada's English and French decisions is an unfortunate consequence of

---

[89] Jesse Saba Kirchner, Justin Nuger & Yi Zhang, "An Extensible Crosslinguistic Readability Framework" in *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora* (Singapore: ACL Anthology, 2009) 11 at 11 ["Crosslinguistic Readability Framework"].

[90] Consider, for instance, a readability formula that relies only on variables that can be measured in any language, like average word and/or sentence length variables.

[91] Consider, for instance, hypothetical variables related to the use of accent marks above particular letters, or the use of contractions, or the use of liaisons that are made between words when a language is spoken. Accent marks are used heavily in French, but not in English. Contractions are used heavily in less formal variations of English, but are not used the same way in French. And, liaisons are often used in spoken French, but are generally not used in spoken English.

[92] Kirchner, Nuger & Zhang, "Crosslinguistic Readability Framework", *supra* note 89 at 11.

the language-specific nature of readability formulas. There is no tool today, nor is there an obvious way to create a future tool, that can measure readability levels of French and English decisions using a common measurement that would allow for meaningful analysis and comparison of these decisions across the language divide. But I candidly admit that this dissertation is less complete, and less useful to the people of Canada, because of my inability to measure the readability of French-language decisions alongside English-language decisions. And I offer this admission here in order to transparently identify the boundary of my research question about how readability should be calculated: this question only strives to determine how the readability of English-language decisions from Canadian courts and tribunals should be calculated.

1.3.3 Subsidiary Question 2: How readable are court and administrative tribunal decisions in a variety of different contexts?

Once I identify a suitable approach for measuring the quantitative readability of legal texts (in Chapter 3), it becomes possible to assess readability levels for large numbers of court and administrative tribunal decisions. Since we currently have almost no data about these readability levels, and since the authors of these decisions seem to be concerned about ensuring that their decisions are understandable to their audiences, a logical next step involves computational measurement of the decision readability levels: only then can interested stakeholders (such as decision-writers, Chief Justices or Tribunal Chairpersons, legislators, etc.) really know whether they are satisfied with the readability levels at which decisions are being written.

However, since different governments, institutions, and even individual authors, might all have distinct readability targets for their decisions, or specific interests in only a subset of decisions, it would arguably be more useful to know how readable Canadian adjudicative decisions are in different contexts than to know how readable they are across an aggregated large sample that captures all court and tribunal types, and all legal subject areas. To answer this research

question, it will therefore be necessary to measure decision readability levels from randomly selected and representative samples of cases drawn from each level of court (inferior, superior, appeal, and apex); across different legal subject areas; and, between administrative tribunals and courts that are writing about the same legal subject areas. With insights about actual decision readability levels that are broken down more precisely along these lines – insights that are contained within Chapter 4 of this dissertation – it may be possible for administrators, Chief Justices, or other policy-makers in different places to accurately assess readability performance of the authors within their areas of influence, or to start thinking about whether they ought to set readability targets in the same way that the BCCRT and the SST have set such targets.

It is also helpful to think about decision readability score differences between different judges, jurisdictions, opinion types, and other factors – which can be done through a targeted selection of cases for study, even if this approach does not allow for statistical inferences to be drawn about cases beyond the ones that are studied within the targeted sample. This approach would facilitate the creation of a large set of readability results that could be user-searched. In other words, it would allow judges and tribunal members whose decisions appear within the sample to see their own readability scores; it would allow other researchers to explore readability scores for subsets of cases (based on, for instance, the gender of a decision's author or the province from which the decision emerged, among other variables); it would allow future decision-makers to see concrete examples of past decisions that are highly readable, or highly unreadable, for different legal subject areas; and, it would allow any interested individual to explore readability levels for different decisions within the larger database of Canadian adjudicative decisions. This initiative – the creation of a personalized database of readability scores for a large sample of Canadian decisions that is available to the public for individual search queries and analysis – is the

focus of Chapter 6. Together, these studies contained in Chapters 4 and 6 shed light (in a statistically robust way within Chapter 4, and in a more individual way in Chapter 6) on the broad topic of decision readability levels in a variety of different contexts.

1.3.4 Subsidiary Question 3: How readable are court decisions in contexts where we might most expect them to be readable?

Although many judges acknowledge that the general public consisting of lay people who are not trained in law form part of the audience for judicial decisions as a general matter, some judges consider their audiences in a more nuanced way, and suggest that audiences may predictably be wider or narrower depending on the circumstances of the case. For instance, Justice Wilson notes that, "[i]f the case is largely factual, or the hearing is an interlocutory application, the audience may not be much broader than the parties."[93] And, as the SCC has noted on several occasions, the requirement to give reasons is closely linked to the need for the parties to understand the basis for a court's decision.[94] In *R v Sheppard*, the Court noted that "the requirement of reasons is tied to their purpose and the purpose varies with the context. At the trial level, the reasons justify and explain the result. The losing party knows why he or she has lost. Informed consideration can be given to grounds for appeal."[95] In other words, a court or tribunal's decision is functionally necessary for the parties – both to understand the final disposition of their case (if there is no possibility of or need to appeal the decision), but also to understand whether they might have grounds to appeal in certain circumstances.

---

[93] Canada, Department of Justice. *The Honourable Steven Wilson's Questionnaire (Questionnaire for the British Columbia Supreme Court Judicial Appointment Process)* (1 Mar 2017), online: <https://www.canada.ca/en/department-justice/news/2019/02/the-honourable-steven-wilsons-questionnaire.html>.
[94] *See, e.g.*, *Baker v Canada (Minister of Citizenship and Immigration)*, [1999] 2 SCR 817, at para 39; *and see R v Sheppard*, 2002 SCC 26, at para 15: "Reasons for judgment are the primary mechanism by which judges account to the parties and to the public for the decisions they render. […]. Trial courts, where the essential findings of facts and drawing of inferences are done, can only be held properly to account if the reasons for their adjudication are transparent and accessible to the public and to the appellate courts."
[95] *R v Sheppard*, 2002 SCC 26, at para 24.

Logically, where the audience may also include others beyond just the parties to a dispute, one might expect the readability level of a decision to vary depending on the author's perceived audience. As Justice Brundrett notes, "[t]he style of legal writing can thus be influenced by the nature of the expected audience. In many cases, this means that legal writing should be understandable to not just legal actors but the public at large. To some extent, the audience will vary depending upon whether the decision is a chambers judgment, a preliminary ruling, or a full opinion on a final verdict. Styles vary."[96] The implication of these comments is that, in at least some cases, there is less need to ensure that the decision is capable of being understood by members of the general public, since the nature of the decision is such that no member of the public is likely to be interested in the decision.

Judges seem to agree, however, that under certain other circumstances, the need to produce decisions that are understandable to the public, or to individuals with less legal sophistication, will be much higher. The public appears to be a key audience in criminal cases,[97] which may necessitate

---

[96] Canada, Department of Justice. *The Honourable Michael Brundrett's Questionnaire (Questionnaire for the British Columbia Supreme Court Judicial Appointment Process)* (2 Nov 2017), online: <https://www.canada.ca/en/department-justice/news/2017/11/the_honourable_michaeljbrundrettsquestionnaire.html>.

[97] Canada, Department of Justice. *The Honourable Warren B Milman's Questionnaire (Questionnaire for the British Columbia Supreme Court Judicial Appointment Process)* (3 Nov 2017), online: <https://www.canada.ca/en/department-justice/news/2017/11/the_honourable_warrenbmilmansquestionnaire.html>: "The public has, in that sense, an interest in the disposition of every case that comes before the courts. This is especially true in criminal cases, where the public is also, in effect, the initiating litigant. The denunciatory element in criminal judgments is, by definition, directed specifically at the public". See also, Canada, Department of Justice. *The Honourable Cynthia Petersen's Questionnaire (Questionnaire for the Ontario Superior Court of Justice Judicial Appointment Process)* (last modified on 30 May 2022), online: <https://www.canada.ca/en/department-justice/news/2022/05/the-honourable-cynthia-petersens-questionnaire.html>: "In matters of criminal law, the main audience of the court's decision typically consists of the accused and the victim (and/or victim's family), though in many cases the public at large will also have a substantial interest in either the outcome and/or reasoning of the court, and may therefore also be part of the primary audience." And see, Canada, Department of Justice. *The Honourable Anne Louise Kirker's Questionnaire (Questionnaire for the Alberta Court of Queen's Bench Judicial Appointment Process)* (13 Nov 2018), online: <https://www.canada.ca/en/department-justice/news/2018/11/the-honourable-anne-louise-kirkers-questionnaire.html>:

> The reasons for decisions in criminal cases must be understood by the accused and those affected by the crime, but are often of interest to the media and the public at large. In this area of law, perhaps more than in others, court proceedings and decisions are the subject of media reports and social media commentary, with

the writing of more readable decisions. Additionally, in cases involving unrepresented or self-represented litigants, judges perceive a need to produce more readable decisions.[98] Research on the prevalence of self-represented litigants suggests that the phenomenon of self-representation is particularly common in cases involving family law (where estimates now suggest that self-represented litigants make up 80% of a court's case load),[99] and before administrative tribunals (where the estimated national baseline for self-representation is 62%),[100] so one might expect decisions to be more readable in these self-representation contexts. Finally, since judges generally acknowledge that their decisions need to address or be written for the parties,[101] it would be useful

---

> people sharing their thoughts and perspectives about the court process and outcomes. With social media revolutionizing the way people communicate and understand the world around them comes an increasingly wide audience of people from all walks of life attempting to comprehend what the courts are doing in some cases […].

[98] Canada, Department of Justice. *The Honourable Michele Hollins' Questionnaire (Questionnaire for the Alberta Court of Queen's Bench Judicial Appointment Process)* (31 Mar 2017), online: <https://www.canada.ca/en/department-justice/news/2017/03/the_honourable_justicemichelehhollinssquestionnaire.html>: "As we continue to see increasing numbers of unrepresented litigants, the public is also an audience for judicial decisions and a compelling reason to make written decisions clear and concise." See also, Canada, Department of Justice. *The Honourable Nicholas Kasirer's Questionnaire (Questionnaire for the Supreme Court of Canada Judicial Appointment Process)* (18 April 2019) (online: <https://www.fja.gc.ca/scc-csc/2019/nominee-candidat-eng.html>): "I support any measure intended to increase the parties' understanding of the reasons, and I note recent initiatives taken by the Supreme Court to limit the risks that its decisions will be misunderstood, especially given the increase in the number of self-represented litigants."

[99] Julie Macfarlane, *The National Self-Represented Litigants Project: Identifying and Meeting the Needs of Self-Represented Litigants – Final Report* (May 2013), online: <https://representingyourselfcanada.com/wp-content/uploads/2016/09/srlreportfinal.pdf>, at 15.

[100] Lilian Ma, et al, *A National Survey of Tribunal Responsiveness to Self-Represented Parties: Measuring Access to Justice for Canadian Administrative Tribunals* (Council of Canadian Administrative Tribunals, 2015), online: <https://www.ccat-ctac.org/wp-content/uploads/2021/11/AJCpaper-FinalPOSTAugust112015.pdf>, at 6-7.

[101] See, e.g., Canada, Department of Justice. *The Honourable Sheilah Martin's Questionnaire (Questionnaire for the Supreme Court of Canada Judicial Appointment Process)* (21 December 2021), at Question 4 ("Who is the audience for Supreme Court of Canada decisions?"). Online: <https://www.fja.gc.ca/scc-csc/2017-SheilahMartin/nominee-candidat-eng.html>: "**Courts must always speak directly to the parties**, who have a personal interest in the outcome of the dispute because their liberty, rights, property, reputation, business and/or family are at stake. Parties merit special attention because they have also shouldered the emotional and financial burdens of moving their cases through the court system" (emphasis added). Although these comments were offered by a candidate for a vacancy at the Supreme Court of Canada (SCC), Justice Martin had been a trial court judge from 2005-2016, and was an appeal court judge from 2016 until she was appointed to the SCC in 2018 (Supreme Court of Canada, "The Honourable Sheilah L. Martin" (15 May 2018), online: <https://www.scc-csc.ca/judges-juges/bio-eng.aspx?id=sheilah-l-martin>). We might infer, therefore, that Justice Martin's conception of the audience for court decisions (specifically that courts must always speak directly to the parties) was informed by the many years she had spent judging at different courts prior to the moment when she expressed her thoughts on this question. *See also* the comments of the Honourable John C

to know how readable decisions are in certain specialized courts where one might expect the parties to be less prepared to understand the decisions, such as in youth criminal courts. In Chapter 5, using the law-specific readability formula that I present in Chapter 3, I provide data and analysis that answers this research question about decision readability levels in two specific contexts where we expect decisions to be more readable: in courts dealing with young persons under the *Youth Criminal Justice Act*,[102] and in court cases where at least one party is self-represented.

1.4 The Interdisciplinary Law and Language Research Methodology

In order to answer my research questions effectively, I am guided by an overarching research methodology that is relevant to problems that lie at the intersection of law and language. My dissertation therefore employs an interdisciplinary law and language methodology,[103] and relies most heavily on what Mertz & Rajah call the "social scientific approach to legal language research" as opposed to the "more humanist and critical approaches."[104] This type of approach tends, methodologically, to rely "on big data empiricism rather than doctrinal analysis; substantially, it treats language not merely as the medium but as the object of study."[105] The benefit of this approach is that it allows for more and easier systemic analyses, as opposed to isolated

---

Bouck, who was a judge of the Supreme Court of British Columbia at the time that he offered this comments (John C Bouck, "Civil Jury Trials – Assessing Non-Pecuniary Damages – Civil Jury Reform" (2002) 81:3 Canadian Bar Review 493 at 519): "Most trial judges write for the parties and not for some judicial data bank." In the context of an administrative tribunal, *see City of Medicine Hat v Medicine Hat Fire Fighters Association*, IAFF Local 263, 2014 CanLII 50000 (AB GAA): "reasons are primarily written for the parties involved; to explain for the particular dispute, how the board considered the issues the parties chose to put forward which, it is reasonable to assume, were of particular interest and importance to the communities and employee groups represented by them."

[102] *Youth Criminal Justice Act*, SC 2002, c 1, s 2(1), sv: "young person", provides that a young person is someone who appears "to be twelve years old or older, but less than eighteen years old."

[103] See, generally, Elizabeth Mertz & Jothie Rajah, "Language-and-Law Scholarship: An Interdisciplinary Conversation and a Post-9/11 Example" (2014) 10 Annual Review of Law and Social Science 169. *See also*, Janet E Ainsworth, "Linguistics as a Knowledge Domain in the Law" (2006) 54 Drake Law Review 651 ["Language-and-Law Scholarship"].

[104] Mertz & Rajah, "Language-and-Law Scholarship", *supra* note 103 at 170.

[105] Hanjo Hamann & Friedemann Vogel, "Evidence-Based Jurisprudence Meets Legal Linguistics: Unlikely Blends Made in Germany" [2017] Brigham Young University Law Review 1473 at 1474 ["Evidence-Based Jurisprudence"].

findings. However, the "challenge for researchers who use law-as-data techniques is to strike a balance between rich, high-dimension representation of legal texts (which can be impractical to analyse) and low-resolution, simplified representations that are tractable but incomplete."[106] My intent is to strike an appropriate balance by sampling widely from the population of adjudicative decisions across Canada in ways that allow for useful statistical generalizations about the full population of cases, without gathering unworkably large amounts of data. I also attempt to add helpful context or 'resolution' to the large-scale representations of decision readability levels that are contained within this dissertation by singling out individual decisions (or extracts from these decisions) for further qualitative analysis where appropriate. In this respect, I rely on qualitative forms of analysis in selective circumstances to offer complementary perspectives that run in parallel alongside the quantitative perspectives that are generally foregrounded within this dissertation.

A particularly practical variant of law and language research that falls within this social scientific approach that I use involves research that takes place at the crossroads of law and linguistics: "[s]tudies looking at linguistic issues in legal settings, especially when viewed as a body of work, […] have a range of potentially impactful applications beyond academia, across policymaking, professional legal practice, social inclusion and the creation and interpretation of law."[107] Law and linguistics research has been described as "a specialized variety of communication studies. Legal linguists explore the relationship of legal practice and language as its own end, studying terminology and grammar, speech theories, semiotics and interpretation

---

[106] Michael A Livermore & Bao Kham Chau, "Studying Judicial Behaviour with Text Analysis" in Lee Epstein et al (eds), *The Oxford Handbook of Comparative Judicial Behaviour* (Oxford: Oxford UP, 2024) 305 ["Studying Judicial Behaviour"].
[107] Alexandra Grey & Laura Smith-Khan, "Bringing Linguistic Research into Legal Scholarship and Practice" (2021) 46 Alternative Law Journal 64 at 65 ["Bringing Linguistic Research into Legal Scholarship"].

methods of law, discourse structures in courts, creation of statutes, or genesis of legal norms in general, for instance."[108]

My research, inspired by the kinds of empirical readability studies that have been undertaken by applied linguistics scholars in the context of general readability,[109] and readability studies in specialized domains,[110] fits within this sub-category of law and linguistics scholarship that falls under the law and language methodological umbrella.[111] Furthermore, like much of the law and linguistics scholarship that has come before this dissertation, my research includes considerations about audience[112] – and specifically, the audience for judicial decisions – within the relevant discussions of readability. It also relies heavily on empirical data (as law and linguistics scholarship often does)[113] – about measured readability levels that are calculated from linguistic information that is extracted from the texts of actual adjudicative decisions – in order to both identify and then employ a law-specific readability measurement.

For the purposes of answering my research questions, this methodological approach is ideal. My objective is to assess how readable Canadian court and administrative tribunal decisions are in the most accurate way that is reasonably possible. Using a law and linguistics approach –

---

[108] Hamann & Vogel, "Evidence-Based Jurisprudence", *supra* note 105 at 1491.
[109] See, e.g., Crossley, Skalicky & Dascalu, "Moving Beyond Classic Readability Formulas", *supra* note 76.
[110] See, e.g., Crossley, et al, "Predicting the Readability of Physicians' Secure Messages", *supra* note 81 (wherein the authors quantitatively studied 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" (2020) 57:1 International Journal of Business Communication 52 (wherein the authors relied upon measures of lexical density, subordinate clause use, and passive voice use, alongside other classic readability formulas, to assess the readability of private-sector business reports).
[111] Grey & Smith-Khan, "Bringing Linguistic Research into Legal Scholarship", *supra* note 107 at 65.
[111] Hamann & Vogel, "Evidence-Based Jurisprudence" *supra* note 105 at 1487: "Legal Linguistics is a theoretical and empirical research approach in the wider field of 'law and language' research."
[112] William D Popkin, "Law and Linguistics: Is There Common Ground" (1995) 73 Wash University Law Quarterly 1043 at 1043.
[113] Popkin, "Law and Linguistics: Is There Common Ground", *supra* note 112 at 1044.

one that "adopts an observational big data strategy by preferring cumulated research over salient yet eclectic findings"[114] – allows me to see the big picture in terms of readability, rather than just the picture that is created by memorable outliers that tend to attract attention.[115] This approach transcends the limitations of many more traditional legal research methods that involve close readings of individual decisions. As Livermore & Chau have noted,

> [e]ven with an army of capable assistants, the amount of data that can be generated in this way is quite limited. Analyses of 'small data' can lead to useful insights, but researchers can now move beyond conventional methods of data collection and analysis. In recent decades, the digitalization of legal texts has enabled empirical legal scholars to use computational text analysis techniques to study large digital corpora of legal documents. These computer-based tools can help researchers study long-standing questions and also open new, previously inaccessible epistemic doors.[116]

My intent is to employ these kinds of large-scale data analysis techniques to expose new knowledge about linguistic complexity levels within Canadian court and tribunal decisions, including knowledge about both how we ought to measure readability levels of adjudicative decisions, and what these readability levels look like for contemporary decisions.

Additionally, in order to identify a law-specific readability formula that is capable of answering my research questions, my research draws upon large corpora of texts in ways that would be difficult if not impossible to accomplish without the assistance of computers. In this sense, my research in this dissertation aligns well with a specific type of law and linguistics approach that Hamann & Vogel call "Computer Assisted Legal Linguistics (CAL^2)."[117] A CAL^2 methodology "focusses on specialized legal vocabularies rather than ordinary language, which was

---

[114] Hamann & Vogel, "Evidence-Based Jurisprudence", *supra* note 105 at 1487.
[115] See, e.g., Mitch Kowalski, "When Judges Make Readable Decisions" Financial Post (5 Apr 2013), online: <https://financialpost.com/legal-post/when-judges-make-readable-decisions>, commenting upon the memorably sarcastic reasons given by the judge in *R v Duncan*, 2013 ONCJ 160.
[116] Livermore & Chau, "Studying Judicial Behaviour", *supra* note 106 at 305.
[117] Hamann & Vogel, "Evidence-Based Jurisprudence", *supra* note 105 at 1495.

the focus for much previous U.S. research in the plain meaning and original intent traditions. […]. CAL^2 analyzes what judges and other jurists write down and how they thus exert power and authority."[118] The advantage of this methodological approach is that it facilitates the "analyzing [of] legal texts, no longer by the handful, but in thousands and millions."[119]

In other senses, employing a law and linguistics methodology allows me to implicitly, or indirectly, suggest law (or judicial/tribunal practice) reforms where such reforms prove to be warranted by my research (as I describe in Chapter 7). As Grey and Smith-Khan note, much law and linguistics scholarship "addresses unequal or unjust experiences and outcomes caused or aggravated by a language issue" within the legal system,[120] like the experiences that a reader of case law might encounter when struggling through a decision that is written at a level that is beyond his or her linguistic capability. My initial hypothesis that many Canadian judicial and administrative tribunal decisions are written less readably than would be necessary to be understood by many of the potential audiences for these decisions, which was a major element of my motivation for undertaking this research into the readability of court and tribunal decisions, is validated within Chapters 4, 5, and 6 of this dissertation. Consequently, my dissertation seems to fit within the class of law and language scholarship that Grey and Smith-Khan call "research motivated by problems of 'linguistic justice' in the exercise of state power."[121]

1.5 Literature Review

[118] Hamann & Vogel, "Evidence-Based Jurisprudence", *supra* note 105 at 1496.
[119] Hamann & Vogel, "Evidence-Based Jurisprudence", *supra* note 105 at 1494.
[120] Grey & Smith-Khan, "Bringing Linguistic Research into Legal Scholarship", *supra* note 107 at 65.
[121] Grey & Smith-Khan, "Bringing Linguistic Research into Legal Scholarship", *supra* note 107 at 66.

There are massive numbers of studies that apply readability formulas within a range of different contexts.[122] Much of the early research took place within the fields of psychology and education,[123] but interest in readability has subsequently expanded to new fields. In recent years, there have been many readability studies relating to public health and pandemic information.[124] Many other studies have assessed the readability of financial disclosure and other business statements,[125] including a study that linked tax avoidance (as measured by relatively lower rates of tax payable by businesses) with the unreadability of business financial statements.[126] Readability has also been assessed within many other vastly divergent domains, ranging from kinesiology,[127] to farming,[128] to cyber-security,[129] to maritime safety.[130] Elsewhere, researchers

---

[122] A simple search using the Google Scholar platform for the term "Flesch-Kincaid" (a well-known readability formula) in documents published between 1985-2023 produces approximately 17,800 results, which equates to approximately 468 documents per year. A cursory review of the first 100 results revealed that they related to readability theory and measurements, and to the readability of foreign language material, web pages, health care information, criminal justice information, grade-school reading material, Twitter posts, and informed consent documents for research studies – among other things.

[123] Walter Kintsch & Douglas Vipond, "Reading Comprehension and Readability in Educational Practice and Psychological Theory" in Lars-Goran Nilsson (ed), *Perspectives on Memory Research* (London: Psychology Press, 1979) 329 at 333.

[124] See, for instance, Corey H Basch et al, "Public Health Communication in Time of Crisis: Readability of On-Line COVID-19 Information" (2020) 14:5 Disaster Medicine and Public Health Preparedness 635; Tanya Serry et al, "Improving Access to COVID-19 Information by Ensuring the Readability of Government Websites" (2022) 34:2 Health Promotion Journal of Australia 595; and, Jessica Kruse et al, "Readability, Content, and Quality of COVID-19 Patient Education Materials from Academic Medical Centers in the United States" (2021) 49 American Journal of Infection Control 690, among thousands of other examples.

[125] See, for instance, Daniel Tschopp, Doug Barney & Passard C Dean, "The Readability of Financial Statement Note Disclosures" (2018) 12:1 International Journal of Business, Accounting, and Finance 63; and, Sinethemba Mankayi, Frank Ranganai Matenda & Mabutho Sibanda, "An Analysis of the Relationship Between Readability of the Chairman's Statement and Firm Financial Performance in South Africa" (2023) 12:8 International Journal of Research in Business & Social Science 205.

[126] Justin Hung Nguyen, "Tax Avoidance and Financial Statement Readability" (2021) 30:5 European Accounting Review 1043.

[127] Jafrā D Thomas, Brian R Flay & Bradley J Cardinal, "Are Physical Activity Resources Understandable as Disseminated? A Meta-Analysis of Readability Studies" (2018) 70:4 Quest 492.

[128] Craig D Morris & Christine Stilwell, "Getting the Write Message Right: Review of Guidelines for Producing Readable Print Agricultural Information Materials" (2003) 69:1 South African Journal of Libraries and Information Science 71.

[129] Tim Smit, Max van Haastrecht & Marco Spruit, "The Effect of Countermeasure Readability on Security Intentions" (2021) 1 Journal of Cybersecurity & Privacy 675.

[130] Floris Goerlandt & Huiyan Liu, "Readability of Maritime Accident Reports: A Comparative Analysis" (2023) Maritime Policy & Management 1, online: <https://doi.org/10.1080/03088839.2023.2166685>.

have studied the readability of media products,[131] political speeches and debates,[132] and literary fiction.[133] Interest in quantitative readability studies (like most of the studies mentioned above, which rely on at least one readability formula) is both widespread across disciplines, and enduring over the years.

My dissertation fits within, but adds multiple novel insights into, a small but well-established body of scholarship that looks in a more focused way at the problem of how sources of law and other legal instruments are understood by their audiences. In one recent study, Benoliel & Becher found that the average Flesch Reading Ease (an older readability formula that was created in 1948) score for the 'click to accept' online contracts within the study was 34.86, "comparable to the score of articles found in academic journals, which typically do not target the general public."[134] The authors also found that 99.6% of the contracts were written below (i.e.: less readably than) the recommended Flesch Reading Ease Score of 60.[135] While the results of this study are concerning (because it is difficult to imagine how a consumer could accept – within the meaning of contract law – a contract that they do not and cannot understand),[136] they are perhaps

---

[131] Jessica F Sparks & Jay D Hmielowski, "At the Extremes: Assessing Readability, Grade Level, Sentiment, and Tone in US Media Outlets" (2023) 24:1 Journalism Studies 24.
[132] Orly Kayam, "The Readability and Simplicity of Donald Trump's Language" (2018) 16:1 Political Studies Review 73.
[133] Yuri Bizzoni et al, "Good Reads and Easy Novels: Readability and Literary Quality in a Corpus of US-published Fiction" in Tanel Alumäe & Mark Fishel, eds, *Proceedings of the 24th Nordic Conference on Computational Linguistics* (Estonia: University of Tartu Library, 2023) 42, online: <https://aclanthology.org/2023.nodalida-1.5.pdf>.
[134] Uri Benoliel & Shmuel I Becher, "The Duty to Read the Unreadable" (2019) 60:8 Boston College Law Review 2255 at 2277-2278.
[135] Benoliel & Becher, "The Duty to Read the Unreadable", *supra* note 134 at 2278.
[136] *See South West Terminal Ltd v Achter Land*, 2023 SKKB 116, at para 18, for a summary of the "*consensus ad idem* / meeting of the minds" principle in Canadian contract law. This principle requires that two parties who enter into a contract must have reached a meeting of the minds, as exemplified through an offer and acceptance of that offer, in order for the contract to be valid. *See Bertolo v Bank of Montreal*, 1986 CanLII 150 (ON CA) – a case involving an elderly woman "of meagre means who has had little schooling and is not fluent in the English language" – for an example of an instance where a court held a contract to be invalid because one of the parties could not possibly have understood what the contract meant. See *Terra Marine Industries Ltd v N & N Electric Motor Specialists Ltd*, 1991 CanLII 13663 (NS SC), for a similar situation.

undermined to a certain extent by a related study that also used the Flesch Reading Ease formula in the context of contract law. Schmitz & Pavillon used excerpts from the terms and conditions for an online video gaming contract that were scored for readability and then assessed against human participants' understanding of the effects of these excerpts.[137] The authors of this study found that "there was no correlation between an excerpt's readability score and the number of correct answers [about what the terms mean], providing preliminary evidence that the Flesch Reading Ease readability formula may not be a good indicator of comprehensibility of gaming terms and conditions."[138] Although these studies seem to reach opposite conclusions about the likelihood of consumers understanding contracts based on the language of the contracts, the conflict is perhaps more attributable to the fallibility of the particular readability formula that was used in the two studies than to any real disagreement about the premise that contracts with difficult language are harder for people to understand.

There has also been a significant amount of commentary on "plain language" in law, especially with respect to legislation,[139] but also in other domains.[140] Professor Blaisie noted in a recent American article that there are now more than 800 codified plain language laws that require

---

[137] Benedikt Schmitz & Charlotte Pavillon, "Measuring Transparency in Consumer Contracts: The Usefulness of Readability Formulas Empirically Assessed" (2020) 9:5 Journal of European Consumer and Market Law 191 ["Measuring Transparency"].
[138] Schmitz & Pavillon, "Measuring Transparency", *supra* note 137 at 199.
[139] See, e.g., Sullivan, "The Promise of Plain Language Drafting", *supra* note 26; David St L Kelly, "Legislative Drafting and Plain English", *supra* note 26; I Turnbull, "Legislative Drafting in Plain Language and Statements of General Principle" (1997) 18:1 Statute Law Review Review 21 (United Kingdom); Jeffrey Barnes, "When Plain Language Legislation is Ambiguous – Sources of Doubt and Lessons for the Plain Language Movement" (2010) 34:3 Melbourne University Law Review 671 (Australia).
[140] See, e.g., Robert P Charrow & Veda R Charrow, "Making Legal Language Understandable: A Psycholinguistic Study of Jury Instructions" (1979) 79:7 Columbia Law Review 1306; Charles R Dyer, et al, "Improving Access to Justice: Plain Language Family Law Court Forms in Washington State" (2013) 11:3 Seattle Journal of Social Justice 1065; Robert W Benson & Joan B Kessler, "Legalese v. Plain English: An Empirical Study of Persuasion and Credibility in Appellate Brief Writing" (1987) 20:2 Loyola of Los Angeles Law Review 301; and, Maria Mindlin, "Is Plain Language Better – A Comparative Readability Study of Court Forms" (2006) 10 Scribes Journal of Legal Writing 55.

the authors of written products to produce understandable material in different contexts.[141] And, it has been noted that readability formulas are often used as yardsticks for assessing plain language results: "several states have also utilized the [Flesch-Kincaid Grade Level formula] in statutes aiming to ensure the readability of specific documents concerning, for example, production contracts, credit life insurance, and health benefit plans."[142] In spite of the growing interest in plain language laws and readable legal instruments, some commentary is still critical of the ways in which readability and plain language concepts are used in law.[143]

With respect to studies of judicial decisions, some have looked qualitatively at the distinctive linguistic characteristics of separate dissenting opinions,[144] at the abnormally high use of emphatic adverbs within judicial decisions[145] (that one might expect to be more dispassionate forms of communication),[146] and at the socio-political message that might be associated with a use of plain language by courts in their decisions.[147] One qualitative study has considered the use of plain language interpretation techniques by the Supreme Court of the United States (SCOTUS) during the tenure of Chief Justice Roberts,[148] although this study seemed to be more about the use

---

[141] Michael A Blaisie, "Regulating Plain Language" [2023] Wisconsin Law Review 687 at 721.
[142] Benoliel & Becher, "The Duty to Read the Unreadable", *supra* note 134 at 2276.
[143] Louis J Sirico Jr, "Readability Studies: How Technocentrism Can Compromise Research and Legal Determinations" (2007) 26:1 Quinnipiac Law Review 147.
[144] Stanisław Goźdź-Roszkowski, "Communicating Dissent in Judicial Opinions: A Comparative, Genre-Based Analysis" (2020) 33 International Journal of for the Semiotics of Law 381.
[145] Edward Finegan & Benjamin T Lee, "Legal Writing: Attitude and Emphasis – Corpus Linguistic Approaches to 'Legal Language': Adverbial Expression of Attitude and Emphasis in Supreme Court Opinions" in Malcolm Coulthard, Alison May & Rui Sousa-Silva, eds, *The Routledge Handbook of Forensic Linguistics*, 2nd ed (New York: Routledge, 2020) 48.
[146] Posner, "Judges' Writing Styles (And Do They Matter?)", *supra* note 67 at 1429, describes the "pure" style of judicial opinion as one that "uses technical legal terms without translation into everyday English, quotes heavily from previous judicial opinions, includes much detail concerning names, times, and places, complies scrupulously with whatever are the current conventions of citation form, avoids any note of levity, [and] conceals the author's personality." Posner (at 1432) suggests that the vast majority of judicial decisions adopt this "pure" and impersonal style.
[147] Iris van Domselaar, "'Plain' Legal Language by Courts: Mere Clarity, An Expression of Civic Friendship or a Masquerade of Violence?" (2022) 10 Theory & Practice of Legislation 93.
[148] David A Strauss, "The Plain Language Court" (2016) 38:2 Cardozo Law Review 651.

of textual, rather than purposivist, approaches to the interpretation of statutes than about "plain language" as the concept is understood by legislative drafters.

Varsava and Craig have separately considered, qualitatively, how judges who place excessive emphasis on linguistic style, and on the production of aesthetically pleasing or memorable judicial decisions, may reflect a measure of disrespect for litigants – by prioritizing the judge's personal voice over the more important institutional voice of the court.[149] These authors make powerful arguments (that will be revisited later in this dissertation) about how judges' decisions ought to be functional documents that "uphold the rule of law, although that might mean passing up opportunities to display their literary skills; to entertain, amuse, and intrigue; and even to persuade their readers."[150] Renowned (now former-) Judge Richard Posner has observed that "[s]ome judicial opinions – those written by the masters – would lose something, and maybe a lot, in being paraphrased" or stripped of their individual styles.[151] But Posner goes on to acknowledge that "their essential meaning would not be lost. Even the best, the most distinctive, the most eloquent judicial opinion could be rewritten in a very different style and yet convey enough of the meaning of the original to be considered a close substitute for it."[152] If Posner is correct in this latter assessment, then Varsava and Craig raise valid and important concerns about judicial decisions that seem to place a priority on expressions of the authors' personal voices: should judges strive to write memorable, or eloquent, or literary decisions? What if these qualities prevent some

[149] See, e.g., Nina Varsava, "Professional Irresponsibility and Judicial Opinions" (2021) 59:1 Houston Law Review 103. *See also*, Elaine Craig, "Judicial Audiences: A Case Study of Justice David Watt's Literary Judgments" (2018) 64:2 McGill Law Journal 309.
[150] Varsava, "Professional Irresponsibility and Judicial Opinions", *supra* note 149 at 173.
[151] Posner, "Judges' Writing Styles (And Do They Matter?)", *supra* note 67 at 1423.
[152] Posner, "Judges' Writing Styles (And Do They Matter?)", *supra* note 67 at 1423.

audiences from understanding the decisions? These are questions that I consider in more detail within Chapter 7.

Other quantitative research has examined the writing styles of famous judges to isolate or identify any linguistic characteristics of their writing that distinguish these judges from their peers.[153] One doctoral student 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.[154] Johnson measured both the length and readability levels (again, using both the Flesch Reading Ease and Flesch-Kincaid Grade Level formulas) of SCOTUS opinions during the periods between 1931-1933 and 2009-2011, and found that the more recent decisions were longer and less readable.[155] Another recent study considered, among other things, whether the SCOTUS tended to copy language directly from lower court decisions if the lower court decisions were written more readably, and found that copying was more likely to have occurred from within decisions that avoided the passive voice, but that Flesch Reading Ease scores had no significant impact on copying.[156] Whalen's large-scale quantitative study measured the readability of 6,206 SCOTUS decisions released since 1946,[157] using the Simple Measure of Gobbledygook (SMOG)[158] readability scores for these decisions, and found that the court's decisions are becoming less readable over time, that judges tend to produce

---

[153] *See, e.g.*, Brady Coleman, "Lord Denning & Justice Cardozo: The Judge as Poet-Philosopher" (2001) 32:2 Rutgers Law Journal 485. See also, Nina Varsava, "Elements of Judicial Style: A Quantitative Guide to Neil Gorsuch's Opinion Writing" (2018) 93 New York University Law Review Online 75.
[154] DeFriez, *Toward a Clearer Democracy*, *supra* note 29.
[155] Johnson, "The Changing Discourse of the Supreme Court", *supra* note 26.
[156] Jennifer Bowie & Elisha Carol Savchak, "State Court Influence on US Supreme Court Opinions" (2022) 10:1 Journal of Law & Courts 139.
[157] Whalen, "Judicial Gobbledygook", *supra* note 8.
[158] SMOG is a quantitative readability formula that is calculated based on the number of three- (or more) syllable words within a 30-sentence sample, first introduced in 1969 (*see* G Harry McLaughlin, "SMOG Grading – A New Readability Formula" (1969) 12 Journal of Reading 639).

less readable opinions as their tenure on the court increases, and that opinions written by conservative judges tend to be slightly less readable than those written by liberal judges.

As the above review of the academic literature relating to the readability of legal texts indicates, there has not yet been any comprehensive study of the readability of Canadian administrative tribunal or court decisions. Furthermore, to the extent that readability studies have been undertaken in respect of non-Canadian decisions, these studies have (1) tended to focus on a narrow range of jurisdictions or levels of court (such as a single state's courts, or the SCOTUS); (2) used older, more primitive readability formulas as their basis of measurement – predominantly the Flesch formulas; (3) not considered how the readability of decisions may vary across legal subject areas and court levels, or between courts and administrative tribunals dealing with the same types of cases; and, (4) not considered how readability levels might be higher or lower in contexts where parties to a case are particularly at risk of not understanding a court's decision (because they are young persons, or are self-represented).

With this picture of the literature landscape in mind, one can begin to see how and why my dissertation fills a void in the existing scholarship. This project does not just use off-the-shelf tools for calculating the readability of legal texts, on the assumption that such tools will work effectively on law-specific texts. Instead, it starts by asking, what tool will work best to achieve the goal of quantifying the readability of our adjudicative decisions? Only after this question is answered does the dissertation proceed further to quantitatively measure readability levels. And the readability measurements, with accompanying analysis, that this dissertation contains are novel in a number of respects: they are the product of studies that are Canada-specific, across both courts and administrative tribunals, from recent (2022) samples of adjudicative decisions, that are coded for interesting and insightful breakdowns based on a number of relevant variables. In short, this

dissertation does what its title suggests that it will do: it puts numbers to words, and allows us all to see for the first time, from multiple perspectives, the readability levels of Canadian adjudicative decisions.

1.6 Access to Justice and the Connection to Readability

My discussion (above at section 1.2) of efforts by the Social Security Tribunal of Canada and the British Columbia Civil Resolution Tribunal to communicate with users of their systems in easily understandable language revealed that these tribunals see a connection between linguistic simplicity, on the one hand, and access to justice (A2J) on the other hand. A2J arguments or rationales for making adjudicative decisions more, or sufficiently, readable to large numbers of people feature frequently throughout this dissertation and often manifest within statements from sitting or prospective judges about how they see their roles.[159] While these types of A2J arguments are not the only ones that can be advanced in efforts to make case law more readable,[160] many of the A2J arguments are compelling. It is therefore worth considering at this point what exactly is meant by the phrase "access to justice," and whether there is a connection between readable case law and A2J.

When we think about A2J, it is important to recognize that people hold many different, and often conflicting, conceptions of justice. At a high level of generality, justice theorists tend to agree that their theories are broadly concerned with "the basic structure of society,"[161] or with ways "of

---

[159] See, for instance, Canada, Department of Justice. *The Honourable Mahmud Jamal's Questionnaire*, *supra* note 48, at Question 4: Justice Jamal says that court decisions should convey meaning to everyone who reads them, whether the readers are lawyers or lay people, and that this "is especially important at a time when many justice system participants are self-represented. Clear and accessible reasons concretely improve access to justice."

[160] See, for instance, Ash, et al, "Translating Legalese", *supra* note 38 at 136 (wherein the authors argue that court decisions can only serve a legitimizing function that strengthens public trust in the courts when these decisions are capable of being understood by the public).

[161] John Rawls, *A Theory of Justice* (Cambridge, MA: Harvard UP, 1999) at 1.

assigning rights and duties in the basic institutions of society."[162] Disagreement between theorists is more evident in the specific details of their theories: some are resource- or welfare-distributive in their focus, and generally strive for equality;[163] some are more concerned with qualitative or process-based considerations (e.g.: the forms of domination and oppression in areas of decision-making or culture that can materialize even where allocative equality is achieved);[164] and others are focused on capabilities (e.g.: to achieve a full and meaningful life,[165] or to access relevant markets and participate politically as citizens.)[166] Some theories have an individualist, liberal focus,[167] while others espouse a more collectivist view of justice.[168] Given the wide divergence in approaches, premises, and conclusions among these theorists, and assuming that they are all discussing a common overarching concept of justice, we can see how any definition of "justice" that is developed inductively from these theories would be very broad. What is clear, however, is that all these theories encompass a subject area that includes far more than just law; justice, in this context, is about the many different factors that must be taken into account if people are to live satisfactorily together within a society.

---

[162] Rawls, *A Theory of Justice*, *supra* note 161, at 4; see also David Miller, "Two Ways to Think About Justice" (2002) 1 Politics, Philosophy & Economics 5 at 6.
[163] See generally Rawls, *A Theory of Justice*, *supra* note 161; Susan Moller Okin, "Justice and Gender: An Unfinished Debate" (2004) 72 Fordham Law Review 1537 (essentially supporting Rawls' egalitarian approach, but adding in a greater emphasis on notions of equality within families, and across the gender divide).
[164] See Iris Marion Young, *Justice and the Politics of Difference* (Princeton: Princeton UP, 2011) at 15-38.
[165] See Martha Craven Nussbaum, *Sex & Social Justice* (Oxford: Oxford UP, 1999) at 39-42 (describing the capabilities that every human ought to have, and without which a life falls "short of being a good human life" – at 42. See also Debra Satz, "Amartya Sen's the Idea of Justice: What Approach, Which Capabilities" (2012) 43 Rutgers Law Journal 277 at 286-292 (describing Amartya Sen's capability-based approach to theorizing about justice).
[166] Elizabeth S Anderson, "What Is the Point of Equality?" (1999) 109 Ethics 287 at 317-19 (describing a theory of "democratic equality" that prioritizes the development of individual capabilities to function "as a human being, as a participant in a system of democratic cooperation, and as a citizen of a democratic state" – at 317).
[167] See generally Rawls, *A Theory of Justice*, *supra* note 161.
[168] See Nicola Lacey, "Theories of Justice and the Welfare State" (1992) 1 Social & Legal Studies 323 (for a social-constructionist communitarian approach toward justice that Lacey advances in opposition to Rawls' and Dworkin's theories of justice – theories that "unarguably take the individual and his or her needs and interests as their starting point, and move from there towards inferred needs for collective provision" – at 330). See also Young, *Justice and the Politics of Difference*, *supra* note 164 at 9 (noting how injustice in the form of oppression often tends to affect groups as well as individuals, and stressing the importance of considering social groups within theories of justice).

The concept of "justice" within much of the discourse on A2J, however, seems to be far narrower in scope. Rhode's treatment of the idea of justice in an A2J context is somewhat prototypical in this regard. She begins by briefly noting how other theorists commonly conflate procedural and substantive justice:

> In theory, "equal justice under law" is difficult to oppose. In practice, however, it begins to unravel at key points, beginning with what we mean by "justice." In most discussions, "equal justice" implies equal access to the justice system. The underlying assumption is that social justice is available through procedural justice. But that, of course, is a dubious proposition.[169]

In the following paragraph, though, Rhode begins a discussion (that fills the remainder of her monograph) about A2J shortcomings in contemporary USA that are all procedural or legalistic in nature: insufficient access to publicly-funded counsel; barriers that prevent non-lawyers from providing legal assistance; a legal system that is incomprehensible to self-represented litigants; and, ineffective *pro bono* regimes – among other problems.[170] Rhode's somewhat superficial acknowledgement that justice is not achieved solely through a legal system is deeply undermined by her near-exclusive focus on the need for changes within the limited boundaries of the American justice system in order to improve A2J. She perpetuates the equation of justice and law, in a way that cannot easily be reconciled with the broader justice theories that I have referred to, above.

Rhode is not unique in this respect. Cappelletti, for instance, begins a general discussion of the global A2J movement by noting the existence of a form of (broad) *injustice*: "the traditional civil and political liberties are a futile promise, indeed a deception for those who, because of economic, social and cultural reasons *de facto*, have no capacity to accede to and to benefit from those liberties."[171] However, the bulk of his discussion describes the successive "waves" of the

---

[169] Deborah L Rhode, *Access to Justice* (New York: Oxford UP, 2004) at 5.
[170] See generally Rhode, *Access to Justice*, *supra* note 169.
[171] Mauro Cappelletti, "Alternative Dispute Resolution Processes within the Framework of the World-Wide Access-to-Justice Movement" (1993) 56 Modern Law Review 282 at 283 ["Alternative Dispute Resolution Processes"].

A2J movement – a first wave attempting to provide better legal aid, insurance, and information to those who might need to access the legal system; a second wave that seeks to mass collective interests (through class actions, for instance) to more efficiently access the legal system; and, a third wave of alternative dispute resolution initiatives designed to streamline some cases into less formal resolution systems.[172] Here, again, we see how the idea of A2J devolves into a discussion about access to courts, or lawyers, or other dispute resolution mechanisms, that have very little to do (at least directly) with notions of distributional resource equality, or one's capability to live a full and meaningful life, or other core ideas drawn from broad theories of justice. Similar trends (whereby A2J becomes predominantly about access to law) can be seen in writing by Engler,[173] Carpenter,[174] and Duggan & Ramsay.[175]

Notwithstanding the gap that often exists between notions of justice in A2J literature and in more theoretical writing about justice, some scholars effectively connect the idea of justice across this conceptual divide. Hughes, for instance, alerts us to the fact that different individuals face vastly different barriers to accessing the justice system, and cautions against adopting one-size-fits-all reform initiatives that will fail to address the needs of certain classes of people.[176] She calls specific attention to cognitive and literacy challenges, and technology challenges (like

---

[172] See generally Cappelletti, "Alternative Dispute Resolution Processes," *supra* note 171.
[173] See Russell Engler, "Opportunities and Challenges: Non-lawyer Forms of Assistance in Providing Access to Justice for Middle-income Earners" in Michael J Trebilcock, Lorne Mitchell Sossin & Anthony J Duggan, eds, *Middle Income Access to Justice* (Toronto: U of T Press, 2012) 145 (for a discussion of A2J deficiencies and enhancements that is completely focused on litigation and legal dispute resolution in courts and tribunals).
[174] See Anna E Carpenter, "Active Judging and Access to Justice" (2017) 93 Notre Dame Law Review 647 (for a discussion of how judges could facilitate A2J enhancements within their courtrooms by taking on more active roles in assisting unrepresented litigants).
[175] See Anthony Duggan & Iain Ramsay, "Front-end Strategies for Improving Consumer Access to Justice" in Michael J Trebilcock, Lorne Mitchell Sossin & Anthony J Duggan, eds, *Middle Income Access to Justice* (Toronto: U of T Press, 2012) 95 (for a discussion of improved ways to promote consumer protection rights that would precede, and ideally prevent, litigation, but that is still focused only on the protection of legal rights, rather than more broad social objectives).
[176] Patricia Hughes, "Advancing Access to Justice through Generic Solutions: The Risk of Perpetuating Exclusion" (2013) 31 Windsor Yearbook of Access to Justice 1 ["Advancing Access to Justice"].

internet and computer access and/or literacy problems) – both of which may have roots in underlying socio-economic disadvantages – that prevent some individuals from fully realizing their legal rights.[177] This discussion, however, implicitly embraces broad understandings of justice: Hughes notes how homogenizing ways of thinking about people and their A2J problems will fail to resolve their problems,[178] and she recognizes how one source of disadvantage (a lack of literacy, or technology, etc) will unsurprisingly exacerbate another source of disadvantage (an inability to effectively access the legal system). Thus, Hughes appears to argue, somewhat unusually (or in reverse, when compared to much other A2J scholarship), that we need to establish better baseline justice (from a capability perspective) in order to positively effect certain people's ability to meaningfully access the legal system. Other scholars, like Cotterrell,[179] Maldonado,[180] and Economides, Timoshanko & Ferraz,[181] all do similar work, but in their own different ways, in linking substantive (broad) ideas of justice to their respective discussions of access to justice.

---

[177] Hughes, "Advancing Access to Justice", *supra* note 176 at 13-20. A similar point has been made, specifically about language-based reform, by Danet, "Language in the Legal Process", *supra* note 31 at 489: "Perhaps the better educated, who are presently more capable of coping, will benefit somewhat, but it is unlikely that the disadvantaged will be significantly better off as a result of linguistic reform."

[178] Hughes, "Advancing Access to Justice", *supra* note 176 at 3: "the [generic, commonly-proposed] reforms are disconnected from a broader understanding or exploration of how the current system reinforces relations of inequality."

[179] See Roger Cotterrell, "Access to Justice, Moral Distance and Changing Demands on Law" (2019) 36 Windsor Yearbook of Access to Justice 193 (for a very explicit recognition of the typical disconnect that exists between A2J discourse and theories of justice developed by philosophers, and for the author's idea of "moral distance" as a concept that measures the gap between the governed and the governing classes. Cotterrell sees moral distance as a "useful organising idea […] because it emphasises not merely technical obstacles to securing justice but also a moral relationship that incorporates the regulators and the regulated in a single community, structured by some degree of trust and links of reciprocity that make fidelity to law worthwhile" – at 207.)

[180] See Daniel Bonilla Maldonado, "The Right to Access to Justice: Its Conceptual Architecture" (2020) 27 Indiana Journal of Global Legal Studies 15 (for a discussion of the roots of A2J within broader political theory, that sees the state's obligation to provide mechanisms for effectively resolving disputes as part of the social contract that allows people to live together cooperatively, in a realm of reason rather than passion).

[181] See Kim Economides, Aaron Timoshanko & Leslie S Ferraz, "Justice at the Edge: Hearing the Sound of Silence" (2020) 41 Adelaide Law Review 39 (for a discussion of a possible next counter-wave in A2J evolution, that could see substantive ideas of justice that currently only exist within non-dominant segments of society – e.g.: giving legal personhood to the natural environment – being drawn centripetally into the justice mainstream).

As the above discussion has shown, it is difficult to define exactly what access to justice entails, since there does not seem to be a consensus on the point.[182] However, Macdonald offers what I find to be the most helpful working construct of the parameters of a system of justice that is accessible: "[a]n accessible system is said to be one that produces: (1) just results, (2) and fair treatment, (3) at reasonable cost, (4) with reasonable speed; and that (5) is understandable to users, and (6) responsive to needs; that (7) provides certainty, and (8) is effective, adequately resourced and well organized."[183] This conception of A2J is broad enough to account for most notions of justice, and it considers questions of access from a variety of important perspectives.

The above discussion about how justice and A2J are understood in different ways by scholars writing on different topics is central to the question of whether court and tribunal decisions need to be highly readable in order to improve A2J, because we need to know what notion of justice is in play in order to answer this question.

If we adopt a broad view of justice, then the readability of adjudicative decisions plays a relatively small role in contributing to A2J. Greater readability of these decisions will help more people (who have the capability to find and read the decisions) to understand the decisions, which, in turn, should help these people to understand any of their legal rights or interests that come up within the decisions. In this sense, decisions that are highly readable likely contribute to a net increase in aggregate abilities to access procedural justice, or the legalistic dimensions of justice. However, if we understand the notion of justice in broad terms, as justice theorists tend to understand the notion, then we are likely focused far less on the legalistic dimension of justice than

---

[182] Ann-Katrin Habbig & Ingrid Robeyns, "Legal Capabilities" (2022) 23:4 Journal of Human Development and Capabilities 611 at 620.
[183] Roderick A Macdonald, "Access to Civil Justice", in Peter Cane & Herbert M Kritzer, eds, *The Oxford Handbook of Empirical Legal Research* (Oxford: Oxford UP, 2010) 493 at 508-509.

on the larger social dimension of justice. We might ask (as I imagine that Nussbaum or Okin might ask), "what good does it do me to understand a judicial decision better, if the substance of the decision reinforces notions of gender inequality that are fundamentally unjust? How does the high level of readability of this decision improve access to justice when the effect of the decision is to perpetuate injustice"? Under a broad conception of justice, the form of a judicial decision is not nearly as important as the substance of the decision; and neither the form nor the substance are dominant instruments for achieving justice. Courts that produce these decisions are perhaps one of the "basic institutions of society"[184] that factor into theories of justice, but they are only one such institution, among many that play a role in allocating resources, or promoting equality, or fostering important capabilities, etc. One ought to be careful not to oversell the impact that the readability of judicial decisions might have on a vision of A2J that incorporate broad notions of justice.

Depending on the specific details of one's broad theory of justice, however, it is possible that readability levels of adjudicative decisions could have a somewhat more significant impact on A2J. For instance, consider Young's idea that "social justice means the elimination of institutionalized domination and oppression."[185] One could argue that uses of technocratic and/or archaic language, legalese, and other forms of writing that make up the legal "genre"[186] (within which judicial decisions are situated) are all exercises in domination on the part of the judiciary and the legal profession, designed to preserve their power. To the extent that adjudicative decisions

---

[184] Rawls, *A Theory of Justice*, *supra* note 161 at 4.
[185] Young, *Justice and the Politics of Difference*, *supra* note 164 at 15.
[186] See Robert A Ferguson, "The Judicial Opinion as Literary Genre" (1990) 2 Yale Journal of Law & Humanities 201; Gerald Lebovits, Alifya V Curtin & Lisa Solomon, "Ethical Judicial Opinion Writing" (2008) 21 Georgetown Journal of Legal Ethics 237; Richard A Posner, "Judges' Writing Styles (And Do They Matter?)" (1995) 62 University of Chicago Law Review 1420; Maurizio Gotti, "Text and Genre" in Lawrence Solan & Peter Meijes Tiersma, eds, *The Oxford Handbook of Language and Law* (Oxford: Oxford University Press, 2012) 52; and Peter Meijes Tiersma, "A History of the Languages of Law" in Lawrence Solan & Peter Meijes Tiersma, eds, *The Oxford Handbook of Language and Law* (Oxford: Oxford University Press, 2012) 13 (all of which describe elements of a unique legal genre or style of writing that is distinct from normal writing, and that is far less comprehensible to non-experts in law).

are not easily readable, people will need to rely upon lawyers (the community from which judges, and often also tribunal members, are selected) in order to understand the decisions – thereby preserving a form of power and dominion of the lawyerly intellectual and socio-economic elite over others in society. Decisions that are not easily readable may also render it difficult for the public to hold legislators and judges accountable for their governmental actions (making and interpreting the law, respectively), which, in turn, could help to entrench powerful individuals within their judicial/legislative roles – because they are protected from scrutiny through obscuring language within judicial decisions.

These arguments that certain forms of language or rhetoric (that may manifest in more or less readable decisions) can hide truths, or build/preserve power, are made to differing degrees in a wide range of scholarly literature.[187] Drawing on Young's perspectives,[188] one might characterize such uses of language by judges and tribunal members – uses that would in many cases be associated with reduced readability levels of the relevant decisions – as being a form of linguistic domination and oppression that creates injustice. In other words, there are at least some circumstances, and some broad theories of justice, that (when combined) can illustrate how the readability of an adjudicative decision impacts both procedural and substantive dimensions of justice, albeit in relatively small ways.

The impact that readability can have on A2J, however, is probably inversely proportional to the breadth of one's underlying definition of justice – which is to say that the readability level of a judicial decision probably has a much greater impact on A2J if one defines "justice" narrowly

---

[187] See Tiersma, "A History of the Languages of Law", *supra* note 186 at 22; Danet, "Language in the Legal Process", *supra* note 31 at 452, 465, 468, 527, and 542; Mark Adler, "The Plain Language Movement" in Lawrence Solan & Peter Meijes Tiersma, eds, *The Oxford Handbook of Language and Law* (Oxford: Oxford University Press, 2012) 67 at 74; and, Guyora Binder & Robert Weisberg, *Literary Criticisms of Law*. (Princeton: Princeton UP, 2000) at 297.
[188] Young, *Justice and the Politics of Difference*, *supra* note 164.

in its more legalistic and procedural sense. Essentially, the argument here is that people need to know and understand the law if they are to be able to govern themselves in accordance with the law, so the law must be sufficiently readable to the people for this purpose: "[i]mportant decisions should be written so that people can easily understand how their rights are affected."[189] Similar points about the importance of fair notice to citizens as a condition of the rule of law are made extensively throughout legal literature.[190] It follows from this argument that, if one equates justice with law, A2J will be helped significantly by the promulgation of more readable adjudicative decisions.

All of the preceding discussion has implied that ideas of "access" and "justice" are positively correlated when we discuss A2J – that an improvement in either "access" or "justice" should translate into increased A2J. It is important to consider the possibility, however, that these two variables might be negatively correlated under certain circumstances. For instance, using humour in a judicial decision could conceivably increase readability of the decision; most humour involves relatively base forms of language that are probably accessible to large numbers of people. However, as a number of commentators have either stated or implied, it may be inappropriate for judges or tribunal members to use humour within their decisions.[191] Such uses of language run

---

[189] Lebovits, Curtin & Solomon, "Ethical Judicial Opinion Writing", *supra* note 184 at 247.
[190] See Paul Weiler, "Two Models of Judicial Decision-making" (1968) 46 Canadian Bar Review 406 at 419 and 465; Thomas R Lee & Stephen C Mouritsen, "The Corpus and the Critics" (2021) 88:2 University of Chicago Law Review 275 at 302; Paul Gowder, "Is legal cognition computational? (When will DeepVehicle replace Judge Hercules?)" in Ryan Whalen, ed, *Computational Legal Studies: The Promise and Challenge of Data-Driven Research* (Northampton, Mass: E Elgar Publishing, 2020), 215 at 226 ["Is legal cognition computational?"]; and, Brian G Slocum, "The Contribution of Linguistics to Legal Interpretation" in Brian G Slocum, ed, *The Nature of Legal Interpretation: What Jurists Can Learn about Legal Interpretation from Linguistics and Philosophy* (Chicago: Chicago UP, 2017) 14 at 31.
[191] See Lebovits, Curtin & Solomon, "Ethical Judicial Opinion Writing", *supra* note 184 at 271-275; Sharyn Roach Anleu, Kathy Mack & Jordan Tutton, "Judicial Humour in the Australian Courtroom" (2014) 38:2 Melbourne University Law Review 621 at 625-29; Benjamin Nathan Cardozo, "Law and Literature" in Margaret Hall, ed, *Selected Writings of Benjamin Nathan Cardozo* (New York: Matthew Bender & Co, 1947) at 349-50.

the risk of showing disrespect to parties, or of making light of the role that society has entrusted to the decision-makers, and could perpetuate feelings of injustice.

A related situation could arise where, in an effort to increase readability, decision-makers use more common and accessible language within their decisions, but in ways that lead to imprecision, and therefore uncertainty in the law. Again, such efforts may increase individuals' ability to conceptually "access" the law, in that people would better understand the linguistic meaning of the relevant judicial decisions, but the efforts may concurrently undermine substantive justice by introducing elements of inconsistency into the law, or by causing confusion about the legal meaning of the decisions.[192]

What, then, can be said about the impact of the readability of judicial decisions on A2J? The effect of readability on A2J seems to be generally positive, in that increases to the readability levels of judicial decisions would normally be expected to assist more people in accessing the procedural or legalistic dimensions of justice more often (by better communicating to the people what their rights, obligations, and interests are in different circumstances). Additionally, increases to readability may also help to break down linguistic barriers between traditionally powerful and powerless people in ways that reduce oppression and domination, and contribute to substantive notions of justice – depending on how, exactly, the readability increases are accumulated. However, in at least some cases, it is possible to envision how readability increases may be achieved in ways that actually weaken perceptions of justice. Overall, readability does not appear to have an overwhelming impact on A2J, regardless of whether one holds a broad or narrow understanding of the term "justice."

---

[192] See Lawrence B Solum, "The Interpretation-Construction Distinction" (2010) 27 Constitutional Commentary 95 (for an explanation of the difference between linguistic meaning and legal meaning, and the importance of preserving the distinction between these two types of meaning).

That being said, most A2J initiatives are relatively modest in scope, and seek to achieve limited gains wherever possible through incremental change; they do not, generally, seek to find a single 'silver bullet'-solution that is capable of solving all A2J problems in a single shot. Consequently, while one ought to be realistic about the potential for more readable adjudicative decisions to improve A2J, there are sufficient indications in the literature to suggest that readability gains will likely have the kind of small, but meaningful, positive impact on A2J that we typically see in A2J intervention strategies. And in any case, regardless of the impact that readability initiatives might have on A2J, there remain other reasons why readable decisions may be desirable that are independent of A2J.

1.7 The Rapidly Changing Worlds of AI and NLP Technologies

It must be acknowledged at the outset that this dissertation was written during a time when generative artificial intelligence (AI) and Natural Language Processing (NLP) technologies were developing at exceedingly fast rates. This contextual reality raises an important threshold question: do readability levels (or even the original texts) of adjudicative decisions matter in a world where machines: (1) can summarize these decisions for readers in customizable ways; (2) can write or edit the decisions for the authors to render them more understandable for readers; or (3) may even replace judges and adjudicators as more efficient alternative instruments for adjudicating cases? In other words, is there any value in worrying about whether people can read and understand the actual content of a court or tribunal decision if increasingly sophisticated and accessible tools exist that can communicate the core meaning of the decision to these people in ways that they will understand?

The impacts that AI technologies might have on the profession and practice of legal adjudication are vast, and one could barely scratch the surface of this topic even in a complete

doctoral dissertation written exclusively on this topic. But we have some experience now with how these technologies are being used in law, and some maturing theories about how judging might change as a result of the adoption of AI tools. From this limited (but constantly expanding) information, I suggest that readability levels of adjudicative decisions matter a great deal, and will continue to matter for the foreseeable future.

To begin with, the practical and normative justifications for retaining a "human-in-the-loop"[193] when it comes to judging cases within common law jurisdictions has been well established. Gowder convincingly notes that flexible considerations about what is good "policy" or "justice" under the circumstances will often need to factor into decision-making, particularly in novel classes of legal cases.[194] Gowder also argues that machines are not as well-suited to weighing these abstract considerations as humans.[195] Morison & Harkens similarly note how the human and social elements of adjudication will prevent machines from replacing judges for some time to come:

> In order for algorithms to entirely permeate this socio-legal milieu, AI technologies would need to reach the point whereby a judge, or at least some form of surrogate judging system, could be produced either semi- or fully autonomously using big data analytics, and that this could act independently to adjudicate on, and intervene in, the governance of human relations and interactions. To create a system like this would require a complete instrumentalisation of all these social aspect of law into something that could be technologically calculated and predicted. […] When the fuller and more dynamic context of a formal courtroom hearing is considered, it becomes even less likely that even the most sophisticated machine learning would work.[196]

---

[193] The expression "human-in-the-loop" in the context of machine learning, artificial intelligence, and algorithmic decision-making contexts refers to the idea of human oversight of these systems: "keeping a human in the loop is a deliberate attempt to maintain human agency and accountability, and to provide legal safeguards and quality control" (Therese Enarsson, Lena Enqvist & Markus Naarttijärvi, "Approaching the Human in the Loop – Legal Perspectives on Hybrid Human/Algorithmic Decision-Making in Three Contexts" (2022) 31:1 Information & Communications Technology Law 123 at 124.

[194] Gowder, "Is Legal Cognition Computational?", *supra* note 190.

[195] Gowder, "Is Legal Cognition Computational?", *supra* note 190 at 220-224.

[196] John Morison & Adam Harkens, "Re-engineering Justice? Robot Judges, Computerised Courts and (Semi) Automated Legal Decision-Making" (2019) 39:4 Legal Studies 618 at 630-631.

As these scholars suggest, society is some distance away from being able to effectively replicate the practical dispute resolution functions of a judge using machines.

But perhaps more importantly, Livermore has articulated a democratic requirement for maintaining human judges over machines.[197] Livermore argues that the legitimacy of the legal system depends in large part on the participation rights of the citizenry within the adjudicative elements of that system, and he suggests that participation must include rights to observe, make arguments, present evidence, and receive reasons for decisions.[198] He notes that "[h]uman reasoning is mutually recognizable, based on notions such as causality, morality, and shared mental models of the world. […] Most broadly, human beings share an *umwelt*[199] that allows for mutual recognition, comprehension, and communication"[200] between individuals that is simply not possible between a human and a machine. Livermore goes on to offer the following conclusion:

> if participation rights at the adjudicatory stage are required, then a human being must serve as the final decision maker. No matter how lifelike and intelligent, machines will not share an *umwelt* with human beings, a fact that fundamentally disrupts legal subjects' confidence that their arguments have been truly heard and considered by the decision maker. In addition, were law-as-computation to eliminate the need for human-led adjudication, it could interfere with the educative function of legal process that helps maintain and perpetuate the cultural and political conditions necessary to maintain the rule of law.[201]

Livermore's forceful defence of human adjudication is compelling: humans who are parties to disputes have feelings that will often need to be validated in order for them to accept a final decision, and a machine that does not have feelings will struggle to provide this validation in an authentic manner.

---

[197] Michael A Livermore, "Rule by Rules" in Ryan Whalen, ed, *Computational Legal Studies: The Promise and Challenge of Data-Driven Research* (Cheltenham, UK: Elgar, 2020) 238.
[198] Livermore, "Rule by Rules", *supra* note 197 at 256-257.
[199] The term "umwelt" is derived from German, and refers to "the environmental factors, collectively, that are capable of affecting the behaviour of an animal or individual" (Collins Online English Dictionary, *sv*: umwelt, online: <https://www.collinsdictionary.com/dictionary/english/umwelt>.
[200] Livermore, "Rule by Rules", *supra* note 197 at 257.
[201] Livermore, "Rule by Rules", *supra* note 197 at 257.

In Canada, although there has not yet been any effort to legislate how members of the judiciary use AI systems, the federal government recently introduced a bill in Parliament that would have, among other things, created a new *Artificial Intelligence and Data Act* (AIDA).[202] Although the bill died on the order paper when the Canadian Parliament was prorogued on January 6, 2025, it is instructive to consider how the bill would have affected Canadian law if it had been enacted and come into force. The proposed AIDA would only have applied to the private sector, but the concerns that the act addressed and the legislative responses to these concerns may foreshadow how future governments would legislate in respect of courts and tribunals. Specifically, the proposed AIDA, at section 8, obligated those responsible for AI systems to "assess and mitigate the risks of harm or biased output that could result from the use of the system." The term "biased output" was defined at subsection 5(1) of the proposed act as any content, recommendation, prediction, or decision that is made by an AI system, and that discriminates in a way that is inconsistent with the *Canadian Human Rights Act*. As this picture of the proposed *Artificial Intelligence and Data Act* indicates, the Canadian government has already shown concerns about the use of algorithmic systems for the purposes of decision-making in Canada's private sector. Future governments may have some of the same concerns about related decision-making in the public sector.

Additionally, some tribunals have issued guidance that seems to impose outright prohibitions on the use of AI tools by members when these members are drafting decisions or analyzing evidence. For instance, the Canadian Human Rights Tribunal's relevant Practice Direction states, "[a]djudication is a human responsibility. Tribunal members hear cases and make

---

[202] Bill C-27, *An Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to make consequential and related amendments to other Acts*, 1st Sess, 44th Parl, cl 39, 2022 (died on the order paper on January 6, 2025).

decisions based on the evidence and submissions. They do not use AI to write decisions or analyze evidence. Tribunal members are fully accountable for their decision-making."[203] Tribunals Ontario – the umbrella organization that encompasses 13 different provincial administrative tribunals – recently published almost identical guidance, with similar categorical language about how members do not use AI tools for core functions.[204]

To be clear, however, there are some jurists who (cautiously) advocate for the adoption of machine-based adjudication systems for simple disputes. For instance, Sir Geoffrey Vos, an appellate judge and the Master of the Rolls of England and Wales, recently asserted that automated tools should be used to adjudicate in lower-level cases and for alternative dispute resolution purposes, "provided that ultimately there is the ability to question an AI driven decision before a human judge. In the first instance, there is no reason why very minor decisions should not be made by the system."[205] And in Argentina, AI-based systems have already been used on relatively small scales for determining the outcomes of cases.[206] But calls for AI-based adjudication in Canada have not yet been heeded, and it is not clear that there is any critical mass of support for the use of AI-based adjudication in Canada, at least in cases involving relatively serious judicial and quasi-judicial decisions. A recent survey of 115 Canadian judges regarding their attitudes toward, and uses of, technology shows that these judges are more preoccupied with having acceptable internet

---

[203] Canadian Human Rights Tribunal, "Practice Direction: Use of Artificial Intelligence (AI) in Tribunal Proceedings", online: <https://chrt-tcdp.gc.ca/en/about-us/practice-direction-use-artificial-intelligence-ai-tribunal-proceedings>.

[204] Tribunals Ontario, "Practice Direction on the Use of Artificial Intelligence (AI) in Tribunal Proceedings", April 2025, online: <https://tribunalsontario.ca/documents/TO/Practice-Direction-on-AI_EN.html>: "Adjudication is a human responsibility. Tribunal members hear cases and make decisions based on the evidence and submissions provided by parties. They do not use AI to write decisions or analyze evidence. Tribunal members are fully accountable for their decision-making."

[205] Geoffroy Vos, "The Future for Dispute Resolution: Horizon Scanning" (Sir Brian Neill Lecture, The Society of Computers and Law, 17 March 2022), online: < https://www.judiciary.uk/wp-content/uploads/2022/03/MR-to-SCL-Sir-Brain-Neill-Lecture-2022-The-Future-for-Dispute-Resolution-Horizon-Scannings-.pdf>, at para 42.
3

[206] Victoria Medizabal, "Courts in Buenos Aires are Using ChatGPT to Draft Rulings," Rest of World (22 Nov 2024), online: <https://restofworld.org/2024/buenos-aires-courts-adopt-chatgpt-draft-rulings/>.

access, video-conferencing tools, and information technology support for remote hearings than they are with more advanced technological questions relating to uses of AI.[207] Consequently, I feel confident in predicting that human judges will continue to adjudicate cases and provide reasons for their decisions in Canada.

There remains the possibility that many judges will use AI tools to either write or edit their decisions in ways that might better communicate the contents of the decisions to the judges' audiences. Scholars have begun to consider the potential consequences of this use of AI technology by adjudicative decision-makers.[208] In one semi-serious and speculative look into the future, Richard Re suggests that the adoption of this technology will proceed in several iterative phases, but that ultimately "[r]hetorical craft will crowd out reason, skill levels will quickly become flattened, and human professionals will struggle to preserve the roles long allotted them."[209] This picture is bleak.

However, a small number of judges in different jurisdictions have already confessed to using generative AI tools to assist in writing decisions, and have endorsed this practice. Lord Justice Birss, a British appellate judge, stated during a conference presentation that he had previously used generative AI to prepare a summary about an area of law, and then incorporated that summary within one of his decisions.[210] It is unclear from the context whether Lord Justice Birss noted in his decision where and how AI had contributed to his reasoning, and the extent to

---

[207] See, generally, Canadian Institute for the Administration of Justice, *Report on the Results of a Survey of Canadian Judges Concerning Technology in Judging Conducted by the Canadian Institute for the Administration of Justice and Athabasca University in 2022*, online: <https://ciaj-icaj.ca/wp-content/uploads/page/2023/08/ciaj-athabasca_2023-report-on-canadian-judges-and-technology_en.pdf>

[208] See, for instance, Maura R Grossman et al, "The GPTJudge: Justice in a Generative AI World" (2023) 23 Duke Law & Technology Review 1 at 34.

[209] Richard M Re, "Artificial Authorship and Judicial Opinions" (2024) 92:4 George Washington Law Review 1558.

[210] This statement was reported in The Law Society (UK), "Generative AI – The Essentials" (17 November 2023), online: <https://www.lawsociety.org.uk/topics/ai-and-lawtech/generative-ai-the-essentials>, at s 2.1.

which he verified the output from the AI tool – although it seems clear that the AI tool contributed only a portion of the text, and did not actually make the decision on behalf of the judge.

In another example, United States Court of Appeals for the Eleventh Circuit Judge Newsom recently used two different generative AI tools to inform his interpretation of the meaning of key terms that were in dispute in that case, and included the results generated by these AI tools within an Appendix to his reasons.[211] Judge Newsom carefully articulated his method for querying these AI tools in a way that promoted replicability. And, his opinion included a measured, but clearly supportive message in favour of the use of these AI tools in judicial decision-making processes.[212] Finally, in Argentina, some courts are openly using ChatGPT to draft rulings for their decisions – albeit on a small scale, and with human experts still involved as 'editors' of the AI-drafted decisions.[213] As these cases suggest, the future has arrived: judges are already using AI-generated writing within their decisions.

But does this fact either reduce or augment our motivation to study the readability of adjudicative decisions? The adoption of AI tools to assist in decision-writing does not take away from existing rationales for studying the readability of adjudicative decisions, since the idea behind these studies is to assess how well a reader might understand the written texts, and this idea remains relevant regardless of who the author is, and regardless of the techniques that the author may have used when writing the text. To be sure, it might be useful to assess whether the use of generative AI tools by judges and tribunal members tends to improve or worsen the quantitative readability scores of their decisions, but this assessment really just adds a different nuance or dimension to

---

[211] *Snell v United Special Insurance Company*, No 22-12581, United States Court of Appeals, Eleventh Circuit (May 28, 2024), online: <https://scholar.google.com/scholar_case?case=9102150573727781811>, *per* Newsom J.
[212] *Snell v United Special Insurance Company*, *supra* note 211, *per* Newsom J.
[213] Medizabal, "Courts in Buenos Aires are Using ChatGPT to Draft Rulings", *supra* note 206.

the underlying broader inquiry about decision readability levels – which remains fundamentally valid and important. So, if anything, we have even more reasons to study decision readability levels in an environment where AI tools have contributed to the production of the decisions.

Lastly, it is possible that the existence of generative AI tools that can summarize adjudicative decisions in personalized ways for different readers has made the issue of (original) decision readability levels moot. After all, if an adequate substitute for the original decision that is understandable to the relevant audience exists, then perhaps it does not matter whether the original decision is understandable.

There is a plethora of different extractive summarization algorithms that can be used to summarize legal decisions. In a recent study, Deroy et al tested 15 of these algorithms against summaries written by human experts concerning the same 50 cases, to identify how well the summarization tools performed.[214] Among other things, these authors found that the algorithmic summaries tended to suffer from lead-bias (in that they drew disproportionately from the beginning of the summarized text), which meant that they tended to frequently exclude sentences from the "Ratio of the decision" part of the summarized text (that typically falls in the middle or end portion of a decision) that human experts included in their summaries.[215] Most AI summaries within this study would therefore be poor substitutes for the original decisions, in terms of explaining the rationales for the decisions – although it would likely be possible for future AI systems to correct for this flaw by drawing on a weighted basis from the relevant sections of a decision in order to summarize the key legal content.

---

[214] Aniket Deroy, et al, "An Analytical Study of Algorithmic and Expert Summaries of Legal Cases" in Erich Schweighofer, ed, *Legal Knowledge and Information Systems* (Amsterdam: IOS Press, 2021) 90.
[215] Deroy, et al, "An Analytical Study of Algorithmic and Expert Summaries of Legal Cases", *supra* note 214.

In an even more recent and relevant study, Ash et al compared participants' objective and subjective comprehension measurements for both AI- and human expert-generated summaries of United States Supreme Court opinions.[216] This study began by noting how most AI language models "do not translate well to the legal domain (as domain shifts are a major obstacle in natural language processing)."[217] The authors also described at length the general need for "prompt engineering" (that is, the careful and iterative development of prompting questions or instructions that feed into a generative AI system in order to ensure that relevant, responsive, and accurate output is produced by the system) when using generative AI systems, and their own extensive process for creating prompts that would yield the desired summaries of the court opinions in their study.[218] The results of this experiment are fascinating. After an initial series of prompts, the summaries generated by the GPT-4 system were substantially less readable than the full court opinions that they summarized, based on Flesch Reading Ease scores.[219] However, after a further prompt that instructed GPT-4 to use simple language within the summaries, the final versions were scored as being "easily [understandable] by 13-15-year-old students."[220] These final summaries were also more readable than ones generated by human experts.[221] This research suggests that AI systems could play a useful role in summarizing adjudicative decisions using language levels that meet the needs of different types of readers.

Participants in Ash et al's study who reviewed the final AI summaries were also better at correctly identifying the outcome of the case and the area of law that the case discussed than

---

[216] Ash, et al, "Translating Legalese", *supra* note 38.
[217] Ash, et al, "Translating Legalese", *supra* note 38 at 137.
[218] Ash, et al, "Translating Legalese", *supra* note 38 at 137-140.
[219] Ash, et al, "Translating Legalese", *supra* note 38 at 141 (Figure 2).
[220] Ash, et al, "Translating Legalese", *supra* note 38 at 141.
[221] Ash, et al, "Translating Legalese", *supra* note 38 at 141 (Figure 2).

counterparts who read human-generated expert summaries.[222] And, particularly for participants who did not have a college education, the participants preferred the AI summaries to human summaries.[223] In this experiment, AI summaries seem to yield tremendous benefits.

However, the authors of the study candidly note two main drawbacks that prevent AI summaries from being fully useful to members of the lay public who might need them most. First, "[s]imple prompts such as 'summarize this court opinion' yield unsatisfactory results. Prompt engineering that draws on legal domain knowledge produces summaries that are clear and crisp. Hence, legal experts play a critical role in the development, evaluation, and deployment of AI assistants in legal applications."[224] And, second,

> summaries are effective at conveying the main points of a legal argument but can miss some subtle nuances that legal professionals may find important. For example, a 7th-grade summary of *Brown v. Board of Ed.* misses discussion of empirical evidence about the psychological impacts of segregation. It also misses the important historical context around how Chief Justice Earl Warren persuaded the rest of the Court to rule 9-0 in favor of *Brown* in part by convincing them that they did not need to explicitly overrule *Plessy v. Ferguson* to reach the result.[225]

The missing nuances noted above may or may not be important to any particular reader of the *Brown* decision – although Ash et al seemed to think that they were important. But Ash et al's broader point was not confined to this one case: "it is not realistic to expect 'perfect' summaries. The 'ground truth of what constitutes a good summary will differ between contexts and even between two people."[226] It is probably unrealistic, then, to expect that every person who wants to understand a court decision will, today, be able to obtain an AI-generated summary that meets that person's needs. These limitations with AI-generated summaries could be significant: if a person

---

[222] Ash, et al, "Translating Legalese", *supra* note 38 at 142.
[223] Ash, et al, "Translating Legalese", *supra* note 38 at 142.
[224] Ash, et al, "Translating Legalese", *supra* note 38 at 143.
[225] Ash, et al, "Translating Legalese", *supra* note 38 at 143.
[226] Ash, et al, "Translating Legalese", *supra* note 38 at 143.

cannot figure out how to make a machine compute the necessary answer, or if a machine can only compute an answer that includes some (but not all) of the information that one needs, then the machine is not optimally useful.

For individuals who are interested in Canadian case law summaries, and who do not have access to subscription-based databases (like the Westlaw or Lexis databases), the CanLII database now includes AI-generated case summaries for almost all case law from Alberta, Saskatchewan, Manitoba, and Prince Edward Island.[227] Ideally, this law-specific AI summary tool would eliminate the need for users to have their own extensive knowledge of how to prompt-engineer the generative text tool, and would also produce reliable and accurate summaries. However, CanLII also recently announced the launch of a user survey initiative, wherein users provide feedback on the usefulness of these AI-generated summaries.[228] Although there is not yet any published research that assesses the accuracy of CanLII's AI-generated summaries, the existence of this survey seems to suggest that CanLII's summaries are in an early stage of maturation. And, in any event, they are currently not available for the majority of – and in particular, the larger – Canadian jurisdictions (like Ontario, and the Federal jurisdiction).

Based on the above general limitations, it seems that AI summaries are not yet capable of standing in the place of full adjudicative decisions for many of the audiences who might be interested in the cases. Well-crafted summaries certainly have a role to play in many contexts. For instance, law firms who wish to produce summaries for their clients about cases in order to show overall trends in the law might benefit from these tools, since the firms would likely have enough

---

[227] Lexum, "AI-Powered Case Analysis Added on CanLII for All Case Law From Alberta, Saskatchewan, Manitoba, and Prince Edward Island" (3 Jun 2024), online: <https://lexum.com/en/blog/ai-powered-case-analysis-added-on-canlii-for-all-case-law-from-alberta-saskatchewan-manitoba-and-prince-edward-island/>.
[228] CanLII, "CanLII Launches Nationwide Survey to Assess AI-Generated Legal Summaries" (1 Oct 2024), online: <https://blog.canlii.org/2024/10/01/canlii-launches-nationwide-survey-to-assess-ai-generated-legal-summaries/>.

in-house legal knowledge to engineer useful prompts for the generative AI system, and their clients may be satisfied with high-level details about a case. But the tools are far less suitable today for the parties to a case (who likely want to understand the full reasoning behind the decision), to prospective litigants (who also need to understand nuances in the law in order to understand their own positions), or to public bodies who might not have dedicated legal advisors but who still need to apply the law as interpreted by the courts. None of these audiences will necessarily be satisfied with summaries that omit legally-relevant details, and none of them are likely to have the legal expertise that is needed to appropriately prompt the machine as required to produce helpful and relevant summaries. Perhaps the proper function of these summaries, as suggested by one of the potential answers to a question about how people use CanLII's summaries,[229] is to simply help a reader determine more quickly whether it is necessary and worthwhile to read the full text of a particular decision.

One could reasonably expect that better AI-generated summaries will soon be available. Improvements in the technology and coverage of these tools is happening quickly. And, even though these tools are not necessarily suitable today as a full replacement for the text of a judicial decision for someone who needs to understand the decision with high degree of accuracy, they are still useful for many other purposes. I return to this point later, in Chapter 7, when I discuss readability-related recommendations for the future.

---

[229] This survey was available online from 1 Oct 2024 to 15 Nov 2024 at the following link: <https://survey-ca.yoursurveynow.com/survey/selfserve/53b/240930?list=1>. After presenting respondents to the survey with both the original text and an AI-generated summary for a statute and a tribunal decision, the survey asked respondents the following question: "After seeing the images and reading the description, which of the following statements would you associate with the AI summaries for caselaw and legislation?" The first possible answer (respondents could select all applicable answers from a list of 13 answers, including one free-text "other" answer) was "Helps me decide whether to read the entire decision or legislation."

At a broader level, the threshold question about whether readability matters in a rapidly-changing world of AI technologies is perhaps contentious. AI tools are already having important impacts on adjudication. But it does not appear as if any of the changing technologies, tools, or methods used in adjudicative decision-making and decision-writing have yet displaced the need for at least some audiences to be capable of understanding the original text of a decision that interests them, across a wide range of circumstances. So, the readability levels of these decisions do matter now, and will likely continue to matter for some time into the future. But this question should perhaps be reassessed on a regular basis, given the fast pace at which changes are happening in this domain.

1.8 Outline of this Dissertation's Contents

The above context provides important information about the social, legal, and academic environments from which this dissertation emerges, and about the overarching questions that it seeks to answer. Before moving into the substance of the dissertation, however, I offer here a brief roadmap of the dissertation that describes the key content within each of the following chapters.

Chapter 2 provides the theoretical framework and background of readability knowledge that one must first understand in order to assess and make meaning of the empirical work that follows within the dissertation. Within this chapter, I describe the evolution of theoretical bases for readability studies, and canvass the strengths, weaknesses, and different approaches (including first-generation "classic" approaches, second-generation "computational linguistics" approaches, and third-generation "machine learning / AI" approaches) toward such studies. This chapter also notes the major functional limitations of readability formulas as tools for enhancing already-written texts, and calls attention to the dangers of misusing such formulas (including the dangers represented by Goodhart's Law, which materialize when performance measures are

inappropriately used to establish performance targets).[230] The purpose of this chapter is to ultimately illustrate the validity and utility of quantitative readability studies.

The dissertation then proceeds with four separate and cumulatively important empirical undertakings that represent the core of my research. These empirical undertakings are contained in Chapters 3-6 of the dissertation, respectively.

Chapter 3 contains what I refer to as my "Formula-Identification Project." I begin by describing the methods for the project, and the theoretical rationale for key choices that were incorporated in the project's design. Within this undertaking, using a combination of expert raters, lay raters, quantitative linguistics, and statistical analysis, I generate and validate a new law-specific readability formula ("*Madden's Adjudicative Decision Readability Score*" or "MADRS") that is based on human-assigned readability scores for a large sample of extracts from actual court and tribunal decisions. I then compare this formula's performance in predicting the readability of decisions against the gold standard of human-assigned scores, and against a number of other general-purpose readability formulas, to identify the best formula for use on adjudicative decisions. The result that emerges from this chapter – the MADRS formula – represents an original contribution to both the legal and linguistics literature, in that no one has ever yet determined what existing or new readability formula ought to be used to measure the readability of texts that are drawn from the legal domain, and that use a unique legal language.[231]

Chapter 4 of this dissertation, containing what I call my "General Decision Readability Studies," applies my newly created readability formula in three different contexts. First, I study decision readability levels based on the level of court (inferior, superior, appeal, or apex) that issues

---

[230] Goriely, "The English Legal Aid White Paper and the LAG Conference", *supra* note 72 at 358-59.
[231] Danet, "Language in the Legal Process", *supra* note 31 at 476-81.

the decision. Next, I look at readability levels across different subject areas, for cases dealing with family, criminal, and tax law matters, respectively. Finally, I study readability level differences between courts and tribunals that are dealing with cases from the same subject areas (social security law, and human rights law, respectively). In each of these three contexts, the MADRS formula is applied to large corpora of Canadian adjudicative decisions to illustrate how well these decisions might be understood by their audiences. The chapter first describes the research methods that I use for the studies, and then presents and discusses the readability results, noting and analyzing the sub-groups of decisions that are more or less readable than expected. The results of this chapter allow me to not only state the readability levels of the decisions, but also to call attention to those instances where decisions appear to be clearly falling below the intended or ideal readability levels.

Chapter 5 contains what I refer to as my "Targeted Decision Readability Studies." This chapter again begins with a short description of the studies' methods. The chapter involves separate applications of the MADRS formula to two different corpora of judicial decisions that are collected from contexts (youth criminal court, and self-represented litigant cases) wherein one might expect the decisions to be more highly readable. Within this chapter, I go on to discuss and analyze the results, indicating how *Youth Criminal Justice Act* decisions and self-represented decisions are being written in comparison with other decisions, from a readability perspective. I conclude this chapter by noting how these decisions might not be meeting the needs of the unique audiences for the decisions.

Chapter 6 of the dissertation includes a final original empirical study of the readability levels of 1621 court and tribunal decisions drawn in a targeted and non-random way from every province, territory, and federal jurisdiction, on a wide range of legal subjects, written by 840

different decision-makers. The results from this study are coded for a number of different variables, including authoring judge or tribunal member's name, which allow for customized searches of the results based on one or more variables. Some of the more noteworthy results from this study are discussed within the chapter, but the primary purpose of this portion of the dissertation is to illustrate how the data can be used by others, and to make it widely available to all those who might want to examine readability levels of Canadian adjudicative decisions from their own perspectives.

Chapter 7 of my dissertation contains my recommendations, some of which are directed toward future researchers, courts and tribunals as institutions, and judges / tribunal members as authors of decisions, respectively. Before making these recommendations, however, the dissertation returns in some ways to Chapter 2's cautions about the limits of readability studies in order to note how quantitative readability scores are really only proxy measures for the probable comprehensibility of a text: readability formulas extrapolate from actual comprehension measurements in order to generalize about the types of text properties that correlate more or less with reader comprehension.[232] As a result, the recommendations within my dissertation about enhancing readability are not aimed directly at improving readability scores (since the scores are only derivatively-useful proxies for the things that we care about much more – reader comprehension). Rather, these recommendations are geared more toward the making of qualitative improvements in the writing of adjudicative decisions (which we would, in turn, expect to produce better readability scores for the future). That said, my recommendations do include some broad classes of writing changes that are most recommended for those who want to improve decision readability scores.

---

[232] Bailin & Grafstein, *Readability*, *supra* note 12 at 53-63.

Chapter 8 of this dissertation offers my brief conclusion and relates this research back to larger questions about how law may need to be communicated (in the broadest sense) to the people who law affects, if it is to be understandable, useable, and helpful to these people.

2. Understanding Readability Theory

2.1 Overview

This dissertation is grounded in readability theory that comes primarily from the fields of applied linguistics, psychology, and education. To explain this theory and how it relates to my dissertation, I will first discuss some of the broad theoretical concepts that are thought to play a qualitative role in reading comprehension and the reading process. I will then briefly describe the evolution and applications of quantitative readability theory over the last two centuries, and through what I refer to as three distinct "generations" of readability approaches: first-generation (or "classic") approaches, involving hand-counting of small numbers of linguistic features to predict readability; second-generation approaches involving computational and corpus linguistics that dramatically expanded the range of linguistic and contextual factors that contributed to readability measurements; and, third-generation approaches involving artificial intelligence and machine learning, that may represent the future of readability classification processes. I then explore some of the major criticisms of quantitative readability studies. Finally, this chapter finishes with a discussion about many of the other non-linguistic dimensions of communication that likely also influence one's ultimate comprehension of written messages, but that are not necessarily well-represented within quantitative readability formulas.

2.2 Readability and Reading Comprehension Generally

There are a wide variety of different theories of reading, and no one comprehensive theory has achieved clear dominance.[1] However, a basic understanding of the mechanical, cognitive, and

[1] Charles Perfetti & Joseph Stafura, "Word Knowledge in a Theory of Reading Comprehension" (2014) 18:1 Scientific Studies of Reading 22 at 22: "There is no theory of reading, because reading has too many components for a single theory. There are theories of word reading, theories of learning to read, theories of dyslexia, and theories of

behavioural processes that contribute to reading and reading comprehension is essential to any meaningful readability study, since "the readability of a text can only be evaluated if one has an explicit model of the processes involved in text comprehension."[2] In other words, it is helpful to know something about how reading comprehension happens within the human mind in order to better appreciate how some texts might be more readable than others.

Most theorists would agree that reading involves the processing of two distinct types of information: "(i) visual information, that is the one that comes from the printed page and (ii) non-visual information, that is, the information that comes from the brain of the reader."[3] The skills needed to process the first type of information could be thought of as mechanical in nature: the reader must "decode" the text on the page by mentally processing individual characters into words, words into sentences, etc.[4] A reader must have knowledge of letters, letter clusters, semantics, and syntax rules, among other things, in order to successfully decode a text.[5] But something more than this mechanical process is needed in order for a reader to derive meaning from the text. "Being able to see sentences in front of our eyes is not enough; we must know something of the language in which the material is written, about its subject matter, and about reading itself."[6] More recent reading theories have therefore tended to focus wholistically on how both the textual information,

---

comprehension at various grain sizes (sentence comprehension, text comprehension), appropriately targeted to a manageable part of reading."

[2] Walter Kintsch & Douglas Vipond, "Reading Comprehension and Readability in Educational Practice and Psychological Theory" in Lars-Goran Nilsson (ed), Perspectives on Memory Research (London: Psychology Press, 1979) 329 at 340 ["Reading Comprehension"].

[3] Maria Novary Ngabut, "Reading Theories and Reading Comprehension" (2015) 5:1 Journal on English as a Foreign Language 25 at 29.

[4] Frederick B Davis, "Psychometric Research on Comprehension in Reading" (1972) 7:4 Reading Research Quarterly 628 at 630.

[5] Perfetti & Stafura, "Word Knowledge in a Theory of Reading Comprehension", *supra* note 1 at 28 (Figure 1).

[6] Ngabut, "Reading Theories and Reading Comprehension", *supra* note 3 at 29.

and the contextual information (from the reader's own experience) that are involved in the reading comprehension process work together to produce meaning.

Bruggink et al offer a simple, but generally representative theory of reading comprehension that encapsulates many of the key concepts that other contemporary reading theorists also describe in their own representations of the comprehension process.[7] Their theory builds on the previous "Reading Systems Framework" theory that was first elaborated by Perfetti & Staffura.[8] According to Bruggink et al, there are 2 critical steps in the reading process: word identification;[9] and, word-to-text integration.[10]

Word identification involves 2 distinct processes. First, the reader must decode the textual representation of the word by "converting strings of letters (orthographic units, known as grapheme) into their corresponding sounds (phonological units, known as phoneme)."[11] Next, the reader must retrieve semantic information about that word from their long-term memory – which is to say, the reader must search their memory for information about what a word means, potentially in relation to other words with similar meanings.[12] For experienced readers, the word identification process happens relatively quickly and automatically, and this automation is helpful because it ensures "that enough cognitive resources remain available for the comprehension processes"[13] that follow word identification.

---

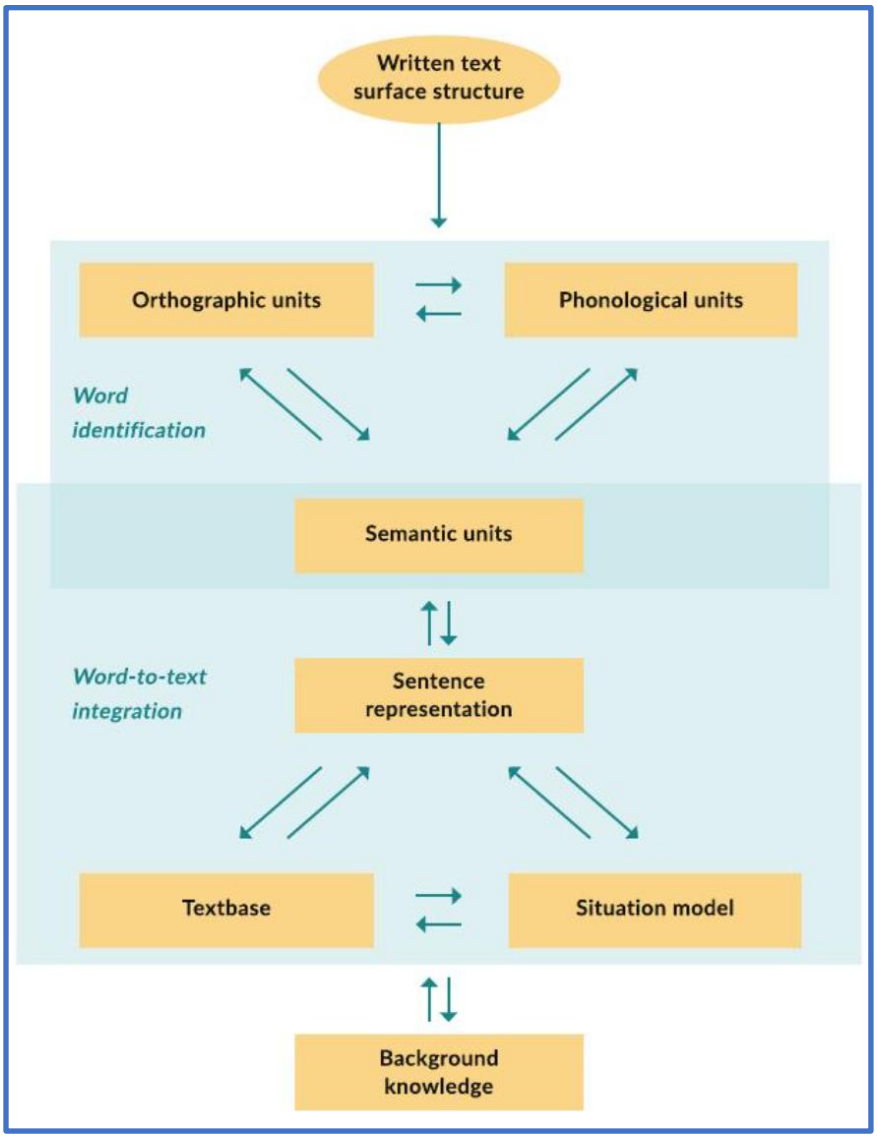
[7] Marian Bruggink, et al, *Putting PIRLS to Use in Classrooms Across the Globe: Evidence-Based Contributions for Teaching Reading Comprehension in a Multilingual Context* (Nijmegen, The Netherlands: Springer, 2022) [*Putting PIRLS to Use*].
[8] Perfetti & Stafura, "Word Knowledge in a Theory of Reading Comprehension", *supra* note 1.
[9] Bruggink, et al, *Putting PIRLS to Use*, *supra* note 7 at 6-9.
[10] Bruggink, et al, *Putting PIRLS to Use*, *supra* note 7 at 9-11.
[11] Bruggink, et al, *Putting PIRLS to Use*, *supra* note 7 at 7.
[12] Bruggink, et al, *Putting PIRLS to Use*, *supra* note 7 at 8-9.
[13] Bruggink, et al, *Putting PIRLS to Use*, *supra* note 7 at 6.

Word-to-text integration is a more cognitively complex process than word identification. It requires a reader "to integrate the individual word meanings into a sentence representation and combine all sentences into a mental model of the text."[14] At its core, this process of word-to-text integration involves the identification of propositions and arguments within a text, and then "combining them into an adequate model of the text."[15] For all texts, a reader's own background knowledge will influence the effectiveness of word-to-text integration, because readers will often be called upon to draw inferences from the text that will depend on their prior knowledge.[16]

But word-to-text integration will likely also be affected, particularly in longer texts, by factors like coherence in the text, and the presence or absence of clear micro- and macro-structures within the text. In this context, "structures" refer to the properties within a text that link it together as a whole, like the repetition or connection of propositions and arguments across different parts of the text.[17] Without this kind of structure, a text could simply be an "arbitrary list of propositions"[18] that have no relation to one another – which would be difficult for readers to integrate into a complete situational model for the text.

Bruggink et al's reading comprehension model is depicted clearly in the following schematic overview, which is extracted from their text[19] and reproduced below:

[14] Bruggink, et al, *Putting PIRLS to Use*, *supra* note 7 at 9.
[15] Bruggink, et al, *Putting PIRLS to Use*, *supra* note 7 at 10.
[16] Bruggink, et al, *Putting PIRLS to Use*, *supra* note 7 at 10.
[17] Kintsch & Vipond, "Reading Comprehension", *supra* note 2 at 343-346.
[18] Kintsch & Vipond, "Reading Comprehension", *supra* note 2 at 344.
[19] Bruggink, et al, *Putting PIRLS to Use*, *supra* note 7 at 13.



[Flowchart reproduced from Marian Bruggink et al, *Putting PIRLS to Use in Classrooms Across the Globe: Evidence-Based Contributions for Teaching Reading Comprehension in a Multilingual Context* (Nijmegen, The Netherlands: Springer, 2022)]

As this visualization aid shows, a reader who perceives a text must first mentally translate letters into words through both sight/spelling (orthographically) and sound/pronunciation (phonologically) in order to identify the words. The reader must then integrate these words (or semantic units) into both sentences, and the textbase as a whole, in order to create situational models for each sentence, and then for the entire text, that are highly influenced by the reader's

own background knowledge and experiences. This complete reading process allows the reader to derive meaning from the text.

In the sections of this chapter that follow, more specific theories of quantitative readability are described. These theories are ultimately aimed at allowing us to predict how well people might understand particular texts based only on the linguistic properties that the texts contain. In other words, quantitative readability theories focus on text-level properties (typically at the word- and sentence-level in the case of older and less sophisticated readability formulas, but often at the whole-text level in the case of more modern formulas). But quantitative readability theories do not generally start from positions that are established within any particular and generalized theory of reading comprehension. This reality has caused some scholars to characterize at least the classic quantitative readability theories as being "atheoretical."[20] I find this characterization to be unfair, and hope that the following sections will paint a picture of (at least the persuasive) quantitative readability theories as being *post-facto* linked to, and defensible on the basis of, more broad reading comprehension theories. In other words, any useful quantitative readability formula should be capable of being explained – at least in part – with reference to more general reading theory. This necessary connection between general reading comprehension theory and more specific quantitative readability theory explains why one must have a working understanding of the former theories as a foundation for understanding the latter theories that preoccupy the majority of this dissertation.

---

[20] Kintsch & Vipond, "Reading Comprehension", *supra* note 2 at 336-337, writing at a time that pre-dated the development of many more advanced and theoretically sound readability formulas, offer a highly critical view of older readability formulas: "the existing work on readability is close to useless from a theoretical perspective. It tells us very little about what makes a text easy or hard to read, and it is totally severed from research in cognitive psychology concerned with the process of reading comprehension. Instead, the traditional approach is an atheoretical one and relies on correlating plausible predictor variables (i.e., text properties) with plausible behavioral criteria of readability."

That being said, none of the quantitative readability theories discussed in the following sections of this chapter, or throughout the remainder of this dissertation, attempt to take into account the very reader-specific considerations of "background knowledge" or personal experience that any individual reader brings to bear on their understanding of a particular text. Consequently, in assessing the usefulness of quantitative readability approaches, we must concede at the outset that these approaches will always fail to fully represent the different factors that contribute to a reader's understanding of a text, because these approaches only consider text-related factors, and not reader-related factors, that influence the reading comprehension processes. However, they still offer us tremendously useful and concrete evidence about how well texts are likely to be understood by readers.

2.3 First-Generation Readability Formulas: The Hand-Counting Era

The first-generation idea of using statistical analysis and the study of formal linguistic characteristics of a text as measures of the quality of a text, and potentially of a reader's ability to understand the text, date back to the late 1800s, and to the work of L.A. Sherman.[21] Sherman focused heavily on sentence length,[22] and noted that "the English sentence was accommodated to the professional men of books, and grew, if not longer, certainly more ponderous and impracticable."[23] In Sherman's view, the literary sentence was in many ways best, and most comprehensible, when it followed the same length conventions as a (typically much shorter) sentence spoken in oral English.[24] However, Sherman's work uses quantitative measurements of sentence length mostly to demonstrate how the writers who he perceives to be exemplary (from a

[21] L A Sherman, *The Analytics of Literature* (Boston, Ginn & Company, 1893).
[22] Sherman, *The Analytics of Literature*, *supra* note 21 at 256-62 (Chapter XIX – The Literary Sentence Length in English Prose).
[23] Sherman, *The Analytics of Literature*, *supra* note 21 at 292.
[24] Sherman, *The Analytics of Literature*, *supra* note 21 at 292.

qualitative perspective) use shorter sentences, rather than to argue that such sentences correlate with improved reader comprehension.[25]

More quantitative studies of readability then emerged in the 1920s. For instance, Thorndike published a book in 1921 that listed the 10,000 words that appeared most commonly in his reference corpus of 41 commonly read texts, including the Bible, newspapers, elementary school books, and literary classics.[26] He was perhaps one of the first to link concepts of word range and frequency[27] with a student's ability to understand a particular word. Thorndike's overall premise appears to have been that "importance" of a word, as measured by reference to range and frequency scores, is relevant information for teachers in determining how and when they should expect students to achieve mastery of the word.[28]

Vogel and Washburne – drawing in part on Thorndike's word list – went on to develop the first quantitative readability formula using regression analysis in 1928, following a method that has been largely emulated by numerous other subsequent readability scholars.[29] These authors used qualitative rankings of the reading difficulty levels of 152 children's books, together with different measurements of linguistic properties relating to each book (e.g.: number of conjunctions in a 1000 word sample) that were compiled by a group of 20 volunteer teachers, to calculate which linguistic properties correlated most strongly with the qualitative reading difficulty levels for the books.[30]

---

[25] Sherman, *The Analytics of Literature*, *supra* note 21 at 292.
[26] Edward L Thorndike, *The Teacher's Word Book* (New York: Columbia University Teachers College, 1921) at iii.
[27] Thorndike, *The Teacher's Word Book*, *supra* note 26 at iii: "*Range* answers the question, 'How many of these forty-one different sources use the word?' or 'How widely is the word used?' *Frequency* answers the question, 'How often is the word used?'" (emphasis in original).
[28] Thorndike, *The Teacher's Word Book*, *supra* note 26 at iv-v.
[29] Mabel Vogel & Carleton Washburne, "An Objective Method of Determining Grade Placement of Children's Reading Material" (1928) 28:5 The Elementary School Journal 373 ["An Objective Method"]. See also James W Cunningham, Elfrieda H Hiebert & Heidi Anne Mesmer, "Investigating the Validity of Two Widely Used Quantitative Text Tools" (2018) 31 Reading and Writing 813 at 814: "Almost all readability formulas, past and present, are regression equations."
[30] Vogel & Washburne, "An Objective Method", *supra* note 29 at 374-76.

The authors then used regression analysis, relying on four of the most strongly correlated linguistic factors – specifically, (1) the number of different words occurring in a 1000 word sample (measuring lexical diversity, or type-token ratio); (2) the number of prepositions in a 1000 word sample; (3) the number of words in a 1000 word sample that were not found within Thorndike's list; and, (4) the number of simple sentences in a 75 sentence sample – to derive an equation that produced a form of readability formula.[31]

Various other readability formulas were later developed throughout the first half of the 20th Century. Dale and Tyler (1934) assessed the actual reading comprehension levels of low-income adults who read passages from 74 different texts relating to health problems, and then used correlations and regression analysis to determine the linguistic properties of texts that best predicted reading comprehension[32] (namely, "the number of different technical words in the selection, the number of different hard non-technical words, and the number of indeterminate clauses"[33]). In 1944, Irving Lorge similarly derived a readability grade-level formula intended for use within schools that was calculated on the basis of average sentence length, the density level of prepositional phrases within the text, and the ratio of uncommon words to total words within the text.[34] In 1948, Rudolph Flesch developed a more streamlined "Reading Ease" formula that only relied on two variables (average syllables per 100 words and average words per sentence, together with a constant), to score the readability of texts on a scale of 1 to 100.[35] Flesch advocated for use of his formula over others in part because it was simpler, and took less time, to calculate.[36] Then,

---

[31] Vogel & Washburne, "An Objective Method", *supra* note 29 at 376-80.
[32] Edgar Dale & Ralph W Tyler, "A Study of the Factors Influencing the Difficulty of Reading Materials for Adults of Limited Reading Ability" (1934) 4:1 The Library Quarterly 384 ["Factors Influencing the Difficulty"].
[33] Dale & Tyler, "Factors Influencing the Difficulty", *supra* note 32 at 401.
[34] Irving Lorge, "Predicting Readability" (1944) 45:6 Teachers College Record 404.
[35] Rudolph Flesch, "A New Readability Yardstick" (1948) 32 Journal of Applied Psychology 221 at 228-29.
[36] Flesch, "A New Readability Yardstick", *supra* note 35 at 222.

also in 1948, and partially in response to weaknesses that they perceived with Flesch's Reading Ease formula,[37] Dale and Chall produced a readability formula based only on average sentence length, and on the ratio words that could not be found within their list of approximately 3000 common words.[38]

Some of these early leaders in the field continued to revise their formulas during the second half of the 20th Century: Flesch and his colleagues developed the Flesch-Kincaid Grade Level Score in 1975, which relied on measures of words per sentence and syllables per word to produce an estimated reading grade level of a text[39] (and which was the formula that both the SST and the BCCRT used within their readability discourse, discussed above); and, Dale and Chall produced an updated word list in 1995 for use within their formula that was both longer, and more reflective of the ways in which words were used at the end of the 20th Century.[40]

As the above discussion illustrates, the first-generation idea of quantitatively measuring the readability of texts was well-established within different academic studies by the mid-1900s. Although many other readability formulas were developed following similar experimental methods during the second half of the century,[41] these formulas tended to differ from their predecessors in only subtle ways: they usually relied on relatively simple linguistic variables that

---

[37] Edgar Dale & Jeanne S Chall, "A Formula for Predicting Readability" (1948) 27:1 Educational Research Bulletin 11 at 12: "On the whole, we found the formula adequate. However, we also found some shortcomings. The most serious shortcoming was the count of affixes, which we found to be rather arbitrary, in the sense that two people making a count on the same sample would usually come out with a different number of affixes."
[38] Dale & Chall, "A Formula for Predicting Readability", *supra* note 37 at 16-18.
[39] J P Kincaid et al, Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch *Reading Ease Formula) for Navy Enlisted Personnel* (Memphis, TN: US Naval Technical Training Command - Research Branch, 1975).
[40] Jeanne S Chall & Edgar Dale. *Readability Revisited: The New Dale-Chall Readability Formula*. (Cambridge, MA: Brookline Books, 1995).
[41] See, e.g., Robert Gunning. *The Technique of Clear Writing* (New York: McGraw Hill, 1952) (creating the Gunning-Fog index, which measured readability on the basis of average sentence length and average number of words with 3 or more syllables); E A Smith & R J Senter. *Automated Readability Index.* AMRL-TR-66-22 (Wright-Patterson AFB, OH: Aerospace Medical Division, 1967) (creating the ARI measure, on the basis of word- and sentence-length).

could be hand-counted by researchers and educators, and these variables only related to the syntactical and lexical dimensions of a studied text's properties.

Although each new formula during this period was often critical of, and was proposed as being superior to, preceding formulas, more contemporary researchers have identified weaknesses in the larger class of formulas that were produced during this period that likely apply to all of the formulas. Kintsch & Vipond, for instance, are highly critical of the readability formulas that had been developed in the pre-computer era. They offer the following assessment of early readability formulas:

> [T]he existing work on readability is close to useless from a theoretical perspective. It tells us very little about what makes a text easy or hard to read, and it is totally severed from research in cognitive psychology concerned with the process of reading comprehension. Instead, the traditional approach is an atheoretical one and relies on correlating plausible predictor variables (i.e., text properties) with plausible behavioral criteria of readability. The trouble is that typically neither set of variables is adequate. We have mentioned the problems with the predictor variables used in current readability formulas. They are concerned with word and sentence properties at a superficial level; **at best, they are correlated with whatever makes a text easy or hard, but they are not the causes themselves**. The problem is that the predictors commonly used do not directly reflect either the content or the organization of a text.[42]

Kintsch & Vipond then go on to offer their own hypothesis about the factors that truly affect readability of a text:

> Reading comprehension is ultimately a process of acquiring information. The nature and structure of that information – that is, the characteristics of the meaning of a text, as well as the processes involved in deriving this meaning from the written text – are, we assume, the real determinants of readability.[43]

Notwithstanding their strong criticism of early approaches to the quantification of readability, it is interesting to note the tentative language that Kintsch & Vipond use ("we assume") to identify "the real determinants of readability." This tentative alternate theory about readability perhaps weakens

---

[42] Kintsch & Vipond, "Reading Comprehension", *supra* note 2 at 336-337 (emphasis added).
[43] Kintsch & Vipond, "Reading Comprehension", *supra* note 2 at 337.

the force of Kintsch & Vipond's criticism, in the sense that they have not provided any greater empirical justification for their own hypothesis about the determinants of readability than the correlation-based justifications that are provided for purely quantitative readability approaches.

Bailin & Grafstein have gone a step further by deconstructing and identifying flaws with each of the early readability formulas individually,[44] but also with the overarching methods and assumptions that all of these formulas rely upon – in particular the fact that these readability measures all seem to be concerned with the same two semantic and syntactic concepts: word difficulty (usually measured by word length or commonness), and syntactic difficulty (usually measured by sentence length or paragraph length).[45] In particular, Bailin & Grafstein refer to a study wherein texts were manipulated and re-written to score better on several classic readability formulas: this study concluded that the re-written texts were generally no more comprehensible to participants (although low-ability participants seemed to benefit from vocabulary simplification changes).[46] Bailin & Grafstein seem to place substantial weight on this study, and rely on it in offering the following conclusion about quantitative readability formulas: "by focusing exclusively on formal, quantifiable properties of texts, readability formulas have failed to provide a theoretical understanding of the properties that affect the readability of texts and consequently are unable to provide effective technical tools for the creation of more readable written works."[47] Bailin & Grafstein suggest that readability should be studied by reference to three more qualitative concepts: (1) a reader's background knowledge and assumptions;[48] (2) correct use of the linguistic

---

[44] Alan Bailin & Ann Grafstein, *Readability: Text and Context* (New York: Palgrave Macmillan, 2016) at 15-52 [*Readability*].
[45] Bailin & Grafstein, *Readability*, *supra* note 44 at 53-58.
[46] Bailin & Grafstein, *Readability*, *supra* note 44 at 59-63, citing MT Duffy & P Kabance "Testing a Readable Writing Approach to Text Revision" (1982) 74:5 Journal of Educational Psychology 733.
[47] Bailin & Grafstein, *Readability*, *supra* note 44 at 62.
[48] Bailin & Grafstein, *Readability*, *supra* note 44 at 178-80.

code, including use of linking words and phrases;[49] and, (3) the organization of a text, including its conformity to conventions or genre expectations.[50]

Although Bailin & Grafstein offer a compelling critique of quantitative readability formulas,[51] many of their concerns could probably be eliminated or reduced through the development of a more precise, relevant, and theoretically sound quantitative method for measuring how well an audience might understand a text (i.e.: the kind of method that my proposed dissertation strives to advance). In other words, to the extent that Bailin & Grafstein's critiques of existing first-generation readability formulas are valid, these critiques perhaps just discourage **misuses** of the formulas – they remind us that formulas only predict, using linguistic features as proxies for true comprehensibility, how difficult a text will be for people to read relative to other texts that are scored using the same formula. But these predictions are not accurate in all cases, for all readers, and the formulas cannot tell us with certainty what specific things need to change within a text in order for it to express the same meaning in a more understandable way.

In this sense, Flesch's calculation of a 0.7 correlation coefficient between his Reading Ease formula and the actual grade level for a text[52] can perhaps be better understood: the formula does a reasonably good job of measuring whether a text will likely be understandable to a group of grade 7 students – but it cannot tell us with precision how many grade 7 students will actually

---

[49] Bailin & Grafstein, *Readability*, *supra* note 44 at 181-84.
[50] Bailin & Grafstein, *Readability*, *supra* note 44 at 184-85.
[51] See, e.g., Bailin & Grafstein, *Readability*, *supra* note 44 at 183 ("[t]here is no reason to think that counting the number of sentences in a text which have linking problems would lead to a useful scale of difficulty"), 184 ("[s]yntactic difficulty of texts clearly cannot be assessed by counting up the number of these difficult syntactic structures"), and 185 ("we cannot effectively account for the relative difficulty of texts by counting up the textual properties that may be problematic for readers and ranking texts along a scale from easy to difficult") for examples of the authors' rejection of the notion that elements of readability can be quantitatively measured.
[52] Flesch determined that a text was suitable for a particular grade level if students within that grade could correctly answer 75% of comprehension questions about the text after having read it: Flesch, "A New Readability Yardstick", *supra* note 35 at 222-223.

understand a text that is classified as a grade 7 text (keeping in mind that we might all have different ideas about what constitutes "understanding" of a text – whether that amounts to the 75% score on a comprehension test that Flesch relied upon,[53] or to some other benchmark). Similarly, the formula cannot tell us the extent to which any particular grade 7 student might understand/misunderstand a text that the formula classifies as a grade 9 text.

The types of weaknesses that Bailin & Grafstein have described with first-generation quantitative readability formulas therefore help to explain why such formulas do not, and likely cannot ever, produce perfect correlations between predicted and actual comprehension levels of texts. Nonetheless, the relatively high correlations that such formulas can achieve suggest that they have useful roles to play in identifying the extent to which texts are comprehensible. In a normal population sample consisting of people with varying degrees of reading competence, a text that scores better on a valid first-generation readability formula would generally be understandable (in real-world testing) to a larger segment of the population than a text that scores poorly using the same formula. This generalization will not hold true in all cases – so other methods for testing comprehension should also be used if a more accurate measurement of reading comprehension is needed for a specific text. A readability formula will also likely fail to isolate the exact reasons why people tend to understand some texts better than they understand other texts – so users of the formula should not rely on it for such purposes. However, as the above discussion illustrates, a readability formula that is used for the purposes of obtaining a general sense of the extent to which a document will be understood by an audience, and with an appreciation for the many limitations that are associated with the readability formula, can provide us with a useful relative assessment of how likely a text is to be understood by readers.

---

[53] Flesch, "A New Readability Yardstick", *supra* note 35 at 222-223.

2.4 Second-Generation Readability Formulas: Leveraging Computers

Some of the other criticisms of quantitative readability formulas like the ones advanced by Bailin & Grafstein could further be mitigated through the use of more sophisticated technological tools – tools that have continued to emerge and mature since the publication of Bailin & Grafstein's manuscript. Since the start of the 21st Century, with advances in computer and Natural Language Processing (NLP) technology, several second-generation sophisticated readability measures have emerged. In 2004, for instance, the Coh-Metrix 1.0 software-based text assessment tool was introduced. The creators of this tool noted that, "[a]s a consequence of these advances in scientific research and language technologies, we can go some distance in automating many of the deeper and global levels of text and language analysis."[54] Thus, instead of just looking at text features related only to syntax and vocabulary, the Coh-Metrix tool measured elements of "texts on multiple levels of language, discourse, cohesion, and world knowledge."[55]

In addition to computing more traditional readability scores (perhaps more reliably, and certainly more quickly, than if such scores were computed by hand), the Coh-Metrix tool also calculated lexical diversity,[56] frequency,[57] concreteness,[58] and cohesion scores.[59] In many ways, the cohesion measurements represented the most original new element to be included within second-generation quantitative readability assessments. The creators of the Coh-Metrix tool explain how cohesion ("an objective property of the explicit language and text")[60] relates to

---

[54] Arthur C Graesser, et al, "Coh-Metrix: Analysis of Text on Cohesion and Language" (2004) 36 Behavior Research Methods, Instruments, & Computers 193 at 193.
[55] Graesser, et al, "Coh-Metrix: Analysis of Text on Cohesion and Language", *supra* note 54 at 201.
[56] Graesser, et al, "Coh-Metrix: Analysis of Text on Cohesion and Language", *supra* note 54 at 198.
[57] Graesser, et al, "Coh-Metrix: Analysis of Text on Cohesion and Language", *supra* note 54 at 197.
[58] Graesser, et al, "Coh-Metrix: Analysis of Text on Cohesion and Language", *supra* note 54 at 196.
[59] Graesser, et al, "Coh-Metrix: Analysis of Text on Cohesion and Language", *supra* note 54 at 199-201.
[60] Graesser, et al, "Coh-Metrix: Analysis of Text on Cohesion and Language", *supra* note 54 at 193.

coherence ("a characteristic of the reader's mental representation of the text content" and "an achievement that is a product of psychological representations and processes").[61] These scholars suggest that cohesion exists in "explicit features, words, phrases, or sentences that guide the reader in interpreting the substantive ideas in the text, in connecting ideas with other ideas, and in connecting ideas to higher level global units" within a given text.[62] Coherence, in contrast, exists in the mind of a reader, rather than in the text itself: "[a] reader perceives a text to be coherent to the extent that the ideas conveyed in the text hang together in a meaningful and organized manner."[63]

The creators of the Coh-Metrix tool noted that greater cohesion within a text tends to benefit readers who have less background knowledge, while lower cohesion within a text tends to benefit readers with more background knowledge (since these readers will use their powers of inference to fill in cohesion gaps).[64] This phenomenon has been called the *reverse cohesion effect*,[65] and is believed to operate as follows: "the high-knowledge readers […] were able to gain from low-cohesion text because it forced them to generate inferences, and that inferencing resulted in a better, or deeper, understanding of the text."[66] These inferences are important for high-knowledge readers: "the inference processes engaged to bridge the conceptual gaps within and between sentences in low-cohesion text result in greater integration between the text and prior knowledge. Thus, low cohesion induces high-knowledge readers to integrate their knowledge with

---

[61] Graesser, et al, "Coh-Metrix: Analysis of Text on Cohesion and Language", *supra* note 54 at 193-94.
[62] Graesser, et al, "Coh-Metrix: Analysis of Text on Cohesion and Language", *supra* note 54 at 193.
[63] Graesser, et al, "Coh-Metrix: Analysis of Text on Cohesion and Language", *supra* note 54 at 193.
[64] Graesser, et al, "Coh-Metrix: Analysis of Text on Cohesion and Language", *supra* note 54 at 194.
[65] Danielle S McNamara, et al, *Automated Evaluation of Text and Discourse with Coh-Metrix* (New York, Cambridge UP, 2014) at 27.
[66] Danielle S McNamara, et al, *Automated Evaluation of Text and Discourse with Coh-Metrix*, *supra* note 65 at 32.

the text, and that process enhances comprehension."[67] However, the reverse cohesion effect is not always identifiable in studies that manipulate cohesion levels in order to assess the relationship between cohesion and comprehension.[68]

In other words, cohesion may have positive or negative effects on text comprehension, depending on the audience. Regardless, the authors suggest, older readability formulas may be unreliable for their failures to take cohesion into account:

> two-parameter multiple regression equations [like the Flesch Reading Ease score] will not go the distance in explaining text difficulty. Even worse, they will end up being misused. Textbook writers are known to shorten sentences in basal readers for the purpose of downsizing the grade levels of their texts. The unfortunate liability of shortening sentences is that the texts end up having lower cohesion and coherence.[69]

Since 2004, the creators have continued to update their tool. The most recent version, Coh-Metrix 3.0, includes a Text Ease and Readability Assessor that can be used to score texts based on five headings: narrativity; syntactic simplicity; word concreteness; referential cohesion; and, deep cohesion.[70]

Although one could continue to debate the extent to which any quantitative readability tool, like the Coh-Metrix tool, can accurately predict a specific reader's ability to understand a text, the more recent second-generation approach to measuring readability seems to be an improvement on earlier approaches, in that it better corresponds with our intuitive understanding of what makes a text understandable. And, as I note below, in section 3.4.4 of this dissertation, the Coh-Metrix readability formula (CML2RI) and another second-generation readability formula (CAREC-M)

---

[67] Tenaha O'Reilly & Danielle S McNamara, "Reversing the Reverse Cohesion Effect: Good Texts Can Be Better for Strategic, High-Knowledge Readers" (2007) 43:2 Discourse Processes 121 at 124 ["Reversing the Reverse Cohesion Effect"].
[68] See O'Reilly & McNamara, "Reversing the Reverse Cohesion Effect", *supra* note 67 at 141-142 (wherein the authors note, and attempt to reconcile, the findings from different studies – including their own – on this point).
[69] Graesser, et al, "Coh-Metrix: Analysis of Text on Cohesion and Language", *supra* note 54 at 194.
[70] Danielle S McNamara, et al, *Automated Evaluation of Text and Discourse with Coh-Metrix*, *supra* note 65 at 76-77 and 84-95.

both empirically outperformed first-generation formulas (Flesch-Kincaid Grade Level, Automated Readability Index, and SMOG) in terms of predicting the actual readability scores that were assigned to adjudicative decision texts by human experts.

Over the last twenty years, other second-generation readability formulas that leverage NLP technology and that require more complex calculations than older 'classic' readability formulas have continued to develop.[71] The Crowdsourced Algorithm of Reading Comprehension (CAREC) (which incorporates range and frequency scores for words and multi-word phrases; concreteness scores; imageability scores; age of acquisition scores, lexical diversity scores; temporal and thematic cohesiveness measures, and sentiment scores, among other things, into its readability calculation),[72] in particular, represents a major leap forward in terms of its sophistication,[73] and the diversity and precision of its constituent components. As the creators of the formula (and its counterpart formula, the Crowdsourced Algorithm of Reading Speed) have observed, "[t]hese formulas, based on linguistic features that better represent theoretical and behavioural accounts of the reading process, significantly outperformed classic readability formulas."[74] Empirically, the addition of these new "non-classic" features within the CAREC readability formula would seem to increase the explanatory power of the formula: one study that compared readability

---

[71] See, for instance, 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" (2020) 13:4 Journal of Communication in Healthcare 344 (wherein the authors quantitatively studied the linguistic properties of 724 secure message 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" (2020) 57:1 International Journal of Business Communication 52 (wherein the authors relied upon measures of lexical density, subordinate clause use, and passive voice use, alongside other classic readability formulas, to assess the readability of private-sector business reports).

[72] Scott A Crossley, Stephen Skalicky & Mihai Dascalu, "Moving Beyond Classic Readability Formulas: New Methods and New Models" (2019) 42 Journal of Research in Reading 541 ["Moving Beyond Classic Readability Formulas"].

[73] Scott Crossley, et al, "A Large-Scaled Corpus for Assessing Text Readability" (2022) Behaviour Research Methods, online: <https://link.springer.com/article/10.3758/s13428-022-01802-x>.

[74] Crossley, Skalicky & Dascalu, "Moving Beyond Classic Readability Formulas", *supra* note 72 at 557.

measurements using only classic (surface-level) linguistic features with readability measurements using a wider variety of both classic and non-classic linguistic features found that the latter group significantly outperformed the former group.[75]

Although classic first-generation readability formulas were (by design) capable of being used without the assistance of computers (i.e.: by hand-counting the relevant linguistic properties within a short sample of text, and by then inputting these properties into a simple equation), the CAREC formula cannot function without computer assistance: the formula requires both the processing power of a modern computer, and the networked access to remote databases/corpora that only a modern computer can realistically provide. In this sense, the CAREC readability formula falls squarely within the group that I refer to as second-generation readability formulas – ones that rely upon computational linguistic techniques.[76]

2.5 THIRD-GENERATION READABILITY: A.I. AND MACHINE LEARNING

Some third-generation readability measurements (although not necessarily formulas) that use machine learning are also now in various stages of relative infancy.[77] These approaches that rely on machine learning and artificial intelligence (AI) processes appear to be motivated by

---

[75] See Thomas François & Eleni Miltsakaki, "Do NLP and machine learning improve traditional readability formulas?" in *Proceedings of the First Workshop on Predicting and Improving Text Readability for Target Reader Populations* (Montreal: Association for Computational Linguistics, 2012) 49 at 56 ["Do NLP and machine learning"]: "Our results indicate that classic features are strong single predictors of readability. While we were not able to show that the non-classic features are better predictors by themselves, our findings show that leaving out non-classic features has a significant negative impact on the performance. The best performance was obtained when both classic and non-classic features were used."

[76] The goals of computational linguistics include "the discovery of processing techniques and learning principles that exploit both the structural and distributional (statistical) properties of language; and the development of cognitively and neuroscientifically plausible computational models of how language processing and learning might occur in the brain": see Lenhart Schubert, "Computational Linguistics" in Edward N Zalta, ed, *Stanford Encyclopedia of Philosophy* (Stanford: Metaphysics Research Lab, 2020) at 3-4.

[77] See, e.g., Muralidhar Pantula & KS Kuppusamy, "A Machine Learning-Based Model to Evaluate Readability and Assess Grade Level for the Web Pages" (2020) 65:4 The Computer Journal 831. I suggest that this example of machine learning for readability purposes is in its infancy because its ambitions are modest (classifying texts into only three categories of "readable", "less readable" and "more readable") and because it bootstraps its method to a classic readability formula (the Flesch-Kincaid Grade Level formula).

concerns about the statistical technique of linear regression that is central to the development of both first- and second-generation readability formulas.[78] In particular, some machine learning advocates question the validity of linear regression calculations when the data that is often used in these calculations do not conform to assumptions (e.g.: about linearity, the normality of the data's distribution, and the absence of significant outliers within the data)[79] upon which the linear regression model is predicated. Other machine learning proponents note that classic and second generation regression formulas are "not suitable for short text applications, which are more widespread in web and social media nowadays."[80]

If machine learning or AI approaches to assessing readability can outperform the approaches that are embodied in classic/first-generation and second-generation formulas, then these machine learning approaches probably represent the most promising future for readability researchers. For the time being, however, machine learning approaches involve several severe limitations. For instance, the neural networks that are used in machine learning processes "use more processing power and require specialized hardware to be trained and used and need massive datasets to achieve superior accuracies."[81] These tools that AI systems depend upon also generate significant environmental demands, in the forms of fossil fuel-generated electricity use, large-scale water consumption, the mining of scarce minerals, and many other similar costs to the

---

[78] Zhiquing Lin, "A Methodological Review of Machine Learning in Applied Linguistics" (2021) 14:1 English Language Teaching 74.
[79] Lin, "A Methodological Review of Machine Learning in Applied Linguistics", *supra* note 78 at 75-6.
[80] Hamid Mohammadi & Seyed Hossein Khasteh, "Text as Environment: A Deep Reinforcement Learning Text Readability Assessment Model" (2019) at 2, online: <https://arxiv.org/pdf/1912.05957.pdf> ["Text as Environment].
[81] Mohammadi & Khasteh, "Text as Environment", *supra* note 80. The authors go on to explain (at 7) how "the use of deep architecture in the proposed [machine learning] model has increased the required memory and computational resources. To be able to train the DRL model, resources such as Graphical Processing Units (GPU) are required, while the previous models are both light-weighted and swiftly trainable on average CPUs."

environment.[82] Others have pointed out how elements of machine learning processes are unhelpfully opaque, and resist easy and transparent explanation by researchers, or by those who wish to interrogate the reported results of a machine learning study.[83]

And, in any case, there is not yet any clear indication that machine learning approaches consistently outperform other second-generation approaches, even if they do outperform classic readability formulas.[84] As a preliminary observation, François & Miltsakaki note how difficult it can be to compare performance between different readability approaches "due to a lack of a common gold standard."[85] These authors review the conflicting studies about whether machine learning approaches outperform other readability formulas,[86] and find that "there is sufficient evidence that the AI paradigm outperforms the **classis** [sic] formulas."[87]

However, the authors call attention to the role that linguistic feature selection might have on the performance of any readability approach: "AI readability systems use NLP for richer feature extraction and a machine learning algorithm. Given that the classic formulas are also statistical, is performance boosted because of the addition of NLP-enabled feature extraction or by better machine learning algorithms?"[88] The authors are also sensitive to the possibility that the choice of a statistical model that is used within a readability approach, rather than any other inherent superiority of AI systems, might explain why some approaches perform better than others: "[a]n

---

[82] See, e.g., Kate Crawford, *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence* (New Haven: Yale UP, 2021) at 41-46 (for a compelling account of the many ways in which supposedly "clean" technology is having vast harmful consequences for the natural environment).
[83] Lin, "A Methodological Review of Machine Learning in Applied Linguistics", *supra* note 78 at 76.
[84] See generally, François & Miltsakaki, "Do NLP and machine learning", *supra* note 75.
[85] François & Miltsakaki, "Do NLP and machine learning", *supra* note 75 at 50.
[86] François & Miltsakaki, "Do NLP and machine learning", *supra* note 75 at 50-1.
[87] François & Miltsakaki, "Do NLP and machine learning", *supra* note 75 at 51 (emphasis added to highlight the unresolved question of whether machine learning approaches outperform more recent second generation - computational linguistic – readability formulas).
[88] François & Miltsakaki, "Do NLP and machine learning", *supra* note 75 at 50.

alternative explanation could be that, by comparison to the simpler statistical analyses that determined the coefficients of the classic formulas, machine learning algorithms, such as support machine vector (SVM) or logistic regression are more sophisticated and better able to learn the regularities in training data, thus building more accurate models."[89] In other words, François & Miltsakaki interrogate whether AI readability approaches are inherently superior to other approaches, or whether performance variances can be explained by reference to other factors that are not necessarily linked to a machine learning approach, such as linguistic feature selection and statistical modeling techniques.

Ultimately, these researchers found that readability approaches using both classical and non-classical linguistic features in tandem generated the best performance.[90] This finding is relevant to both machine learning approaches and to second-generation (computational linguistics) approaches, since both generally rely upon a variety of classic and non-classic linguistic features in producing readability assessments;[91] however, the finding does not suggest that either machine learning or computational approaches are inherently superior. Similarly, François & Miltsakaki found that linear regression (used in both classic and second-generation readability formulas) and support machine vector (the analogous statistical technique that is commonly used in machine learning readability approaches) performed almost equally "in accounting for the variance of text difficulty."[92] Again, this finding suggests that neither third-generation machine learning nor

---

[89] François & Miltsakaki, "Do NLP and machine learning", *supra* note 75 at 51.
[90] François & Miltsakaki, "Do NLP and machine learning", *supra* note 75 at 56.
[91] Compare, for instance, the machine learning approach in Renu Balyan, Kathryn S McCarthy & Danielle S McNamara, "Applying Natural Language Processing and Hierarchical Machine Learning Approaches to Text Difficulty Classification" (2020) 30 International Journal of Artificial Intelligence in Education 337 with the computational linguistics approach used in Crossley, Skalicky & Dascalu, "Moving Beyond Classic Readability Formulas", *supra* note 72: both approaches rely on non-classic linguistic features related to imageability and age of acquisition for content words, in addition to classic features related to word length.
[92] François & Miltsakaki, "Do NLP and machine learning", *supra* note 75 at 55.

second-generation computational linguistic approaches to readability assessments are inherently superior.

As all of the above context indicates, machine learning likely represents the next bound forward in readability research. However, the drawbacks of opacity, environmental impacts, and processing power demands that are currently associated with machine learning approaches are substantial. Until these challenges with machine learning approaches can be addressed – in ways that allow a wider community of scholars and stakeholders to access the knowledge and tools that machine learning requires – readability scholars will continue to have good reasons to develop new second-generation readability measurements. Such second-generation approaches are built upon a well-accepted foundation in applied linguistics that dates back over a century, while still taking advantage of automated parsing, tagging, and counting features that networked personal computers can perform – so they seem to represent a reliable middle ground between accepted and emerging approaches to the problem of measuring readability.

2.6 Readability Critics and Limitations of the Quantitative Approach

Despite advances in readability measurement approaches in recent years, several criticisms of readability measurements remain. Davison & Kantor, for instance, caution against the idea of re-writing texts to simplify them solely on the basis of early readability formulas' criteria (e.g.: by using shorter sentences or words).[93] Through an empirical study, they found that such efforts at simplification did not necessarily improve actual reading comprehension, since much more goes into the concept of comprehension than basic word/sentence length factors.[94] This finding is

---

[93] Alice Davison & Robert N Kantor, "On the Failure of Readability Formulas to Define Readable Texts: A Case Study from Adaptations" (1982) 17:2 Reading Research Quarterly 187 ["On the Failure of Readability Formulas"].

[94] Davison & Kantor, "On the Failure of Readability Formulas", *supra* note 93 at 207:

> [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

echoed by others: "readability formulas are useful for measuring the factors *reflecting* a text's readability. When adapting texts, it was more important that adaptors were paying attention to the discourse structures of sentences, paragraphs, and the whole text rather than simply trying to shorten words and sentences."[95] The Government of Canada's "Canada.ca Content Style Guide" – a document that contains "the rules to create web content" for government web pages – makes a similar point: "[d]on't rely on readability tools. Readability tools are based on algorithms that measure things like word length. These tools cannot tell you if your audience will find your content easy to understand and use. Readability tools can be a helpful first step to identify some areas that need work […]. They can also […] give you a rough sense of the complexity of a text."[96]

Other studies have affirmed the theoretical reasoning about the multi-dimensional nature of the reading comprehension process, by assessing the extent to which the visual layout and presentation of a text facilitates reader processing of that text[97] – and a more complete discussion about how non-linguistic factors like layout might affect ultimate comprehension is included in the following section of this dissertation. But with increased acceptance of the extent to which non-linguistic factors affect a person's ultimate comprehension of a written message, there is likely a corresponding reduction in the perceived power of purely linguistic readability formulas to predict how well readers might understand a document.

> 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 […].

[95] Rebekah George Benjamin, "Reconstructing Readability: Recent Developments and Recommendations in the Analysis of Text Difficulty" (2012) 24 Educational Psychology Review 63 at 64 (emphasis in original) ["Reconstructing Readability"].

[96] Canada, "Canada.ca Content Style Guide" (6 Aug 2024), online: <https://design.canada.ca/style-guide>, at s. 2.1.

[97] See, e.g., Mark Sableman, "Typographic Legibility: Delivering Your Message Effectively" (2017) 17 Scribes Journal of Legal Writing 9 at 15-17; Mary Alton Mackey & Marilyn Metz, "Ease of Reading of Mandatory Information on Canadian Food Product Labels" (2009) 33 International Journal of Consumer Studies 369 at 371-72; Khaled Moustafa, "Improving PDF Readability of Scientific Papers on Computer Screens" (2016) 35 Behaviour & Information Technology 319.

More recently, other authors have observed that there may be reason to suspect the validity of the benchmarks against which quantitative readability formulas are normed, or the criterion variables – such as grade levels that have been assigned to texts (potentially by the publishers) for other purposes, or by particular scores that readers would need to achieve on Cloze[98] tests to signify that they understood the texts.[99] As Cunningham, Hiebert & Mesmer correctly note, "[i]f the criterion lacks validity as a measure of the difficulty of sample texts for readers [then] it does not matter how well the [readability] equation predicts that criterion."[100] Kintsch & Vipond point out how many of the different approaches to testing reader comprehension may not truly be measuring comprehension – a fact that would undermine any readability formula that is based on these comprehension benchmarks.[101] They suggest, for instance, that asking readers to fill in a blank space with the missing word in a sentence is more about memory than comprehension: "[t]he cloze procedure […] is probably actually misleading. It measures the statistical redundancy of a text, which is a far cry from its comprehensibility."[102] Even typical reading comprehension test questions are not necessarily useful measurement's of a person's understanding of a text.[103] So, if the formulas are based on flawed measurements of human comprehension because the

---

[98] Cloze tests are reading comprehension tests in which every fifth word is replaced by an underlined blank space that the test subject is asked to fill in: see John R Bormuth, "Cloze Test Readability: Criterion Reference Scores" (1968) 5:3 Journal of Educational Measurement 189 at 189.

[99] James W Cunningham, Elfrieda H Hiebert & Heidi Anne Mesmer, "Investigating the Validity of Two Widely Used Quantitative Text Tools" (2018) 31 Reading and Writing 813.

[100] Cunningham, Hiebert & Mesmer, "Investigating the Validity of Two Widely Used Quantitative Text Tools", *supra* note 99 at 815.

[101] Kintsch & Vipond, "Reading Comprehension", *supra* note 2 at 337-338.

[102] Kintsch & Vipond, "Reading Comprehension", *supra* note 2 at 337.

[103] Kintsch & Vipond, "Reading Comprehension", *supra* note 2 at 338:

> It makes no sense to simply ask questions - any kind of question that occurs to the researcher - about a text. Much misuse has occurred in this way (e.g., when comprehension of a story is indexed by asking for some irrelevant detail). Asking questions is meaningful only if we have a fairly precise theory of text structure and text processing and if we know the role of the information tapped by our questions in terms of the total structure and process. Without such a theoretical understanding of a text, question answering is of dubious value. **Indeed, being able to answer some irrelevancy correctly may sometimes be a sign of not understanding rather than a point to be scored for comprehension**. (emphasis added)

"measurement of comprehension has historically been plagued by problems,"[104] then the formulas themselves will be flawed in predicting comprehension from linguistic features.

Another problem that manifests particularly in first-generation readability formulas is their lack of accounting for whole-text level features like organization and structure. Kintsch & Vipond note that these "higher-level features of text organization are probably the most serious omission from readability formulas."[105] They attribute this problem to the relatively short reading comprehension passages that were used as the baseline texts for measuring human comprehension in many early readability formula development efforts: "the McCall-Crabbs Test Lessons, the prototypic criterion, were too short to reveal such factors even if they had been there. In this sense there was from the beginning a bias against including higher-level features in the formulas."[106] A similar criticism of readability formulas appeared in a relatively recent Government of Canada blog post: "formulas don't measure the right things. They were not created for technical documents. They assume that short words are always the better words. They don't work with many documented features of plain language [like lists and headings]."[107] Although this type of criticism does not apply to all formulas, it does highlight the need for careful planning in the development of a baseline body of texts to be used for formula development, to ensure that appropriate linguistic properties that truly impact reading comprehension can be considered as potential formula variables.

---

[104] Steven J Amendum, Kristin Conradi & Elfrieda Hiebert, "Does Text Complexity Matter in the Elementary Grades? A Research Synthesis of Text Difficulty and Elementary Students' Reading Fluency and Comprehension" (2018) 30 Educational Psychology Review 121 at 144.
[105] Kintsch & Vipond, "Reading Comprehension", *supra* note 2 at 336.
[106] Kintsch & Vipond, "Reading Comprehension", *supra* note 2 at 336.
[107] Canada, "Readability Formulas, Programs and Tools: Do They Work for Plain Language?" (25 Jul 2022) *Our Languages Blog*, online: <https://www.noslangues-ourlanguages.gc.ca/en/blogue-blog/readability-formulas-eng>.

As a last criticism, Bailin & Grafstein have also identified what they refer to as the "increment issue" with readability formulas: "[c]ounts of formal properties do not translate into units of reading difficulty. If one text has an average sentence length of ten words and another of 15 words, this does not correlate to a difference of some function of five units' difference of difficulty"[108] – because reading comprehension is not affected in a linear fashion by any one variable. In other words, even though readability formulas purport to be quantitative, it is not clear that they are always coherently scaled to represent actual differences between text complexity levels.

Even though quantitative readability approaches clearly have some limitations, they are still used – perhaps now more than ever, as a result of freely-available web-based tools – to measure the complexity of texts in a broad range of fields,[109] for different private[110] and governmental[111] purposes. The Flesch Reading Ease formula also continues to be used within popular word processing software applications, such as Microsoft Word and Google Docs.[112] Thus, while quantitative readability studies cannot provide perfect truth on questions relating to the comprehensibility of a text, they can offer some useful predictors – particularly in terms of comparative assessments of readability levels between different texts, or text sets.

---

[108] Bailin & Grafstein, *Readability*, *supra* note 44 at 53-54.

[109] See, e.g., Matthew R Edmunds, Robert J Barry & Alastair K Denniston, "Readability Assessment of Online Ophthalmic Patient Information" (2013) 131:12 JAMA Ophthalmology 1610 (health care); see also, George R Milne, Mary J Culnan & Henry Greene, "A Longitudinal Assessment of Online Privacy Notice Readability" (2006) 25:2 Journal of Public Policy & Marketing 238 (consumer privacy); and see, Scott W Davis, et al, "Say What? How the Interplay of Tweet Readability and Brand Hedonism Affects Consumer Engagement" (2019) 100 Journal of Business Research 150 (social media marketing).

[110] See, e.g., Gene E Burton, "The Readability of Consumer-Oriented Bank Brochures: An Empirical Investigation" (1991) 30:1 Business & Society 21.

[111] See, e.g., Alexandre Deslongchamps, "Readability and the Bank of Canada," Bank of Canada Staff Analytical Note 2018-20, June 2018, online: <https://www.bankofcanada.ca/2018/06/staff-analytical-note-2018-20>.

[112] *See* Richard Johnson, "How to Apply the Flesch Kincaid Readability Formula to Your Content," February 23, 2021, online: <https://www.optimonk.com/how-to-apply-the-flesch-kincaid-readability-formula-to-your-content/> [https://perma.cc/6ZDM-DWND], for a description of how to access this readability formula in both Microsoft Word and Google Docs.

As the above review of general readability theory and practice illustrates, much academic literature supports the use of readability studies, with different formulas that each have their own strengths and weaknesses, for various purposes. In this respect, the theoretical foundation for my proposed dissertation is not, itself, original. However, as I demonstrate below (in Chapters 3-6), this dissertation incorporates significant originality with respect to: (1) my somewhat under-studied subject area for an application of readability theory (i.e.: Canadian administrative tribunal and court decisions);[113] (2) the new, law-specific formula that I identify to measure readability (a legal domain-specific readability measurement); and, (3) the results of my empirical readability study of different classes of decisions that showcase how my new legal domain-specific readability formula generates novel and important insights about the ways in which decisions are written.

2.7 Beyond Language Features: Other Influences on Comprehension

Although the above discussion has focused narrowly on readability (that is, on measures of how understandable the language of a text might be to a reader), there are a number of other important and related concepts that also contribute in different ways to the burdens that texts impose upon readers. Because many of these burdens could, under certain circumstances, ultimately be expected to have an impact on how well a reader understands written material – in ways that cannot easily be captured by linguistically derived readability formulas – it is worthwhile to discuss some of these burdens briefly here.

2.7.1 Text Length

[113] See above, at section 1.5 of this dissertation (wherein I conclude, after having reviewed the relevant literature, that there has not yet been any comprehensive study of the readability of Canadian administrative tribunal or court decisions, and that other readability studies undertaken in respect of non-Canadian decisions have tended to be jurisdictionally narrow while relying on older – and less theoretically sound – readability metrics).

Text length is perhaps the most obvious characteristic that might influence ultimate comprehension of a text, notwithstanding the linguistic readability of that text. Even without delving into research about the speeds at which people generally read documents,[114] we know intuitively that it takes longer to read 10 pages of text than it does to read 1 page of text, other things being equal. In other words, a longer text imposes a greater time burden on its readers than a shorter text,[115] even if the language within the two texts are identically "readable."[116] It therefore stands to reason that text length might have an impact on ultimate comprehension of a document.

Within certain limits, one might expect a motivated reader to be equally capable of understanding shorter or longer texts that are equivalently readable, because non-linguistic factors like boredom, distraction, attention deficits, or reading efficiency losses that result from time breaks between reading sessions of a long document, are not necessarily going to have significant effects on a reader's ultimate understanding of the documents. To the extent that text length might affect ultimate comprehension, that effect would not necessarily be seen until a certain threshold length differential between a shorter and a longer text is reached.

Indeed, this theory about text-length's lack of effect on comprehension across smaller text-length differences is borne out by multiple empirical studies. For instance, Mehrpour & Riazi conducted a study that used both original and shortened reading comprehension text passages to assess whether text length affected comprehension levels of the 100 advanced students who read

---

[114] Marc Brysbaert, "How Many Words do we Read per Minute? A Review and Meta-analysis of Reading Rate" (2019) 109 Journal of Memory and Language 104047, at 21, suggests that adult readers generally read nonfiction (silently, to themselves) at between 175-300 words per minute.

[115] Quinten Steenhuis, Bryce Willey & David Colarusso, "Beyond Readability with RateMyPDF: A Combined Rule-based and Machine Learning Approach to Improving Court Forms," in *Proceedings of International Conference on Artificial Intelligence and Law (ICAIL 2023)* (Association for Computing Machinery: New York, 2023), online: <https://doi.org/10.1145/3594536.3595146>, at s 2.3.1 ["Beyond Readability with RateMyPDF"].

[116] I use the term "readable" here, as I use it throughout this dissertation, to mean, "capable of being understood by a reader based on the language within the text."

and answered questions about the passages.[117] These authors found that "there was no statistically significant difference between the performances of the students on the two versions of the test" – despite the fact that the shortened versions of each passage were slightly less readable than the original versions (based on the Flesch Reading Ease formula).[118] However, even though the shortened text passages in this study were approximately 2/3 shorter than the original passages, the originals were also extremely short (at 346, 333, and 371 words, respectively)[119] when compared to a typical court decision. So, text length does not seem to affect ultimate comprehension within narrow ranges at the shorter end of the text-length spectrum.

In Kristen Beach's doctoral dissertation,[120] however, a similar finding emerged across much greater text-length differences. Beach found that "there was no statistically significant difference between those reading a 600-word passage and those reading a 2,000-word passage in terms of their reading comprehension scores."[121] These 600- to 2000-word text lengths are much closer to representing the text lengths of different judicial decisions than the texts that were used in Mehrpour & Riazi's study[122] – although the average decision length for adjudicative decisions within the real-world 1621-file corpus that is described and reported on below in Chapter 6 was 5752 words in length, and these decisions ranged between 408 and 110404 words in length.[123] So, Beach's research indicates that text length is not likely to affect ultimate comprehension across a

---

[117] Saeed Mehrpour & Abdolmehdi Riazi, "The Impact of Text Length on EFL Students' Reading Comprehension" (2004) 6:3 Asian EFL Journal.
[118] Mehrpour & Riazi, "The Impact of Text Length on EFL Students' Reading Comprehension", *supra* note 117.
[119] Mehrpour & Riazi, "The Impact of Text Length on EFL Students' Reading Comprehension", *supra* note 117.
[120] Kristen L Beach, *The Effect of Media, Text Length, and Reading Rates on College Student Reading Comprehension Levels*, (unpublished Doctor of Business Administration Dissertation, University of Phoenix, 2008) [*The Effect of Media*].
[121] Beach, *The Effect of Media*, *supra* note 120 at 67.
[122] Mehrpour & Riazi, "The Impact of Text Length on EFL Students' Reading Comprehension", *supra* note 117.
[123] These word length results, alongside readability results and results for several other variables, are all available within the xlsx spreadsheet file that is publicly available online: <https://doi.org/10.7910/DVN/NIQ2KJ>. A much more complete discussion of the results is included below, in Chapter 6.

wide range of adjudicate decision-lengths, but her research does not tell us whether comprehension levels would drop off, for instance, as one makes their way through the epic length of Justice Wooley's 110,404-word decision in *Remington Development Corporation v Canadian Pacific Railway Company*,[124] a real property case that represents the longest decision for which readability scores are calculated in subsequent chapters of this dissertation.

As the above discussion suggests, there is no existing research that allows us to definitively conclude how text length might affect ultimate comprehension of judicial decisions across very wide text-length spectrums (e.g.: for texts that differ in length by 3 or 4 orders of magnitude). We may continue to believe that comprehension would be adversely affected by massive text-length differences, and future research may eventually support this hypothesis. But this question does not need to be answered for the purposes of my dissertation. Wide text length differences may, or may not, have an impact on ultimate comprehension of texts, and none of the readability formulas that have been discussed in this chapter (nor any of the formulas that are employed later in this dissertation) would be reliably capable of measuring the effect that text length might have on ultimate comprehension – because each existing formula has been generated using texts of relatively homogenous lengths.[125] This is the key point to understand about text length: current quantitative readability formulas do not account for any potential impact that large text length variations might have on a reader's ability to understand different texts. And we have at least some

---

[124] 2022 ABKB 692, online: <https://www.canlii.org/en/ab/abkb/doc/2022/2022abkb692/2022abkb692.html>. Incidentally, Justice Woolley's readability score in this decision was somewhat more readable than the average score for superior court decisions, using *Madden's Adjudicative Decision Readability Score* (the MADRS formula) that I develop and explain in Chapter 3, and considering the results reported below in both Chapters 4 and 6.

[125] For instance, Crossley, Skalicky & Dascalu, "Moving Beyond Classic Readability Formulas", *supra* note 72 at 547, used Wikipedia articles that were all approximately 200 words long in building their CAREC readability formula. Similarly, the *McCall-Crabbs Standard Test Lessons in Reading* that formed the basis for the Flesch, Dale-Chall, and Gunning readability formulas were each only 150 words in length (approximately): Milton D Jacobson, C Eric Kirkland & Ramsay W Selden, "An Examination of the McCall-Crabbs Standard Test Lessons in Reading" (1978) 22:3 Journal of Reading 224 at 224-225.

theoretical bases for wondering whether very long texts are harder to understand than much shorter texts. Readability calculations like the existing formulas discussed in this chapter, and the formulas that I derive and/or test in Chapter 3 of this dissertation, do not help us to resolve this question. Moreover, these formulas may perform inaccurately in predicting ultimate comprehension (as opposed to just linguistic "readability") of very long texts if there truly is a length-effect associated with very long texts that pulls down reading comprehension levels for those texts.

While it would likely be possible to derive a readability formula from texts of vastly different lengths in order to assess whether text length has a statistically significant effect on a reader's ability to understand a text, no such study of which I am aware has yet set out to achieve such a task. I hypothesize that such studies have not yet been undertaken due to the practical difficulty (in terms of time commitment) that researchers would encounter when asking real-world readers to read and demonstrate their comprehension (or otherwise signal the readability) of very long texts. Where text length has been found not to have a statistically significant effect on reading comprehension scores across texts of substantially (but not massively) different lengths,[126] I did not find it necessary within the study that I conducted in Chapter 3 to develop a law-specific readability formula that could assess how text length might affect the perceived comprehensibility of an adjudicative decision text.

2.7.2 Text Display and Legibility

There is an abundance of research that explores the extent to which text display choices may affect both the legibility of a text, and also one's ability to understand the text. I use the phrase "text display choices" here to refer to what others might call "typography"[127] – that is, to several

---

[126] Beach, *The Effect of* Media, *supra* note 120 at 67.
[127] Joseph A Rosenberg, "Why Lawyers Should Write for the Digital Reader" (2023) 27 Journal of the Legal Writing Institute 111 at 133: "Typography determines the readability of a writing in the choices made about the visual

different font and layout-related characteristics of a text, including things like font style and size, white space, line spacing, and line length or column width. This research is filled with contradictions.

Soleimani & Mohammadi review much of the contradictory literature about how font choices (such as using serif or sans serif fonts, using 10-point or 12-point fonts, and using different font types) sometimes seem to affect or not affect reading speed and comprehension.[128] In their own empirical study, however, these authors find that size 12 font was read more quickly than size 10 font, but that font size had no statistically significant effect on reading comprehension.[129] They also found that neither font type (Arial or Old Bookman) nor line spacing (set solid,[130] or double) had a statistically significant effect on reading comprehension.[131] By the authors' own admissions, many of these findings supported similar findings from other studies, while also contradicting similar findings from still other studies.[132]

Other research suggests that obscure (to most of us) factors like the x-height of a particular font can also affect how well readers understand text that is written in that font.[133] The x-height refers to difference in height between lower case letters with either ascenders (like h, t, f) or descenders (like g, p, j), on the one hand, and lower case letters with neither ascenders nor

---

appearance of text. These choices may involve the distance between lines of text, letter spacing, alignment, hierarchical relationships, and typeface."

[128] Hassan Soleimani & Elham Mohammadi, "The Effect of Text Typographical Features on Legibility, Comprehension, and Retrieval of EFL Learners" (2012) 5:8 English Language Teaching 207 at 207-209 ["The Effect of Text Typographical Features"].

[129] Soleimani & Mohammadi, "The Effect of Text Typographical Features", *supra* note 128 at 211-213.

[130] "Set solid" line spacing is described elsewhere as "type set with leading equal to the point size (i.e., 12/12, 10/10, etc.) or, in other words, without any additional space between successive lines. One caveat about setting type solid is that the descenders of one line may touch—or appear to touch—the ascenders of the line beneath it" (PrintWiki: The Free Encyclopedia of Print, "Set Solid", online: <https://printwiki.org/Set_Solid>.

[131] Soleimani & Mohammadi, "The Effect of Text Typographical Features", *supra* note 128 at 212-214.

[132] Soleimani & Mohammadi, "The Effect of Text Typographical Features", *supra* note 128 at 214.

[133] Sableman, "Typographic Legibility: Delivering your Message Effectively", *supra* note 97 at 13-15.

descenders (like x, a, c) on the other hand.[134] Greater x-height will tend to increase the white space between successive lines of text, which may facilitate reading comprehension.

Still other research has shown that text written in all capital letters takes an average of 12-13% longer to read than text written in sentence case.[135] Underlined and italicized text also take longer to read (8-11%, and 4.5%, respectively) than text without emphasis, although bold text does not appear to hinder reading speed.[136] Reading speed is also apparently affected by line length or column width – with the optimum line length for text printed in 12-point font being somewhere between 2.75 and 4 inches.[137] This recommendation seems to apply more to printed text[138] than to text that will be read on-screen, however, and does not seem to apply differently depending on whether the printed text is found within a book or a sheet of computer-printed paper[139] (e.g.: an 8.5" x 11" sheet of paper). McMullin et al found, in their study of how individuals read text from web pages, "that very [short] line lengths (25 characters per line) slowed reading and led to less comprehension than relatively longer line lengths (55 and 100 characters per line). These convergent findings contradict the print literature."[140]

As the above discussion suggests, the layout and visual presentation details associated with a text are thought to have an impact on reading speed, and potentially on reading comprehension,

---

[134] Sableman, "Typographic Legibility: Delivering your Message Effectively", *supra* note 97 at 13-15.
[135] Ruth Anne Robbins, "Painting with Print: Incorporating Concepts of Typographic and Layout Design into the Text of Legal Writing Documents" (2004) 2 Journal of the Association of Legal Writing Directors 108 at 115 ["Painting with Print"].
[136] Robbins, "Painting with Print", *supra* note 135 at 118-119.
[137] Robbins, "Painting with Print", *supra* note 135 at 122-123. Reviewing several other studies, Anuj A Nanavati & Randolph G Bias, "Optimal Line Length in Reading – A Literature Review" (2005) 39:2 Visible Language 120 at 139, suggest that printed text should come in line lengths of between 2.3 and 5.5 inches.
[138] Jess McMullin, et al, "Effects of Surrounding Information and Line Length on Text Comprehension from the Web" (2002) 28:1 Canadian Journal of Learning and Technology 1 at 2 ["Effects of Surrounding Information"].
[139] Nanavati & Bias, "Optimal Line Length in Reading – A Literature Review", *supra* note 137 at 139. This article takes a very creative form, wherein the authors present their argument in gradually increasing or decreasing line lengths on each page.
[140] McMullin, et al, "Effects of Surrounding Information", *supra* note 138 at 8.

of the text. However, studies tend to diverge in their findings about what these impacts might be. Again, the point of raising this factor here is not for the purpose of resolving debates about how things like font choices or white space might affect ultimate comprehension of a text. Rather, the purpose of this discussion is merely to familiarize readers of this dissertation with the broad contours of the debates, and to distinguish between linguistic factors that might affect ultimate comprehension (and that can be well-captured within quantitative readability formulas that have historically isolated out the effects of language on comprehension) on the one hand, and other non-linguistic or presentation-related factors that could potentially affect ultimate comprehension (but that are likely not accounted for within existing quantitative readability formulas) on the other hand.

2.7.3 Print and Electronic Reading Media

The question of whether text will be better understood if it is presented and read on-screen as opposed to in print form is also heavily studied. From a theoretical perspective, there are reasons to believe that print reading will facilitate better comprehension: "screen texts lack effective contextual cues for readers to recall from memory and search for information,"[141] whereas print texts include contextual cues like "perceiving the thickness of a book based on senses of sight and touch, realizing the approximate location of information of the text in book pages, and measuring how much text remains to be read and how much has already been read."[142] Additionally, the almost-inevitable requirement to scroll when reading electronically can hinder comprehension: "[s]crolling is known to hamper the process of reading, by imposing a spatial instability which

[141] Chih-Ming Chen & Yu-Ju Lin, "Effects of Different Text Display Types on Reading Comprehension, Sustained Attention and Cognitive Load in Mobile Reading Contexts" (2016) 24:3 Interactive Learning Environments 553 at 554 ["Effects of Different Text Display Types"].
[142] Chen & Lin, "Effects of Different Text Display Types", *supra* note 141 at 554.

may negatively affect the reader's mental representation of the text and, by implication, comprehension."[143]

However, there are again conflicting studies about how comprehension is affected by reading on-screen compared to in print media. In their literature review of several prior studies that compare reading performance in print and electronic media, Delgado et al noted that results were inconsistent: some studies found that paper-based reading produced a small advantage, while other studies found that electronic reading produced a small advantage.[144] These authors then conducted their own meta-analysis of 54 studies, and concluded that screen reading was associated with a small reduction in reading comprehension when compared to print reading.[145] But a more recent meta-analysis of 32 studies conducted by Schwabe et al that focused on narrative texts instead of on more academic texts reached a different conclusion.[146] Schwabe et al did not find any statistically significant differences in reading comprehension rates between texts read electronically or in print form.[147] And, their "results suggest a small advantage of the narrative text's digital version compared to the printed text when multimedia / interactive functions were used,"[148] which may indicate that well-designed digital documents offer comprehension advantages over print equivalents.

Readers will begin to notice a pattern here: there is no clear winner in the contest between print and electronic media when it comes to facilitating reading comprehension. The media factor

---

[143] Anne Mangen, Bente R Walgermo & Kolbjørn Brønnick, "Reading Linear Texts on Paper Versus Computer Screen: Effects on Reading Comprehension" (2013) 58 International Journal of Educational Research 61 at 65.
[144] Pablo Delgado, et al, "Don't Throw Away Your Printed Books: A Meta-Analysis on the Effects of Reading Media on Reading Comprehension" (2018) 25 Educational Research Review 23 at 24-25 ["Don't Throw Away Your Printed Books"].
[145] Delgado, et al, "Don't Throw Away Your Printed Books", *supra* note 144 at 33-34.
[146] Annika Schwabe, et al, "No Negative Effects of Reading on Screen on Comprehension of Narrative Texts Compared to Print: A Meta-analysis" (2022) 25:6 Media Psychology 779 ["No Negative Effects"].
[147] Schwabe, et al, "No Negative Effects", *supra* note 146 at 791.
[148] Schwabe, et al, "No Negative Effects", *supra* note 146 at 791.

seems to be yet another non-linguistic variable that probably affects comprehension – but we do not seem to be able to pinpoint the direction and strength of this effect with consistency across different studies and meta-studies. It is sufficient for the purposes of this dissertation, however, to simply note that the media within which a text is presented may impact ultimate comprehension of the text, and that linguistic readability analyses like the ones contained in the remainder of this dissertation do not tend to capture any such impacts on comprehension.

2.7.4 Reading Comprehension and Listening Comprehension

Ultimate comprehension of a text might also be influenced by the different sensory processes that a person can use in receiving and processing the text. To that end, one might question whether the concept of readability – a concept that studies how textual inputs communicate a linguistic message – is only partially helpful in considering how to facilitate comprehension of the message. Perhaps it would be more productive to consider whether auditory substitutes for the process of reading adjudicative decisions would be more effective at communicating the contents of the decisions.

Older research comparing reading and listening comprehension approaches to communicating material has revealed that there is "no consistent data indicating a general superiority of one over the other as a method of learning."[149] In an empirical study that controlled for time (that is, where readers and listeners were presented with material in written or spoken form at exactly the same rate and for the same limited duration), it was found that comprehension

[149] Robert Q Young, "A Comparison of Reading and Listening Comprehension with Rate of Presentation Controlled" (1973) 21:3 AV Communication Review 327 at 327.

was essentially the same for both modalities (written or spoken), and that performance improved in proportion to time rather than in relation to the modality of the presentation.[150]

More recent research has investigated the extent to which college-educated adults understand material better or worse depending on whether it is presented to them in their preferred learning style (visual/written, or auditory/spoken).[151] This study was designed to interrogate the "widespread belief among educators and the general public alike that individuals learn better when they are presented instruction in the modality that capitalizes on their learning style preference."[152] Ultimately, the authors noted that "differences in preferred learning style (auditory, visual word) were not found to significantly predict differences in learning aptitude (listening vs. reading comprehension),"[153] but that people whose preferred learning style was visual tended to achieve better reading **and** listening comprehension levels than people with a preferred auditory learning style.[154] This research is relevant because it does not suggest that listening comprehension is superior or inferior to reading comprehension as a general matter, and because it suggests that providing people with material in the format that they prefer (written or spoken) will not necessarily have any impact on how well people understand the information in that material.

It might be more useful to know whether a categorical comprehension advantage can be achieved by providing people with information in either written or spoken form, regardless of their preferences. However, it is difficult to isolate the different factors that contribute to comprehension: "[w]ithout information on listening comprehension, it is difficult to determine the

---

[150] Young, "A Comparison of Reading and Listening Comprehension with Rate of Presentation Controlled", *supra* note 149 at 329-334.
[151] Beth A Rogowsky, Barbara M Calhoun & Paula Tallal, "Matching Learning Style to Instructional Method: Effects on Comprehension" (2015) 107:1 Journal of Educational Psychology 64 ["Matching Learning Style"].
[152] Rogowsky, Calhoun & Tallal, "Matching Learning Style", *supra* note 151 at 75.
[153] Rogowsky, Calhoun & Tallal, "Matching Learning Style", *supra* note 151 at 76.
[154] Rogowsky, Calhoun & Tallal, "Matching Learning Style", *supra* note 151 at 76.

cause of the poor reader's problem. Poor reading comprehension may reflect a decoding problem, a general comprehension problem, or a combination of both these factors."[155] In other words, a person's failure to understand a written message (if they have weak reading skills but otherwise strong language skills) could perhaps be rectified by simply reading the message aloud to that person. Listening comprehension (the ability to understand a spoken message) and reading comprehension levels are not necessarily identical for any particular person, so it may not be possible to identify the overall best modality for communicating information.

This idea that reading and listening comprehension skills develop at different rates is encapsulated within "dual comprehension process" theories of decoding language,[156] which suggest that there are "processing differences, linguistic differences, social-situational differences, contextual differences, and task differences that make comprehension under oral and written modes a more complex and multifarious process than was once envisioned"[157] under earlier "unitary comprehension process"[158] theories. Specifically, it has been suggested that spoken communications contain more contextual information that may be relevant to the decoding process: "[s]peech has rising and falling pitch, stress, and pauses and discontinuities. These prosodic[159] features of speech give expression to the voice but are not represented or are poorly represented in writing."[160] One might expect, with more contextual information from the spoken word than the written word, that spoken communications would facilitate greater general comprehension.

---

[155] Rosalind Horwitz & S Jay Samuels, "Reading and Listening to Expository Text" (1985) 17:3 Journal of Reading Behavior 185 at 185-186.
[156] Horwitz & Samuels, "Reading and Listening to Expository Text", *supra* note 155 at 187-188.
[157] Horwitz & Samuels, "Reading and Listening to Expository Text", *supra* note 155 at 187.
[158] Horwitz & Samuels, "Reading and Listening to Expository Text", *supra* note 155 at 186-187.
[159] Prosody refers to sound-related patterns in language, like rhythm and tone.
[160] Horwitz & Samuels, "Reading and Listening to Expository Text", *supra* note 155 at 187.

But there continue to be theoretical disagreements about whether "comprehension" is a single skill that does not fluctuate depending on the medium through which information is communicated (orally or in writing),[161] or whether there are modality distinctions that ought to be made between reading and listening comprehension skills.[162] Empirically – within a study of 2nd and 3rd grade students who were presented with written-only texts, or with written texts together with accompanying audio – it seems that there is some overlap between the reading and listening comprehension processes: "part of the reading and listening comprehension process taps a general comprehension skill that operates regardless of modality."[163] However, only 40% of the proportion of variance in individuals' reading comprehension skills could be explained by their listening comprehension skills, and only 34% of the proportion of variance in their listening comprehension skills could be explained by their reading comprehension skills.[164] This means that a majority of the variance in reading / listening skills cannot be explained by the opposite skill, so these skills are likely "modality-specific and cannot be subsumed under a general comprehension skill."[165]

If reading and listening comprehension skills are truly distinct, then one way for adjudicative bodies to better reach their audiences might involve providing decisions through both the written and spoken modalities. This could perhaps be done by having the judge or tribunal member read their decision aloud to the parties, but such an approach would not necessarily maximize the potential benefits of tapping into both modalities.

---

[161] MC Wolf, et al, "The Relationship Between Reading and Listening Comprehension: Shared and Modality-Specific Components" (2019) 32 Reading and Writing 1747 at 1748 ["The Relationship Between Reading and Listening Comprehension"].
[162] Wolf, et al, "The Relationship Between Reading and Listening Comprehension", *supra* note 161 at 1749-1750.
[163] Wolf, et al, "The Relationship Between Reading and Listening Comprehension", *supra* note 161 at 1761.
[164] Wolf, et al, "The Relationship Between Reading and Listening Comprehension", *supra* note 161 at 1756-1758.
[165] Wolf, et al, "The Relationship Between Reading and Listening Comprehension", *supra* note 161 at 1760.

To begin with, the advantages of leveraging listening comprehension skills seems to be most pronounced when written and spoken messages are paired together, such as when text-to-speech or other similar technologies are employed.[166] The benefits of using paired modalities may flow from the activation of visual memory when both written and spoken messages are received.[167] But benefits might also accrue because an accompanying written message compensates for challenges associated with a listening-only process.[168] Written material is static, and allows individuals to read at their own pace, and reread particular sections as often as needed; "[a]ural material, by contrast, is not static; it is presented at a rate controlled by the speaker rather than by the listener and is constantly 'moving' from beginning to end. Listeners have no opportunity, in most cases, to rehear parts of the message, but must learn as much of the material as possible in one presentation."[169] So, having a decision read aloud to the parties, at a pace that they cannot control, and where they do not also have the written decision in front of them, would not necessarily aid in comprehension. And even if there were benefits to this approach, they would only help the parties – not any other audiences for the decisions who might only encounter the texts in written form.

But using reading and listening techniques concurrently may noticeably improve general comprehension rates. One recent meta-analysis of studies that assessed how text-to-speech and read-aloud technologies might assist students with learning disabilities to understand written texts

---

[166] Sarah G Wood, et al, "Does Use of Text-to-Speech and Related Read-Aloud Tools Improve Reading Comprehension for Students With Reading Disabilities? A Meta-Analysis" (2018) 51:1 Journal of Learning Disabilities 73 ["Text-to-Speech and Related Read-Aloud Tools"].
[167] Wolf, et al, "The Relationship Between Reading and Listening Comprehension", *supra* note 161 at 1751.
[168] Wolf, et al, "The Relationship Between Reading and Listening Comprehension", *supra* note 161 at 1751. Disadvantages of a listening-only process include the "fleeting nature of the presentation of the information and the fact that comprehenders cannot control the pace of information themselves."
[169] Young, "A Comparison of Reading and Listening Comprehension with Rate of Presentation Controlled", *supra* note 149 at 328.

showed that there was a positive effect of moderate size associated with these technologies – regardless of the students' grade levels.[170] This research suggests that courts and tribunals may wish to consider how they could communicate their decisions in formats that can easily be used with text-to-speech or similar technologies. Such a step would allow those who access the decisions to leverage two different modalities for their understanding of a decision, while still offering them the time benefit that allows them to proceed through the decision at their own pace, and as often as they need to in order to understand it.

In any case, the above discussion is not intended to prescribe the modality or modalities that adjudicative bodies should use to communicate their decisions as much as it is intended to demonstrate how non-linguistic modality factors might contribute to ultimate comprehension of the decisions in ways that cannot be well-captured within readability formulas. To be sure, the linguistic content of a decision will theoretically have a significant impact on how well the decision is understood regardless of the modality that is used to communicate the decision. But linguistic content is not a complete answer to the question of how well one might understand a message, since other factors (in this case relating to differences between reading and listening comprehension skills) are also likely to have an impact on ultimate comprehension.

2.7.5 Reader-specific and Other Factors

In addition to all the text-, display-, medium-, and modality-specific factors that might affect ultimate comprehension of a written message, it is clear that other audience- or reader-specific variables will also influence comprehension. For instance, we intuitively appreciate that the average 6 year-old will not understand some language that will be easily understood by the

[170] Wood, et al, "Text-to-Speech and Related Read-Aloud Tools", *supra* note 166 at 81.

average 17 year-old. And, readability theorists have affirmed this intuition: "selecting appropriate texts for a population of readers requires some understanding of both the reader and the text, and different methods may be more or less appropriate for different types of texts and different populations of readers."[171] A brief discussion of some of the less obvious reader-specific influences on comprehension are included below for illustrative purposes.

Pishghadam & Abbasnejad argue that a reader's "emotioncy", or the extent to which a reader has experienced a word or concept with increasing levels of sensory or emotional contact (progressing from no experience, to auditory experience, to visual experience, to tactile experience, to internalized emotional experience, etc) will affect the reader's ability to understand language.[172] This theory accords with Steenhuis, Willey & Collarusso's theory (from another context) that the emotional burden of a text (for instance, a court form that requires individuals to recount their experiences of trauma in unnecessary detail) will have an impact on the overall complexity of that text from the reader's or user's perspective,[173] and with other research suggesting that the sentiment of a text factors into its readability.[174] Emotions seem to clearly factor into comprehension processes.

Pishghadam & Abbasnejad explain how they assess emotioncy using the formula "emotioncy = sense (frequency + emotion)" in their empirical study, using the following example. "[I]magine a student who has expressed his emotioncy for the word sea as follows: I have gone to sea and swum in it (sense score: 4) many times (frequency score: 5) and I feel extremely positive

---

[171] Benjamin, "Reconstructing Readability", *supra* note 95 at 64.
[172] Reza Pishghadam & Hannaneh Abbasnejad, "Emotioncy: A Potential Measure of Readability" (2016) 9:1 International Electronic Journal of Elementary Education 109 at 111-113.
[173] Steenhuis, Willey & Colarusso, "Beyond Readability with RateMyPDF", *supra* note 115.
[174] At least one recent readability study found that the use of positive adjectives was strongly and positively correlated with readability results, and relied upon this sentiment-related variable within the reading comprehension formula that the study produced: Crossley, Skalicky & Dascalu, "Moving Beyond Classic Readability Formulas", *supra* note 72 at 553 (Table 2).

about sea (emotion score: 5). His total emotioncy score would be 4(5+5) = 40, which shows that the student is completely involved with the notion of sea."[175]

These authors then contrasted classic readability formula scores for different texts with student-reported difficulty levels for the texts, and found that the differences between these difficulty levels could be explained by the students' emotioncy levels: "it is not only the cognitive connections which tie individuals to words in a passage, but also the emotional dependency on the text itself. That is to say, individuals' interests, their background knowledge, and motivation, noted among paramount factors disregarded in readability formulae, are the basic requirements of emotioncy which tries to approach readers' minds through their hearts."[176] As this study suggests, individual experiences and characteristics are highly relevant to questions about reading comprehension, even though they are not particularly well-represented in different quantitative readability formulas.

In other domains, it has been suggested that a multitude of contextual factors beyond just language contribute to comprehension. For instance, health care researchers have noted that "factors related to the document (layout, color, font, spacing, legibility, and grammar); person (education, comprehension, health literacy, motivation, prior knowledge, information needs, anxiety levels); and style of writing (cultural sensitivity, context, comprehensiveness, and appropriateness)" all affect a patient's ability to understand the document.[177] Motivation, in particular, seems to be a weighty factor. As Meade & Smith observe, "the individual with a kidney transplant who goes home with several medication instruction sheets may exhibit high motivation

---

[175] Pishghadam & Abbasnejad, "Emotioncy: A Potential Measure of Readability", *supra* note 172 at 114.
[176] Pishghadam & Abbasnejad, "Emotioncy: A Potential Measure of Readability", *supra* note 172 at 117.
[177] Pranay Jindal & Joy C MacDermid, "Assessing Reading Levels of Health Information: Uses and Limitations of Flesch Formula" (2017) 30:1 Education for Health 84 at 86.

and interest in the information because he/she knows that the correct administration of the medication means the difference between healthy living or organ rejection."[178] These authors suggest that motivation differences (and writing techniques that are designed to increase a reader's motivation) can play a key role in facilitating comprehension.[179]

As this discussion illustrates, there are many reader-specific factors that will influence how well a particular person understands a particular written text. It is also evident that authors of written documents have almost no control over these personal factors that are unique to individual readers. It is likely for these reasons (the almost infinite number of combined personal factors that make some people more or less likely to understand a document, and the inability of writers to produce texts that will be universally more understandable because of how they respond to all of the different personal factors) that quantitative readability studies have disregarded almost entirely the reader-centric perspective of reading comprehension processes, and have focused instead on the text-centric properties of documents that tend to affect how well the documents are understood.

2.7.6 Contemplating a more global "quantitative comprehension" formula

The above discussion within this section has surveyed several different considerations that could have an impact on a person's ability to understand a written message – ranging from considerations about the ways in which the message is presented visually, to ones about auditory

---

[178] Cathy D Meade & Cyrus F Smith, "Readability Formulas: Cautions and Criteria" (1991) 17 Patient Education and Counseling 153 at 156. For a similar observation, *see* Edgar Dale & Jeanne S Chall, "A Formula for Predicting Readability", *supra* note 37 at 20:

> To say that a given article on chemistry is comfortable reading for average adults because it has a predicted grade level of VII-VIII, is giving an incomplete picture. For those readers who have no interest or no background in chemistry, the article will probably not be comfortable reading and they may get very little meaning from it. For other readers who are interested in chemistry and do considerable reading in the subject, the same article will probably be most comfortable reading. This difference in ease of reading and comprehension may exist even though both groups of readers have completed approximately eight and one-half years of schooling and have the same general reading ability on a standardized reading test.

[179] Meade & Smith, "Readability Formulas: Cautions and Criteria", *supra* note 178 at 156.

supplements for a written message, to ones that are unique to each reader's mental state and background. Although I have stressed that linguistically derived readability formulas are not designed to, and do not, take into account these other considerations when predicting how well readers will understand a text, this does not mean that it would be impossible for more comprehensive "quantitative comprehension" formulas to be created. Below, I explain briefly how it might be possible to create such a formula. I also offer several hypotheses as to why no one has yet accomplished this task, and explain why it is beyond the scope of this dissertation to undertake such an effort.

Quantitative readability formulas are a type of comprehension formula – but they focus narrowly on one class of variables that relate to a reader's comprehension of a text: linguistic variables. There is no principled reason, however, why the overarching approach that is used to derive a linguistic readability formula could not also be used for a more wide-ranging comprehension formula.

As will be shown by example in Chapter 3 (and as noted above in this chapter), the typical method for creating a quantitative readability formula involves several steps. First, real-world comprehension scores are assigned by human subjects to a group of texts. Next, the differences between these texts, in terms of their scores on a range of variables (linguistic variables, specifically, for readability formulas) are measured. Correlations are identified between each variable and the real-world comprehension scores for a given text. The variables that are significantly and most strongly correlated with real-world comprehension scores are then processed through a statistical regression (or similar type of) analysis to create a weighted formula that is best capable of predicting the real-world scores for the texts. This formula is then generally

tested against an independent data set (of human-scored texts) to see how well the formula predicts the human-assigned scores.

Many of the non-linguistic variables that are discussed above – like font size, line length, overall text length, amount of white space on a page, and x-height – could each be quantified as scale variables that are capable of being assessed statistically in the same way that linguistic variables are assessed for the purposes of creating readability formulas. It would therefore be possible to measure these types of variables alongside other quantitative linguistic variables to verify whether any of the non-linguistic variables are significantly and strongly correlated with real-world comprehension scores for texts. And, if such correlations existed, then regression analysis (or a similar technique) could be used to produce a weighted predictive formula that might rely on these non-linguistic variables – a more comprehensive comprehension formula than simple linguistic readability formulas.

Other non-linguistic variables that do not lend themselves to quantification on a scaled basis could potentially still be assessed statistically, by coding them using "dummy" numerical values (e.g.: 0 = decision was not read aloud; 1 = decision was read aloud at the speaker's pace; 2 = decision was read aloud at the listener/reader's pace, etc). It is possible that variables like this could then be factored into the process of statistically computing a more global quantitative comprehension formula.

Where there are sound theoretical reasons for believing that some or all of these types of non-linguistic variables could have an impact on a person's ultimate ability to understand a written message, it seems obvious that it would be beneficial to measure the variables, and to verify our theoretical hypothesis that they could have the anticipated effects. But no such undertaking – that

simultaneously and empirically considers the impacts of both linguistic and non-linguistic variables on ultimate comprehension of a written message – has yet been published.

I hypothesize that there are several plausible reasons why a more global quantitative comprehension formula has not yet been created. First, and perhaps most likely, it would take several times more work to set up and execute such a study when compared to designing and running a study that focuses only on a single class of variables. For each variable that is being studied or manipulated, samples would need to be created that contain the full range of desired variations in terms of that variable, for assessment by real-world comprehension raters. That is, instead of just presenting assessors with samples that have varied linguistic properties, one would need to ensure that these samples also include sufficiently varied fonts, font sizes, x-heights, line spacing, line lengths, and margins – among any other factors that might be thought to have an impact on comprehension. And the work in creating this widely varied sample set would be substantial when compared to the work involved in collecting linguistic samples that are widely varied. Natural language organically contains a high degree of linguistic variation: each unique (non-repetitive) sentence is naturally different when compared to all other sentences. Each unique text is naturally different from every other text. However, sample texts would likely need to be deliberately (although perhaps computationally / automatically) manipulated to produce the necessary variations that would allow each non-linguistic variable to be studied. In other words, the number of sample texts that are presented to human assessors for scoring would likely need to be much larger for the type of global comprehension study that is being discussed here, so that enough texts in each different font style, size, etc, is scored by the human raters to allow for statistical assessment of the relationships between these variables and human comprehension.

The problem of sample size for the set of texts to be scored for this type of study is one that would impact upon human participants in the study in addition to the researchers: more texts to score means more time and energy inputs from human participants in the study. Practically speaking, these additional demands on participants' time (or on the number of participants needed within the study in order to keep each participant's time commitment to a reasonable amount) may present a barrier to the research that is difficult to overcome.

As this discussion suggests, the scope of a project would expand significantly with each new class of variable that one wants to study for the purposes of predicting different variables' effects on comprehension. This project-expansion problem could explain why no one has yet sought to develop a comprehensive quantitative comprehension formula.

A second hypothesis about why more global comprehension formulas have not yet been created relates to subject-matter expertise. The knowledge that would be needed to undertake a study capable of creating such a formula spans several academic disciplines. Ideally, it would involve applied linguists (for the linguistic readability dimension), educational researchers and/or educational psychologists (for the auditory dimension, and potentially others), and communications or visual design specialists (for the presentation and layout dimensions) – although other specialists may also be needed. Interdisciplinary academic research is far more common today than it was in the past, but is still less common than discipline-specific research in many fields. This historical, and current, tendency for the production of single-discipline research in academic fields can potentially explain why an interdisciplinary global quantitative comprehension formula has not been developed.

A final hypothesis draws on the relatively high degree of statistically predicted accuracy of many linguistic readability formulas in terms of explaining the variance in reading comprehension

levels. Specifically, I suggest here that researchers have not made efforts to develop more wide-ranging comprehension formulas because linguistic readability formulas are already capable of predicting sometimes more than half of the variance in reading comprehension differences across texts. Under these circumstances, perhaps researchers feel that linguistic formulas are good enough for their comprehension-assessment purposes, and no additional classes of new variables need to be considered beyond linguistic variables.

Consider, for instance, Flesch's readability formula, which was generated in the manner described above through statistical regression analysis. This formula produced a multiple correlation coefficient of R = 0.7047.[180] As Cohen explains, the squared value of the correlation coefficient (R^2) provides us with a measure of the proportion of variance.[181] In simple terms, this means that the R^2 value from a regression equation that creates a readability formula tells us the percentage of variation in the 'dependent variable' (the readability score) that can be explained by the 'independent variables' (the quantitative linguistic variables, such as *average sentence length*, that contribute to the relevant readability formula).[182] In the case of Flesch's formula, the R^2 value would be $(0.7047)^2$, or 0.497. So, Flesch's formula is statistically capable of explaining 49.7% of the variance in "Reading Ease" based on measurements for the two variables in the formula: *average word length* (in syllables) and *average sentence length* (in words).[183] In other words, according to Flesch and the statistics that flow from his formula, essentially half of reading comprehension variances in relation to texts can be explained by these two variables.

---

[180] Flesch, "A New Readability Yardstick", *supra* note 35 at 225.
[181] Jacob Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed (New York: Lawrence Erlbaum Associates, 1988) at 78.
[182] Cohen, *Statistical Power Analysis for the Behavioral Sciences*, *supra* note 181 at 78-79.
[183] Flesch, "A New Readability Yardstick", *supra* note 35 at 224-225.

Dale & Chall similarly found that their readability formula had a multiple correlation coefficient of $R = 0.7$,[184] which would produce an R^2 value of 0.49. However, their formula outperformed this prediction in real-world testing. The authors collected scores regarding 55 text passages about health education topics that were assessed by readability experts, and that were also presented to adults and children who were then asked a series of comprehension questions about the passages.[185] The adults and children were deemed to understand a text if they answered 75% of the comprehension questions about that text correctly.[186] When the Dale-Chall readability formula scores were compared with the judgements of readability experts, the two values correlated strongly ($r = 0.92$).[187] The same was true for the formula's correlation with the reading grades of adults and children who were able to correctly answer the required number of comprehension questions ($r = 0.9$).[188] Similar results were obtained for a separate set of 78 texts about foreign affairs ($r = 0.9$).[189] In each of these cases, the R^2 value would be approximately 0.81 – which is to say, Dale & Chall's two-variable linguistic readability formula explained 81% of the variance in human-assessed comprehension of different texts. Again, these results suggest that linguistic factors are overwhelmingly capable of explaining reading comprehension variations between texts, in ways that do not leave much room for other variables to have large and statistically significant relationships with comprehension. Perhaps this is why researchers have not yet sought to build more multi-factorial quantitative comprehension formulas.

---

[184] Dale & Chall, "A Formula for Predicting Readability", *supra* note 37 at 18.
[185] Dale & Chall, "A Formula for Predicting Readability", *supra* note 37 at 18.
[186] Dale & Chall, "A Formula for Predicting Readability", *supra* note 37 at 18.
[187] Dale & Chall, "A Formula for Predicting Readability", *supra* note 37 at 18.
[188] Dale & Chall, "A Formula for Predicting Readability", *supra* note 37 at 18.
[189] Dale & Chall, "A Formula for Predicting Readability", *supra* note 37 at 18.

While I acknowledge that it would likely be useful to study whether non-linguistic variables like the ones that are discussed in this section have a statistically identifiable relationship with human-assessed comprehension of a text, it is beyond the scope of this dissertation to undertake such a study. For reasons relating to time, expertise, and efficiency, this dissertation, like other quantitative readability studies, will focus on linguistic considerations that affect comprehension. But again, the idea that many other extra-textual factors also have an impact on ultimate comprehension is raised here as a caution: quantitative readability studies like the ones that follow in this dissertation are useful, and offer valuable information about how well texts might be understood by some or most readers. But they cannot offer complete truths about how well any specific person will understand any specific text, because there are simply too many individualized and reader-specific factors, and other non-linguistic factors relating to each text, that are not accounted for within quantitative readability formulas.

2.8 Reading, Understanding, and Using Adjudicative Decisions

There are important differences between the ideas of reading, understanding, and using sources of law, including adjudicative decisions. I appreciate that one could argue that it is possible to engage in the mechanical act of "reading" a text without understanding any of the meaning that text. However, within this dissertation, the idea of "reading" (whether in reference to concepts and theories of reading comprehension, or of linguistic readability) necessarily implies some additional component of understanding. Reading comprehension theorists suggest that word-to-text integration is a fundamental part of the reading process, which leads a reader to create a mental situational model of the text that ultimately gives the text meaning for the reader.[190] And even some

[190] Bruggink, et al, *Putting PIRLS to Use*, *supra* note 7 at 6-13.

of the earliest developers of quantitative readability formulas relied on reading comprehension test scores from student readers as their baseline for determining the grade levels of texts that were used in producing their formulas.[191] In other words, like the reading comprehension and linguistic readability scholars whose work I have relied upon in this dissertation, I reject definitions of "reading" that describe the process as being purely mechanical, and being delinked from any notion of comprehension. Reading includes at least some degree of comprehension of the text.

But comprehension is an amorphous concept. How much does a reader need to correctly understand about a text for us to say that the reader has achieved comprehension of that text? Scholars tend to have different views on this question. For instance, Flesch used a set of standard reading comprehension passages as the basis for his Flesch Reading Ease formula.[192] Each passage was accompanied by 8-10 comprehension questions. Flesch determined the reading grade level for each of these passages when students from a particular grade were able to correctly answer 75% of the comprehension questions.[193] Dale & Chall, in contrast, used the same standard reading comprehension passages to create their readability formula, but they determined the reading grade level for each of these passages when students from a particular grade were able to correctly answer only 50% of the comprehension questions.[194] Others have used altogether different measures – such as a person's ability to correctly fill in the missing word within a text 35% of the time, for at least 50% of the people at that grade level.[195] As these studies suggest, there is no consensus about what degree of comprehension is needed in order for a text to be understood.

---

[191] See, for instance, Flesch, "A New Readability Yardstick", *supra* note 35, and Dale & Chall, "A Formula for Predicting Readability", *supra* note 37.
[192] Flesch, "A New Readability Yardstick", *supra* note 35 at 222. Specifically, Flesch used the *McCall-Crabbs' Standard Test Lessons in Reading*.
[193] Flesch, "A New Readability Yardstick", *supra* note 35 at 222-223.
[194] Dale & Chall, "A Formula for Predicting Readability", *supra* note 37 at 15.
[195] J P Kincaid et al, Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch

Based on the varying standards for assessing comprehension in these past studies, that each fell below a 100% correct test score, it is safe to say that additional and higher degrees of understanding must also be possible. There must be other levels of understanding that go beyond the minimum that a reader would need to have to "understand" a text. We might therefore differentiate between a linguistic understanding of a text (that relies on a minimum but clearly less-than-perfect degree of comprehension) on the one hand, and higher or deeper levels of textual understanding, on the other hand. With respect to case law, this distinction might manifest when, for instance, a reader understands the linguistic content of an adjudicative decision, but does not necessarily understand the deeper legal content of the decision.

Consider, for instance, a decision dealing with an arbitrary detention under section 9 of the *Canadian Charter of Rights and Freedoms*. A reader might understand this decision in a linguistic sense because the reader is familiar with both the words "arbitrary" and "detention" in everyday English. But this does not mean that the reader would necessarily have a strong legal understanding of the decision, if the reader does not also realize that, according to Canadian case law, a detention can take place psychologically (as opposed to just physically),[196] and that a detention is arbitrary if police do not have lawful authority to detain a person[197] (even if there are other good reasons why police want to detain the person). As this example highlights, understanding an adjudicative decision – depending on how one defines and frames the concept of "understanding" – could involve much more than just being able to comprehend the linguistic meaning of the text of the decision. And in law, where common English words are often given unique or nuanced

---

Reading Ease Formula) for Navy Enlisted Personnel (Memphis, TN: US Naval Technical Training Command - Research Branch, 1975) at 11.

[196] *R v Grant*, 2009 SCC 32, at paras 30-31.

[197] *R v Grant*, 2009 SCC 32, at para 54.

meanings,[198] the gap between what is needed for a linguistic understanding of a decision and a deeper legal understanding of a decision will often tend to be large.

It must be understood that quantitative readability formulas like the ones discussed in this chapter, and throughout the remainder of this dissertation, do not assess deeper levels of comprehension. These formulas are only capable of predicting the level of comprehension that the researchers relied upon when creating the formula. So, a formula-based assessment that an adjudicative decision is readable to a person with a particular education level is not a signal that the person would likely understand the full legal content of the decision – only the linguistic content.

If we move forward from ideas of linguistic readability and deeper legal comprehension of a text, we should also recognize that, in the real world, people who seek out and try to read case law will often have some specific purpose in mind that takes them beyond just wanting to read and understand the content of an adjudicative decision. Presumably, once these people understand what a decision means they will want or need to do something with that knowledge. In other words, the usability of a decision (or the extent to which a decision can help them to achieve their purposes) is a related and relevant consideration.

I am deliberately adopting a definition of "usability" here that is broader and distinct from more common definitions of usability in legal studies (that tend to explore the concept from a web-based user interface perspective, to assess how easy it is for people to navigate through a site and find the information they seek).[199] The usability of a court or tribunal decision, in my view, would

---

[198] Brenda Danet, "Language in the Legal Process" (1980) 14 Law & Society Review 445, at 469 and 476-77.
[199] See, for instance, Kathleen Darvil, "Increasing Access to Justice by Improving Usability of Statutory Code Websites" (2023) 115 Law Library Journal 123 at 127-131.

be much more closely connected to ideas of legal capabilities and functionings,[200] in the sense that all of these concepts consider whether a person can and will act to enlarge their opportunities through use of law. This idea of usability is also related to the concept of functional (legal) literacy that Salyzyn, et al, have described in their study of user-encountered difficulties with typical court forms:[201] usability, like functional literacy, asks whether a reader can meet task demands based on what they have read, rather than simply asking whether the reader has understood what they have read. A usable adjudicative decision is therefore not only understandable, but it is also one that could empower or enable a reader to act so as to preserve or improve upon their legal state of affairs within their surrounding world.

As this conception of usability that I have proposed here suggests, a court or tribunal decision must be more than linguistically readable to be usable. But there is no unique way to use an adjudicative decision. Different classes of potential readers of adjudicative decisions will often have different motivations for reading the decisions and will aspire to use them in distinct ways for various purposes.

In spite of this likelihood, we might conceptualize the world of potential readers for adjudicative decisions by separating these readers into categories based on the extent to which they are directly affected by a decision.

The first tier of potential readers might therefore include the parties, their lawyers, any government departments or offices who must implement the decision in a particular case, any victims, and any directly affected community members (such as, for instance, church congregation

---

[200] Ann-Katrin Habbig & Ingrid Robeyns, "Legal Capabilities" (2022) 23:4 Journal of Human Development and Capabilities 611 at 612-615.
[201] Amy Salyzyn, et al, "Literacy Requirements of Court Documents: An Under-Explored Barrier to Access to Justice" (2016) 33 Windsor Yearbook of Access to Justice 216.

members who would be prevented from attending their church when a COVID-related restriction is unsuccessfully challenged in court).[202] Each set of these readers may need to first understand a decision, and then use that understanding to guide their future actions in relation to the case that the decision involved. The parties and their lawyers would, in some circumstances, want to use weaknesses in a decision as a basis for appealing that decision. In other circumstances, they would need to use the decision to do whatever is required from them to comply with the decision. Government departments may need to use a decision as the basis for legislative reform proposals that displace the impacts of that decision. A victim may want to use a decision that convicts or sentences an offender as the starting point for a civil suit seeking damages from the offender.

For all these people or groups, a decision will need to be locatable and understandable, at a minimum, in order to be usable. But for some of these stakeholders, more might be required. Victims who intuitively sense that they want, or should be able, to use a criminal court's decision to help launch a civil law suit may not find the decision to be usable if nothing in the decision itself, or in the surrounding legal ecosystem (that consists of, among other things, a court's website, public legal education and information resources, interactions that may have taken place between a victim and a prosecutor, etc), addresses the victim on this point.

A second tier of potential readers for an adjudicative decision could include people who are less directly or immediately affected by the decision. It could include people who want or need to understand and use the decision not in relation to the scenario from which the decision emerged, but rather in relation to a future scenario. Readers who would fall within this tier include other judges or adjudicators; other lawyers who need to keep abreast of the case law; self-represented litigants who research past case law in relation to their current problems; law students, professors,

---

[202] *Springs of Living Water Centre Inc. v. The Government of Manitoba*, 2020 MBQB 185.

and other legal researchers; and, government officials who may need to change broad ranges of actions or policies for the future because of how a past case has been decided.

As for the first tier of potential readers, those who fall within the second tier may also have different requirements in terms of what makes a decision usable. Legal researchers, for instance, may need decisions to be batch-downloadable and machine-readable so that these researchers can analyze the decisions at scale. Self-represented litigants, in contrast, may need particularly influential decisions to be more clearly drawn to their attention from within the giant population of Canadian cases writ large.

A third and final tier of potential readers of decisions would include all others who want or need to read a decision even though their connections to the decision may be more attenuated. People within this tier could be described as those who are interested in, but not particularly affected by, a decision. This class of potential readers would include community members at large, and those who want to see what courts and tribunals are doing in their adjudicative work – perhaps as a form of democratic or popular oversight. People within this tier may need certain decisions to be reported on and linked within media stories in order for the decisions to be usable. These readers may also need companion tools that summarize or explain what is happening within a decision, since they may be less familiar with the circumstances surrounding the decision than other stakeholders who are more directly affected by the decision would be.

The point here is not to itemize all the different properties that a decision must have in order to be usable to every different kind of potential reader for the decision. Instead, my goal is to demonstrate how each user of a decision (depending on the person's identity, level of background knowledge, and motivation for reading a decision) will require distinct things in order to make the decision usable for that person.

More broadly, my intent is also to show that neither linguistic readability, nor deeper / legal comprehension of a decision is likely to be the ultimate end that a reader of a decision is hoping to achieve. My sense is that most people who start reading a court or tribunal decision do this because they hope to achieve something outside of the decision, in the real world, that will advance (or prevent a decline in) their situation.

With these distinctions between reading, understanding, and using an adjudicative decision in mind, it becomes obvious that focusing on decision readability in isolation will not necessarily improve how well members of society understand these decisions in deeper or legalistic ways – since something more than a linguistic understanding of the decision is needed to achieve this deeper comprehension. And such a focus will not necessarily increase the usability of the decisions, if other overwhelming barriers to usability exist.

But striving to measure, and (where appropriate) improve upon linguistic readability levels of adjudicative decisions is arguably an important first step to be taken on any path toward improving deeper comprehension or usability levels of these decisions. I say this because linguistic readability is a threshold issue. It is a necessary, but not sufficient, condition for achieving deeper comprehension and usability of adjudicative decisions. If a person wants or needs to read an adjudicative decision, and that person can understand the surface-level language of the text, then it is logically possible that the person could possess the necessary ability to understand that text at a deeper level, and then use the text to advance personal goals. In contrast, if a reader cannot even understand the language of a text at a surface-level, then there is arguably no logical way that the person will have more complex or nuanced levels of textual understanding. Linguistic readability is a precondition for higher order interactions with a text, like the development of legal understanding, or the ability to use the text in furtherance of a person's aims.

2.9 Potential and Pitfalls: Readability Theory Guides the Way

This chapter provides readers with the theoretical knowledge that is necessary to understand and assess what follows within the dissertation, in terms of my development and applications of a law-specific readability formula. Within this chapter, I have tried to candidly and transparently explain the significant potential that readability theory holds for studies in law, but also the many limitations and qualifications that must be considered when using readability to study law.

I embark in the next chapter on the first step toward answering this dissertation's research questions. Specifically, the next chapter identifies a law-specific quantitative readability formula that is best able to generate insights (subject to the limitations discussed in this chapter) on questions about both general and specific readability levels of Canadian court and tribunal decisions. This formula is then applied in later chapters to report on the results of studies that put the law-specific readability formula into operation.

3. Identifying a Law-Specific Readability Approach

3.1 Overview

In order to ultimately determine how quantitatively readable Canadian adjudicative decisions are, one must first decide on the yardstick, or scale of measurement, that will be used to score or assess these decisions. However, this measurement-identification task presents a problem. On the one hand, there are many existing readability formulas (of varying degrees of notoriety) that could be used to score decisions for their readability.[1] On the other hand, none of these existing formulas were built for the purpose of measuring the readability of law-specific texts, and none of them used legal (or even particularly sophisticated) documents as the basis from which the formulas were created.[2] So, even though there are lots of existing formulas that might be capable of producing readability scores for Canadian court and tribunal decisions, any experiment that uses one of these formulas to score a group of adjudicative decisions would be extending that formula's use to documents and types of language that are clearly different from the documents and types of language that were used to create the formula.

Statistically, such a shift in a readability formula's use from one type of document (e.g.: grade-school readers or Wikipedia articles) that was used to create the formula, to another type of document (i.e.: court or tribunal decisions) with different language properties that is the object of

---

[1] For instance, the Automated Readability Tool for English (ARTE) web-based readability application calculates readability scores for texts using 9 different general-purpose readability formulas (that are each described on the site's "About Readability" page). Online: <https://nlp.gsu.edu/home>.

[2] For instance, the Flesch Reading Ease formula was created from the *McCall-Crabbs Standard Test Lessons in Reading*, which are short reading comprehension passages and associated comprehension questions for students between grades 2-12: Rudolph Flesch, "A New Readability Yardstick" (1948) 32 Journal of Applied Psychology 221 at 222-223. The Crowdsourced Algorithm of Reading Comprehension was created from a selection of Wikipedia articles that were read and ranked based on relative reading difficulty levels by crowdsourced raters: Scott A Crossley, Stephen Skalicky & Mihai Dascalu, "Moving Beyond Classic Readability Formulas: New Methods and New Models" (2019) 42 Journal of Research in Reading 541 ["Moving Beyond Classic Readability Formulas"].

the formula's predictions, may weaken the predictive power of the formula. Regression formulas, including all of the major first- and second-generation formulas that were described in the previous chapter, are most reliable when used on new data that falls within the same range as the original data that served as the foundation for the formula. So, for instance, if the documents that were used to create the Flesch Reading Ease formula (which is calculated based on average word and sentence lengths)[3] had average sentence lengths of between 4 and 35 words, then the formula would not necessarily be reliable in predicting the readability of documents with much larger average sentence lengths of between 30 and 90 words. The use of regression formulas on data that falls outside of the scope of the formula's original data is called extrapolation, and the predictive unreliability of such formulas when used for extrapolation is well-recognized by statisticians: a regression formula's output may be biased or imprecise when extrapolating, and may become even more so depending on the size of any data differences between the original and the subsequent data sets.[4]

This inherent limitation associated with regression formulas when used beyond their original purposes seems to show itself in readability contexts. For instance, research suggests that general-purpose readability formulas do not necessarily perform effectively on texts drawn from specific and technical domains (like the health-care domain) that may use a different form of English than that used in general-purpose texts.[5] So, there are reasons to question whether any existing readability formula will produce valid readability results on a corpus of law-specific texts, like adjudicative decisions.

---

[3] Flesch, "A New Readability Yardstick", *supra* note 2 at 229.
[4] See, generally, Gerald J Hahn, "The Hazards of Extrapolation in Regression Analysis" (1977) 9 Journal of Quarterly Technology 159.
[5] Jiaping Zheng & Hong Yu, "Readability Formulas and User Perceptions of Electronic Health Records Difficulty: A Corpus Study" (2017) 19:3 Journal of Medical Internet Research.

Furthermore, even if no statistically-based argument or prior research called into question the reliability of existing general-purpose readability formulas for use on adjudicative decisions, one would still need to determine which of the existing formulas is best-suited for the task. No one has yet attempted to identify the best formula for measuring the readability of court and tribunal decisions.

The above general context sets the stage for the first of three major empirical undertakings within this dissertation. Specifically, this context demonstrates the need for Undertaking #1 (my Formula-Identification Project). In this project, that I describe in detail below, I rely on expert human ratings (that are cross-validated against ordinary, non-expert human ratings) of extracts from actual court and tribunal decisions to establish a sound basis for two related purposes: (1) creating a new law-specific readability formula; and, (2) testing this new formula, and 6 other general-purpose readability formulas, against my human-generated readability scores to find out which readability formula's (predictive) scores best correlate with actual assigned scores for those texts. In this manner, I conclude Undertaking #1 by identifying for the first time how Canadian case law readability[6] should be quantitatively assessed, going forward.

3.2 UNDERTAKING #1 – THE BROAD CONCEPT

In general terms, the task of finding the readability formula that has the greatest statistical likelihood of accurately predicting human-assigned comprehension scores first requires me to establish "gold standard" human-generated assessments[7] of the readability levels of a corpus of

---

[6] Recall, from section 1.2 (above), that "readability" for the purposes of this dissertation means "a quantitative assessment of a text that is designed to predict how well the text might ultimately be understood by a generic reader, based solely on the linguistic features that the text contains."

[7] There is widespread acknowledgement in readability studies that some form of human-generated reading difficulty assessment is needed as the gold standard from which follow-on predictive models can be derived. See, e.g., Sasikiran Kandula & Qing Zeng-Treitler, "Creating a Gold Standard for the Readability Measurement of Health Texts" [2008]

documents that have similar linguistic properties to the documents that are the focus of what I am studying in this dissertation (i.e.: real Canadian court and adjudicative decisions). Next, by splitting this "gold standard" data set into two groups, I can use one sub-group of the data to derive a new law-specific readability formula. Specifically, I will calculate scores for over 1000 individual linguistic variables for each document within the first sub-group, and then identify through correlations and regression analysis which of these variables (in which combination and with which weighting) might best predict the original "gold standard" readability scores for the documents within the sub-group. In this manner, I create a new law-specific readability formula.

Finally, using the other "gold standard" sub-group of texts, I calculate readability scores for every document within the second sub-group using both my new law-specific readability formula, and 6 other existing general-purpose readability formulas. I then assess the correlations between scores produced using each of these formulas and the "gold standard" scores that were actually assigned by my expert raters, to identify which formula's scores most strongly correlate with the "gold standard" scores. In this way, I select the best formula for further use within my dissertation based on its ability to quickly and easily assign readability scores to adjudicative decisions texts, where these scores are strongly associated with scores that a human expert would be expected to assign to the same texts. In executing all of these research tasks, my approach parallels established methods that have been used to create readability formulas by applied linguists over the last century.[8]

American Medical Informatics Association Symposium Proceedings 353 (wherein the authors rely upon expert ratings performed by humans to generate readability scores) ["Creating a Gold Standard"]; see also Orphée de Clercq, et al, "Using the Crowd for Readability Prediction" (2014) 20:3 Natural Language Engineering 293 at 299-304 (wherein the authors take as a given that human-generated assessments are needed to develop a gold standard, but debate the merits of using expert raters as compared to non-expert – crowdsourced – readability raters).

[8] Mabel Vogel & Carleton Washburne, "An Objective Method of Determining Grade Placement of Children's Reading Material" (1928) 28:5 The Elementary School Journal 373 ["An Objective Method"]. See also James W Cunningham,

3.3 Step 1 – Establish "Gold Standard" Readability Scores

3.3.1 Method – Establish "Gold Standard" Readability Scores

In order to generate benchmark readability scores for adjudicative texts, I follow a method that represents a blending of the methods used by Crossley, Skalicki & Dascalu in creating their general-purpose CAREC readability formula,[9] the methods used by Crossley, et al, in creating their healthcare-specific Model of Text Readability in Physicians (MoTeR-P) readability formula,[10] and the methods used by Kandula & Zeng-Treitler in their comparison of expert readability measurements with classic readability formula measurements for healthcare-related texts.[11] I do this through the following sub-steps – each of which is described in more detail, below: A) building a diverse body of texts for my experts to score; B) obtaining expert readability scores for each text; and, C) obtaining "lay" (i.e.: ordinary or non-expert) readability scores for each text. At the end of this process, I will have established a discrete readability score (based on the average of expert scores) for each document within my "gold standard" corpus of adjudicative decision text files.

3.3.1.1 Sub-Step 1A – Build the Corpus of Reference Texts

My method involves the creation of human-generated readability measurements for a reference corpus of law-specific texts, consisting of excerpts from judicial and administrative tribunal decisions that were published in 2021 – the most recent full calendar year of decisions prior to the year (2022) from which decisions are drawn for this dissertation's subsequent readability studies. Readability research suggests that stylistic variations tend to occur over time,

---

Elfrieda H Hiebert & Heidi Anne Mesmer, "Investigating the Validity of Two Widely Used Quantitative Text Tools" (2018) 31 Reading and Writing 813 at 814.

[9] Crossley, Skalicky & Dascalu, "Moving Beyond Classic Readability Formulas", *supra* note 2.

[10] 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" (2020) 13:4 Journal of Communication in Healthcare 344 ["Physicians' Secure Messages"].

[11] Kandula & Zeng-Treitler, "Creating a Gold Standard", *supra* note 7.

and that texts that may have been more readable to audiences at the time when the texts were first published are less readable to today's audiences.[12] Since the focus of my dissertation is on the readability of contemporary adjudicative decisions, my reference corpus consists of texts that were recently published, and that are drawn from a period that is close in time to (i.e.: within 2 years of) the decisions that are studied within subsequent chapters of this dissertation.

It is perhaps worth noting here why I have chosen to focus this dissertation on the readability of **contemporary** adjudicative decisions. I acknowledge that people who want to understand the law may have a need to read case law from much older periods (for instance, the *R v Oakes*[13] case from the Supreme Court of Canada remains the leading case on when it is constitutionally permissible for the government to limit a person's *Charter* rights, and that case was decided in 1986). So, there are compelling reasons why we might want to know the readability levels of these older decisions. However, the authors of such decisions, and the authors from all times past who are no longer in positions where they still write adjudicative decisions, cannot be influenced in ways that would make them write more readable decisions in the future – if this dissertation reveals that past readability levels are not where society needs them to be. In contrast, authors of decisions from 2022 are far more likely than authors from 1986 (or earlier periods) to still be serving in adjudicative decision-making roles. If this dissertation gives these authors cause to think that they should write more readably, going forward, then they are well-positioned to start making such efforts during, and for the remainder of, their adjudicative careers.

Beyond just the time frame for decisions within the initial reference corpus of texts, I had several other objectives in building the corpus. First, I wanted to ensure that it included a wide

---

[12] See generally, Wayne A Danielson and Dominic L Lasorsa, "A New Readability Formula Based on the Stylistic Age of Novels" (1989) 33:3 Journal of Reading 194.
[13] [1986] 1 SCR 103.

spectrum of the types of language that can be found in Canadian court and tribunal decisions. Second, I wanted to make sure that the corpus was large enough for my purposes, even after it was split into formula-building and formula-testing sub-groups. Third, given that volunteer human raters were reading and scoring the readability of each document in the corpus, I wanted to ensure that the documents, and the overall work obligations required of the volunteers, were not unduly long. These considerations all factored into my decision-making about how to build the corpus.

I built a corpus of 375 text files, where each file contained approximately 600 words. Based on an estimated reading time for each volunteer of approximately 3 minutes for every 600-word file, using a corpus of this size would require each expert volunteer rater to dedicate approximately 18.75 hours to the task of reading and scoring the 375 files in the corpus, for an estimated total time commitment of about 20 hours per volunteer (including time spent during initial meetings, emails, etc). I felt that it would be inappropriate, and likely unachievable, to seek anything greater than this 20-hour time commitment from expert volunteers participating in my research.

My 375-file corpus contained less files than the 600 files that were used by Crossley, Skalicki & Dascalu,[14] and less than the 724 files that were used by Crossley, et al.[15] However, it contained a total of 233,527 words – which was more than the approximately 120,000 words within Crossley, Skalicki & Dascalu's corpus,[16] and more than the approximately 213,500 words in Crossley, et al's corpus.[17] Based on the number of words and files within my corpus, I assessed that its size was sufficient for my purposes. Additionally, based on a related study from a slightly

[14] Crossley, Skalicky & Dascalu, "Moving Beyond Classic Readability Formulas", *supra* note 2 at 547.
[15] Crossley, et al, "Physicians' Secure Messages", *supra* note 10 at 346.
[16] Crossley, Skalicky & Dascalu, "Moving Beyond Classic Readability Formulas", *supra* note 2 at 547.
[17] Crossley, et al, "Physicians' Secure Messages", *supra* note 10 at 347.

different context,[18] my expectation was that by using relatively longer texts (albeit within a corpus containing a lower number of texts) than some others had recently used in conjunction with Natural Language Processing techniques to derive readability formulas, I would be in a better position to assess correlation strengths between linguistic features in these texts, and the overall readability levels of the texts.

In order to generate a corpus that was representative of Canadian case law as a whole, I ran 13 different search queries[19] using the CanLII database,[20] and then selected randomly from the results of each query to identify decisions that would contribute to my corpus. I have kept a complete history of my search and case-identification techniques on file, but describe this technique here for 2 of my 13 queries (i.e.: for my queries intended to ensure that I had diverse legal subject areas within my corpus, and a minimum number of appellate decisions), for the sake of illustration:

Diverse Legal Subject Areas

- Search text = "law";
- Time = 2021;
- Tribunal = all courts and tribunals;
- Filter subject = "business; insurance; IP; international; property & trusts; taxation";
- Sort by "most pages";

---

[18] Scott A Crossley, "Limitations of Natural Language Processing Tools for Data Mining: Text Length and Grade Level Differences" in Kristy Elizabeth Boyer & Michael Yudelson, eds, *Proceedings of the 11th International Conference on Educational Data Mining* (Buffalo, NY: International Conference on Educational Data Mining, 2018), 630 at 634. In this study, the author used Natural Language Processing techniques to assess whether linguistic variables within an online math tutoring environment correlated with student success on associated math questions. The author found that correlation strengths increased for the linguistic variables being studied in proportion to the overall length of the text. It seems from this study that one can acquire a more refined picture of the relationship between linguistic variables and other variables being studied if the sample of language that is used is relatively larger.

[19] These search queries were intended to, for instance: generate broad jurisdictional coverage of texts from all federal and provinces/territorial (except Quebec) courts; include some likely more readable decisions (from tribunals identified in my research as ones with interest in producing readable decisions); include some likely less readable decisions (on subjects like tax law and intellectual property law that would likely be unfamiliar and abstract to many readers); ensure that a minimum number of frequently-cited decisions were included; ensure that a minimum number of administrative tribunal decisions were included; etc.

[20] Online: <http://CanLII.org>.

- Take every 10th decision until 30 decisions have been identified;

Ensure Minimum Number of Appellate Court Decisions

- Search text = "law";
- Time = 2021;
- Tribunal = All appeal courts;
- Sort by "newest";
- Take every 10th decision until 40 decisions have been identified.

If, during the selection process, a decision that should have been selected was found to contain less than 600 words (or if a decision that had already been included based on a previous search query was due for selection a second time), then the next decision on the search result list was identified instead, so that it would be possible to extract an original 600-word sample from the decision.

For each decision that was identified based on my search queries and selection processes, I then opened the decision, and selected quasi-randomly[21] a passage of approximately 600 words of text from within the decision. Specifically, I drew approximately evenly from what I assessed to be the first third, middle third, and last third of the decisions when identifying passages for inclusion in my corpus, in an effort to ensure that my corpus contained a balance of language from Introduction/Overview/Facts sections (typically in the first third of a decision), Analysis sections (usually in the middle third), and Conclusion/Remedy/Order sections (typically in the last third of

[21] I say that the passages were selected "quasi-randomly" here because I did not use computational or mathematical approaches to ensure randomness in the selected passages. My method in selecting passages was as follows: I selected from approximately the first 1/3, second 1/3, or final 1/3 of a decision in a repeating pattern of 3's, such that the 1st, 4th, 7th, …, etc, decisions would each contribute an extract from the first 1/3 of the decision, and the 2nd 5th, 8th, …, etc, decisions would each contribute an extract from the middle 1/3 of the decision, and so on. For each decision, I opened the decision file in my web browser. I looked roughly at the size of the decision by assessing the height of the scroll bar for vertical scrolling that is located at the rightmost edge of my browser window, and scrolled to somewhere in the relevant 1/3 portion of the decision from which an extract was to be drawn. Where ever my cursor landed, I looked for the beginning of the next sentence after the cursor, and selected what appeared to be a 600-word segment of text. I generally erred on the side of selected more, rather than less, than 600-word samples. I then added or deleted from the extracted text sample until it fell somewhere within the 550–650 word range, while still ending in a complete sentence. There was no empirical rationale for using completely random text extracts, and my quasi-random process contributed sufficiently to my objective of creating a broad and diverse corpus of text extracts.

a decision) – on the assumption that different styles of language may be included in these different sections.

For each decision, I then copied and pasted the selected text sample into a blank Microsoft Word document, and verified its length (adding or deleting full sentences as necessary to produce a roughly 600-word extract from the decision) using Microsoft Word's "word count" function. I then saved the file as a plain text file within my corpus. Each file was saved using arbitrary combinations of letters and numbers as the file name, so that when the files were sorted alphabetically in a computer folder, they would be listed in a random order. This file naming convention helped to ensure that a volunteer participant who was reading and scoring files based on their alphabetical order in a file folder would not come across the files in any particularly relevant order, and would therefore encounter files from each of my 13 search query groups more or less at random.

At the end of this process, I had a database/corpus of 375 discrete files each containing an approximately 600-word text sample drawn from a judicial or administrative tribunal decision that had been identified in order to ensure that the corpus had broad coverage of the types of decisions that make up the complete body of Canadian case law.[22]

3.3.1.2 Sub-Step 1B – Obtain Expert Readability Scores for Each of the Reference Texts

Like Crossley, et al[23] and Kandula & Zeng-Treitler,[24] I decided to rely upon panels of expert raters to generate my "gold standard" human readability assessments using a Likert Scale. I sought

[22] This dataset of 375 text files, including the case name and citation for the decision from which the text was extracted, is publicly available: Mike Madden, "MADRS_MaddenDissertation_Chapter3_375FilesScored" (30 May 2024), online: <https://doi.org/10.7910/DVN/J9Y3MH>, Harvard Dataverse, V2.
[23] Crossley, et al, "Physicians' Secure Messages", *supra* note 10.
[24] Kandula & Zeng-Treitler, "Creating a Gold Standard", *supra* note 7.

and obtained ethical approval to involve expert raters as human subjects within my research.[25] Using expert raters offers several benefits, but a major advantage of this approach is that it allows a researcher the ability to rapidly and efficiently rate a large number of texts,[26] especially when compared to an approach that attempts to more directly measure a particular person's comprehension of a text that they have just read through, for example, short answer or fill-in-the-missing-word questions about the text. Even scholars who opt to use other methods for generating readability scores on a reference corpus (like De Clercq, et al, who used a non-expert, crowd-sourced approach alongside an expert approach within their study) concede that using expert scoring is the "gold standard,"[27] and that other approaches are mainly desirable in order to compensate for the "expensive and time-consuming" dimensions of expert scoring.[28]

Regardless of whether expert, or ordinary/crowd-sourced human raters are used to score a reference corpus as the "gold standard" for subsequent readability purposes, there is a growing consensus that these subjective approaches to scoring the readability of texts is preferred.[29] DeClercq, et al, suggest that the typical way to generate these human scores is "to let people assign

---

[25] University of Ottawa Research Ethics Board file S-06-23-9339.

[26] Kandula & Zeng-Treitler, "Creating a Gold Standard", *supra* note 7 at 357:

> there is no gold standard for evaluating existing readability formulae or developing new ones in the health domain. Compared to the Cloze test, the expert panel approach makes it feasible to assess a much larger sample of documents. In lieu of using documents of certain types (e.g. blogs or journal articles) as the gold standard for easy or difficult materials, expert ratings do not rely on the presumption of a particular type of document being inherently difficult or easy, and recognize the varying levels of difficulty within a type.

Although there are drawbacks to using expert raters, the benefits may outweigh the costs (Kandula & Zeng-Treitler, "Creating a Gold Standard", *supra* note 7 at 357): "The gold standard we created relies on expert knowledge, rather than direct user testing. Ideally, we would like to test each document on a representative sample of users. However, adult health care consumers are extremely diverse. Given the formidable cost of testing a large number of documents on a large number of patients, we believe the experts' opinion can serve as a reasonable proxy."

[27] DeClercq, et al, "Using the Crowd for Readability Prediction", *supra* note 7 at 299.

[28] DeClercq, et al, "Using the Crowd for Readability Prediction", *supra* note 7 at 299.

[29] DeClercq, et al, "Using the Crowd for Readability Prediction", *supra* note 7 at 297.

absolute scores to each text and use the resulting mean readability score."[30] This is precisely the approach that I use within my dissertation.

My intention was to select as expert raters a small group of individuals who had both legal expertise and experience, and some additional expertise about the needs and capabilities of the parties to legal disputes. In order to solicit potential experts with these skills, I approached executives at both the federal Social Security Tribunal, and the British Columbia Civil Resolution Tribunal (two tribunals that have focused their members on writing readable decisions, and whose members tend to interact directly with unrepresented participants much of the time) about providing volunteers. I also approached several provincial legal aid executives about providing legal aid staff lawyers, and several provincial law (reform) commissions about providing legal staff. Finally, I approached several other individuals (e.g.: a retired Chairperson of the Social Security Tribunal, staff of the National Self-Represented Litigants Project, etc) about participating as volunteer experts.

In the end, I succeeded in recruiting the following six experts: three legal aid staff lawyers, one federal tribunal member, one provincial tribunal member, and one law (reform) commission lawyer (who had deep access-to-justice experience in his current role, as well as several years of legal aid experience in a previous role). Each of these experts has multiple post-secondary degrees (including at least one law degree), and legal practice experience. All of the experts also have extensive experience in working with participants in the legal system who have varying degrees of legal sophistication and legal literacy. It was this combination of each person's own ability to understand law at its highest level of complexity (based on the person's education and professional practice experience), together with their experiences in helping less sophisticated users of the legal

---

[30] DeClercq, et al, "Using the Crowd for Readability Prediction", *supra* note 7 at 297.

system to understand and access law, that I relied upon to establish these individuals as expert raters for the purposes of my study.

This approach to identifying and selecting experts is consistent with approaches that others have used in the past. For instance, DeClercq, et al, used a combination of 36 teachers, writers, and linguists as expert readability raters for texts of different genres (e.g.: administrative texts, news articles, and miscellaneous others).[31] Crossley, et al, used a large group of teachers who had been recruited by email as expert raters for scoring a corpus of texts commonly found in the kindergarten to grade 12 classroom environments, from both informational and literary genres.[32] Kandula & Zeng-Treitler used the following experts to rate the readability of health-related literature: "a health literacy consultant with extensive experience in content development and health communication education, a health communication researcher with background in education psychology and nursing, a patient education content developer and nurse, a practicing nurse and certified diabetes educator, and a librarian from a hospital-based consumer health library."[33] And, Steenhuis, Willey & Collarusso used "expert reviewers who variously work in the field of plain language, participate in form committees, and regularly work with self-represented litigants" in their study that assessed the usability of court forms.[34] As these examples illustrate, expert raters may acquire their expertise through experience with language or teaching (for general-purpose texts or texts drawn from multiple fields), but they may also acquire their expertise by having a heightened understanding of both the specialized subject of the text, and of

---

[31] DeClercq, et al, "Using the Crowd for Readability Prediction", *supra* note 7 at 302.
[32] Scott Crossley, et al., "A Large-Scaled Corpus for Assessing Text Readability" (2023) 55 Behavior Research Methods 491 at 494.
[33] Kandula & Zeng-Treitler, "Creating a Gold Standard", *supra* note 7.
[34] Quinten Steenhuis, Bryce Willey & David Colarusso, "Beyond Readability with RateMyPDF: A Combined Rule-based and Machine Learning Approach to Improving Court Forms," in *Proceedings of International Conference on Artificial Intelligence and Law (ICAIL 2023)* (Association for Computing Machinery: New York, 2023), online: <https://doi.org/10.1145/3594536.3595146>.

the challenges that readers of this text might face (for specialized, technical, or discipline-specific texts). My expert raters fall into this latter category, in that they have law-specific expertise – particularly with respect to the needs of self-represented and legally unsophisticated users of the legal system – for scoring the readability of law-related texts.

Once the experts had been identified, I sent them 5 text files of approximately 600 words each, containing excerpts from adjudicative decisions that I had selected (non-randomly) in an effort to cover a range of text readability levels across the 5 files. My objective in selecting these files was to provide the experts with texts that I thought were likely to fall at different points on the scoring scale (in order to cause the experts to think about how they might apply a range of different scores on the scale), so I deliberately selected files that I thought were easy, middling, and difficult to read. I also sent the experts my preliminary (1-page) guidance to them about how to score the files based on a 7-point Likert scale that I provided to them.[35] I asked each expert to read and be prepared to discuss how they would score each file based on the provided scale. We then met together for a training / scoring calibration session that took place over approximately one hour, on 7 September 2023.

Over the course of this scoring calibration session, the experts discussed each file and how they would each score the files. They first sought clarification from me about whether they ought to assume that readers of the files would be adults (above the age of majority). Since children would generally have counsel and/or litigation guardians assisting them in accessing the legal system, and since I am primarily interested in the question of how readable Canadian case law is to adults, I instructed the experts to assume that readers of the files would be adults. The experts

---

[35] A 7-point Likert scale was similarly used to assess readability in Jean-Charles Chebat, "Testing Consumers' Motivation and Linguistic Ability as Moderators of Advertising Readability" (2003) 20:7 Psychology & Marketing 599 at 607.

also asked whether they should assume that readers of the files are random people who happen to have the files placed in front of them, or are instead people who have sought out the files because they have some motivation or purpose for reading and understanding the file. I told them to assume that their readers have some motivation or purpose for finding and reading the documents (e.g.: they want to know more about a case that was in the news; a friend was involved in a legal proceeding and they want to read the decision from that friend's case, etc). Finally, the experts sought clarification about what was meant by the phrase "secondary school" within the scoring scale, since this term meant different things in the different provinces where the experts were located. I advised them to take "secondary school" as meaning "grades 9-12". The experts then scored the 5 sample files and talked about why they assigned the scores that they did. There was widespread agreement (a maximum of 1-point scoring variations across the 6 experts for each file) about how each file should be scored. On this basis, I determined that the experts were ready to independently score each of the 375 files in my corpus.

I provided each expert with a revised copy of the scoring guidance document that incorporated the clarifying direction that I had issued to the experts during our scoring calibration session (a copy is included below at Figure 3.1). I also sent each expert an electronic copy of all 375 files to be scored, and a blank Microsoft Excel scoresheet that required them to input a score of between 1 and 7 for each file. The experts then commenced their independent scoring, and returned their completed scoresheets to me. I received the last scoresheet in January 2024.

Figure 3.1 – Expert Scoring Guidance Document

EXPERT RATER SCORING GUIDANCE

You have been provided with access to a database consisting of 375 discrete text files. Each file contains an extract of a recent judicial or administrative tribunal decision. Some extracts are from the beginning of a decision, while others are from the middle, and others from the end.

PLEASE READ AND ASSIGN A READABILITY SCORE TO EACH FILE, USING THE SCALE SHOWN IN THE TABLE BELOW.

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

* "secondary school" means grades 9-12.

Please assign scores using only whole numbers, guided by the above table.

Recall:

This study aims to measure the readability of texts drawn from adjudicative decisions, but not the extent of legal knowledge that would be needed to fully understand all of the implications of the text. In other words, you are being asked to assess surface-level readability: who is likely to be able to understand the language of the text sample?

If you have questions, please reach out to me.

3.3.1.3 Sub-Step 1C – Obtain Lay Readability Scores for Each of the Reference Texts

Although my method involved using expert scores as the ultimate "gold standard" measurement of readability for each text within my 375-file corpus, I also used non-expert raters who produced scores that allowed me to assess the validity of the expert ratings when compared against non-expert ratings. I sought and obtained ethical approval to have non-expert raters participate in my research as human subjects.[36] The idea of using non-expert raters alongside

[36] University of Ottawa Research Ethics Board file S-06-23-9339.

expert raters follows the method used by Kandula & Zeng-Treitler, who used a patient advocate (a person whose family member had diabetes) to score a subsection of files that had also been scored by experts, and who found that the correlation between the patient scores and the expert scores "provided an extra validation to the experts' rating."[37] Ideally, within my Undertaking #1, there would be substantial agreement between the broad scores assigned by both experts and non-experts for each file – anticipatorily defined as agreement above the level of $W = 0.61$ when measured using Kendall's Coefficient of Concordance (Kendall's W) for inter-rater agreement among the expert raters,[38] and a large effect sized correlation of greater than 0.5 to describe the rank-order relationships between expert and non-expert scores for files in the dataset.[39]

My intention was to select a group of grade 12 students as lay raters, since people at this education level would fall close to the middle of the education range that is captured within my 7-point scoring scale that the expert raters were using. These individuals would therefore be in an ideal position to provide a non-expert assessment of whether a text was understandable to a person with approximately their education level, or a substantially higher or lower education level.

By disseminating my call for volunteers through my own social media and personal email networks (which included a significant number of parents with children in grade 12), I successfully recruited 18 different students in grade 12 who were willing to volunteer between 3-9 hours of time to score files. I then sent each student the same 5 files that I had sent to my experts, and a

---

[37] Kandula & Zeng-Treitler, "Creating a Gold Standard", *supra* note 7.
[38] J Richard Landis & Gary G Koch, "The Measurement of Observer Agreement for Categorical Data" (1977) 33:1 Biometrics 159, at 165.
[39] Jacob Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed (New York: Lawrence Erlbaum Associates, 1988) at 80-81.

copy of a one-page "lay scorer's guidance document" (containing a 3-point Likert Scale[40] for the students to employ), in preparation for a training / scoring calibration session.

During the last week of September 2023, I conducted two one-hour scoring calibration sessions (where each student volunteer attended one of these sessions) to discuss scoring of the sample files using the scoring guidance document, and to answer any questions from the students. At the end of these meetings, I assessed that the students were ready to independently score their respective files based on their apparent understanding of the scoring criteria and the work to be performed.

I then sent each student electronic copies of the files (ranging from 33-99 files per student, depending on the number of volunteer hours that they were willing to offer) for scoring, a copy of the lay rater scoring guidance document (included below at Figure 3.2), and a blank Microsoft Excel scoresheet requiring them to input a score of between 1-3 for each of their files. The students then commenced their independent scoring, and returned their completed scoresheets to me. I received the last scoresheet in January 2024.

One of these student volunteers did not complete any scoring (without explanation), so ultimately, only 17 students scored files (with 3 of the remaining 17 students graciously agreeing to share the extra scoring responsibilities of the 18th student). In the end, each of the files within my 375-file corpus was independently scored by 3 students.

[40] A 3-point Likert scale was similarly used to assess readability in Emilia da Silva Pons et al, "Users' Preferences and Perceptions of the Comprehensibility and Readability of Medication Labels" (2019) 14:2 PLoS ONE e0212173, online: <https://doi.org/10.1371/journal.pone.0212173>, at 3.

Figure 3.2 – Lay Rater Scoring Guidance Document

LAY RATER SCORING GUIDANCE

You have been provided with access to a database consisting of several text files. Each file contains an extract of a recent court or tribunal decision. Some extracts are from the beginning of a decision, while others are from the middle, and others from the end.

PLEASE READ AND ASSIGN A READABILITY SCORE TO EACH FILE, USING THE SCALE SHOWN IN THE TABLE BELOW.

This text could be likely be understood by an adult who is fluent in English and:	
who has a lot more education than I have.	3
who has the same or slightly more education than I have.	2
who has less education than I have, or who is less than fluent in English.	1

* Please assign scores using only whole numbers, guided by the above table.

Recall:

This study measures readability, not legal or technical knowledge. In other words, you are just being asked to assess surface-level readability: who is likely to be able to understand the language of the text sample that you are reading?

There are no right or wrong answers; rather, this study will be most helped by your honest assessment of each sample.

If you have questions, please reach out to me.

3.3.2 Results – Establish "Gold Standard" Readability Scores

After obtaining both expert and lay scores for each document within my 375-file corpus, I assessed the results of these scoring processes to determine whether I had sufficiently reliable and valid data for the purposes of this research undertaking.

3.3.2.1 Expert Scoring Results – Discussion

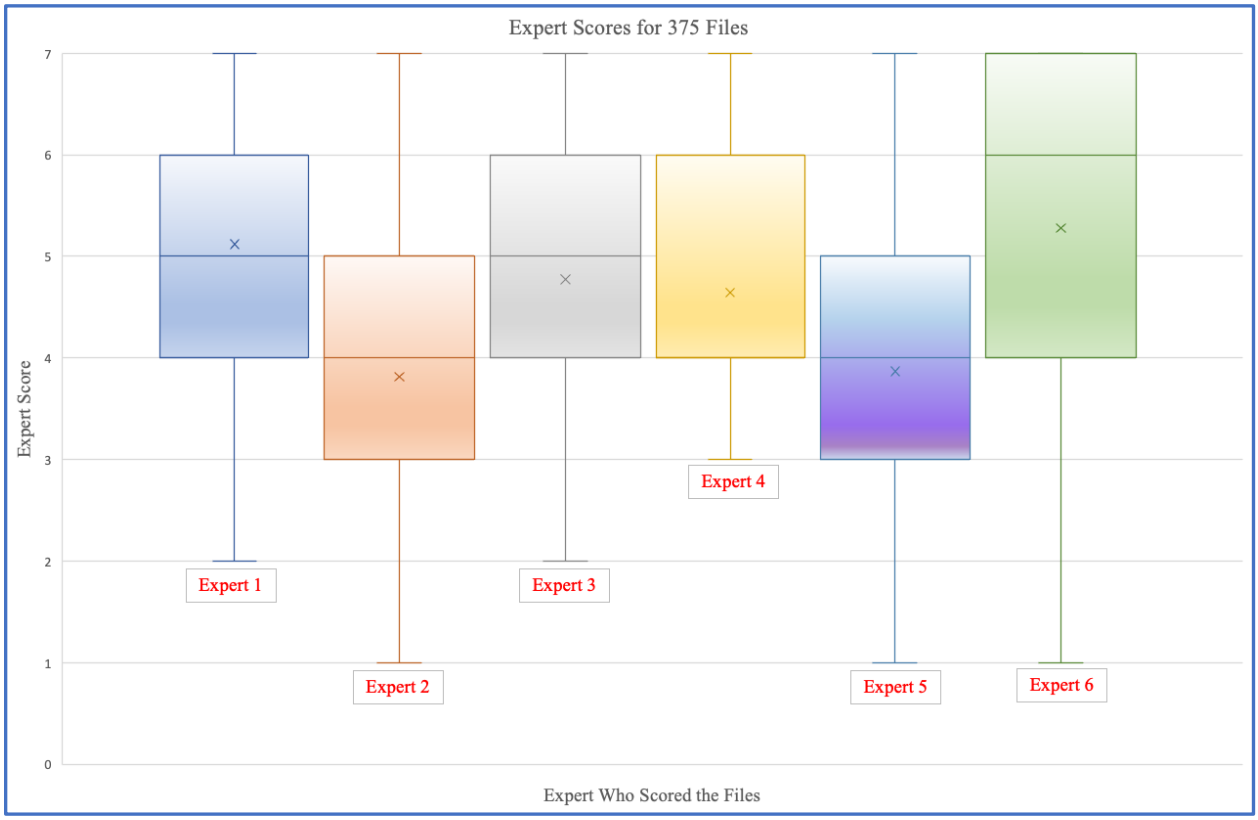
Each expert scored each of the 375 files in my corpus, for a total of 2250 independent scoring observations by the 6 experts, that involved approximately 120 hours of collective volunteer time.[41] The average score across all 2250 observations was 4.6 (requiring an education

[41] A spreadsheet containing individual and average expert scores for all 375 files is publicly available: Mike Madden, "MADRS_MaddenDissertation_Chapter3_375FilesExpertScores" (30 May 2024), online: <https://doi.org/10.7910/DVN/NQ0EYO>, Harvard Dataverse, V1.

level of between "some post-secondary education" and "an undergraduate degree or higher level of non-legal education" from the reader in order to understand the text). Individually, the experts assigned average scores of between 3.8 (for the expert with the lowest average text score) and 5.3 (for the expert with the highest average score).[42] The expert scores are represented visually below, in Chart 3.1, in the form of boxplots. For each expert who scored the set of 375 files, the "box" represents the interquartile range[43] for that expert's scores. The middle horizontal line represents that expert's median, or middle, score (the score that has an equal number of higher and lower scores within the sample).[44] The "x" represents the expert's average score. The "whiskers" extending above and below each box extend to show the full range of scores assigned by the expert.

[42] In ascending order, the experts' individual average assigned scores were as follows: 3.8 / 3.9 / 4.6 / 4.8 / 5.1 / 5.3.
[43] "Interquartile range" refers to the range between the 25th and 75th percentiles; it offers a view of the middle 50% of the data points, and is less sensitive to outlier points than other dispersion measures: Michael O Finkelstein & Bruce Levin, *Statistics for Lawyers*, 3rd ed (New York: Springer, 2015) at 25.
[44] For Expert 4, the median line is not visible, because the median score assigned by this expert was 4. Since the median score and the lower extremity of the interquartile range are both "4" for this expert, the median line does not appear within the boxplot (or appears directly on top of the line showing the lower boundary of the inter-quartile range).

Chart 3.1 – Expert Scores for 375 Files

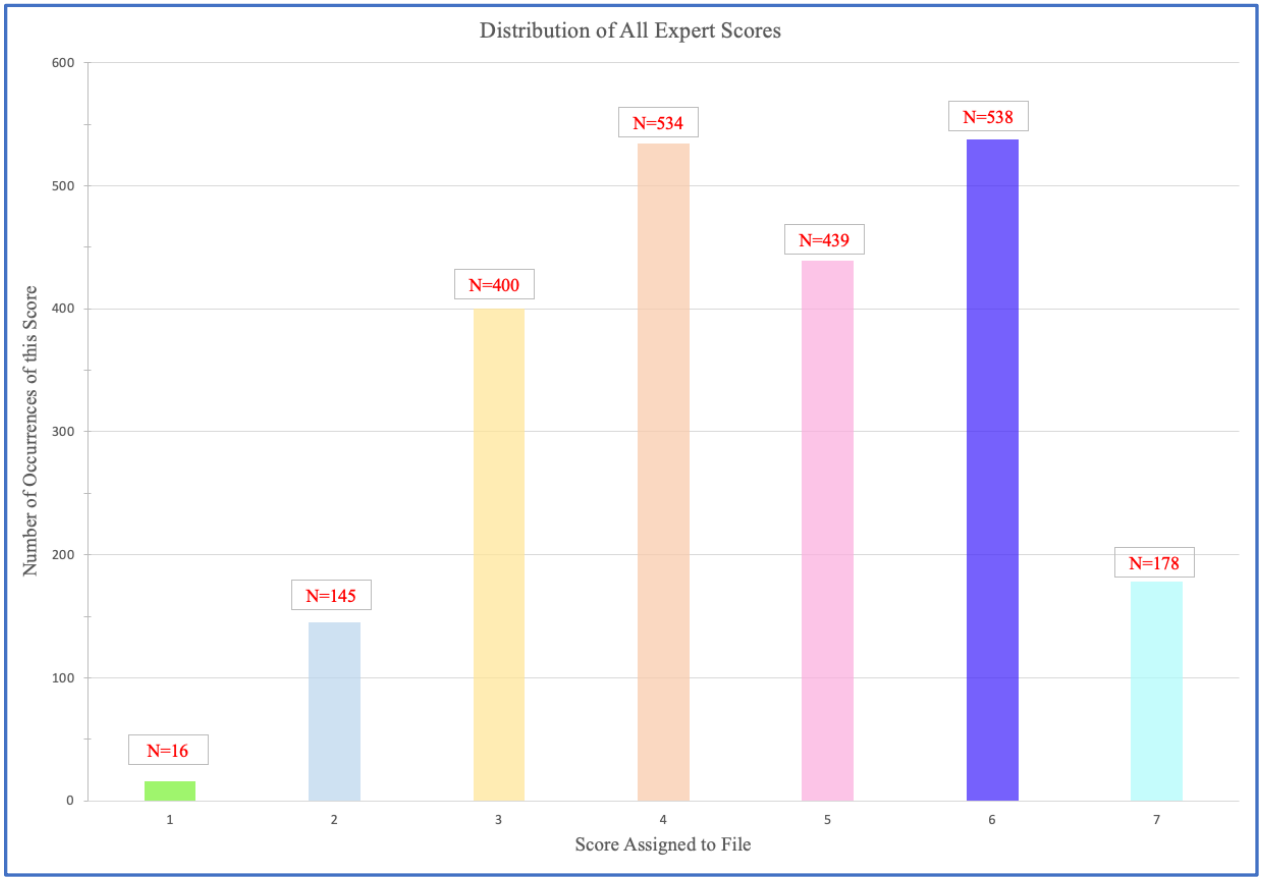


A score of "1" (signaling that the text could be understood by an adult with no secondary school education, or who is less than fluent in English) was assigned only 16 times, and this score was only used by 3 of the 6 experts. The extremely low number of "1" scores was somewhat surprising, because 40 of the 375 decisions within my corpus had been selected from tribunals where I expected the decisions to be highly readable, and because at least some scores of "1" could otherwise have been expected to materialize organically from my process of building what I considered to be a diverse corpus of adjudicative decision texts. My objective at this stage of the dissertation was not to assess the readability results for Canadian adjudicative decisions solely from the expert scores that were assigned to files within my reference corpus (since studies within subsequent chapters of this dissertation will attempt to make this type of assessment on the basis

of large samples of complete decisions). However, the fact that scores of "1" were assigned so infrequently (only 0.7% of the time) suggests that less-educated adults and individuals with more limited fluency in English may struggle to read and understand adjudicative decisions. I will examine this hypothesis more closely in discussing the results within both Chapters 4 and 6 of this dissertation.

The full distribution of occurrences of each score assigned by the experts is shown below in Chart 3.2.

Chart 3.2 – Distribution of All Expert Scores



These scores seem to reflect what researchers call "central tendency bias" in Likert scoring responses.[45] This bias, in theory, "inclines participants to avoid the endpoints of a response scale and to prefer responses closer to the midpoint," and is perceived as one of the most persistent response biases in psychological research.[46] However, as Douven argues, this perceived bias ("something that, ideally, we would like to be absent from our data")[47] may actually be misrepresented. Douven theorizes that it is possible "we have been wrong in thinking of it as a bias, threatening the validity of our research, and as something for which we should try to statistically correct."[48] Douven suggests instead that "we can conclude that data from Likert scale questions may be perfectly in order if they reveal a central tendency bias"[49] because this central phenomenon often accurately reflects our subjective views about the things that we observe around us.

In the present case, the original 7-point Likert scale that I developed for use by my experts was intended to include scoring extremes that were likely to describe even the most and least readable adjudicative decision texts. I designed the scale this way to ensure that no decision would clearly exceed the scale (and therefore be inaccurately rateable) at either the high or low end of the spectrum – even if only a low number of decisions would receive extreme scores. Consequently, I am not surprised by the overall central tendency phenomenon that appears in the experts' scores (notwithstanding my surprise at the very low number of "1" scores), nor do I find it to be concerning from a data validity or reliability perspective.

---

[45] Igor Douven, "A Bayesian Perspective on Likert Scales and Central Tendency" (2018) 25:3 Psychonomic Bulletin & Review 1203.
[46] Douven, "A Bayesian Perspective on Likert Scales and Central Tendency", *supra* note 45 at 1203.
[47] Douven, "A Bayesian Perspective on Likert Scales and Central Tendency", *supra* note 45 at 1209.
[48] Douven, "A Bayesian Perspective on Likert Scales and Central Tendency", *supra* note 45 at 1209.
[49] Douven, "A Bayesian Perspective on Likert Scales and Central Tendency", *supra* note 45 at 1209.

In order to assess the measure of agreement between my 6 expert raters, I calculated Kendall's Coefficient of Concordance (Kendall's *W*). "This statistic assesses the overall degree of agreement in a set of rankings given by several individuals."[50] Specifically, Kendall's *W* was run using SPSS software and the individual scores from each expert for each file to determine if there was agreement between the 6 experts' judgement on the readability of the 375 text files that they scored. The 6 experts statistically significantly agreed in their assessments, $W = 0.67$, $p < 0.001$. Although there is no intuitive meaning for the value of this statistic, Landis & Koch suggest that a value of between 0.61 and 0.8 for a statistic like a *W*-value signifies "substantial" strength of agreement between raters.[51] Others have suggested that "strong consensus" exists for *W*-values of 0.7 and above, while "moderate consensus" exists for *W*-values in the range of 0.5 to 0.69.[52] However, in the context of social science and behavioural research, specifically, Kraska-Miller has said that "Kendall's W can be interpreted as a correlation with weak, moderate, and strong agreement denoted as 0.10, 0.30, and 0.50, respectively."[53] Based on these characterizations of consensus strength from different contexts, and the calculated *W*-value, I assess that there is substantial agreement between the expert raters such that their scores are sufficiently reliable for the purposes of Undertaking #1.

3.3.2.2 Lay Scoring Results – Discussion and Validation of Expert Scoring

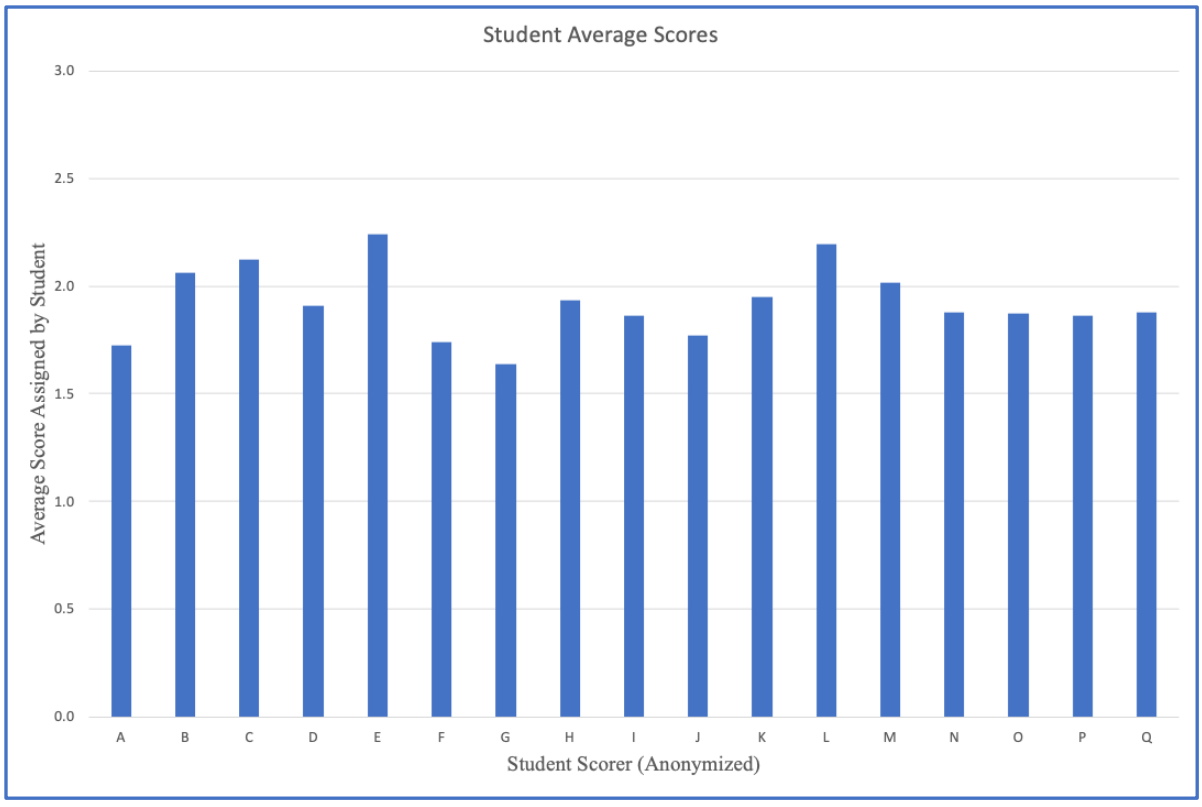
As noted above, the design of Undertaking #1 also included lay rankings for each of the 375 files within my corpus. These lay scores were intended for use as an additional means of

---

[50] Ángel García-Crespo, et al, "A Qualitative Study of Hard Decision Making in Managing Global Software Development Teams" (2010) 27 Information Systems Management 247 at 248.
[51] Landis & Koch, "The Measurement of Observer Agreement for Categorical Data", *supra* note 38 at 165.
[52] Roy Schmidt, et al, "Identifying Software Project Risks: An International Delphi Study" (2001) 17:4 Journal of Management Information Systems 5 at 13.
[53] M Kraska-Miller, *Nonparametric Statistics for Social and Behavioral Sciences* (Boca Raton, FL: Taylor & Francis, 2014), at 191. *See also*, Riccardo Russo, *Statistics for the Behavioural Sciences: An Introduction* (New York: Taylor & Francis, 2003) at 201: "As a rule of thumb, if W is less than 0.5 there is no substantial agreement between judges."

validating the experts' scores. Each the 375 files in my corpus was read and scored by 3 different Grade 12 student volunteers, for a total of 1125 independent scoring observations by the 17 student volunteers, that involved approximately 103 hours of collective volunteer time.[54] The average score across all 1125 student observations was 1.9 (suggesting that the text could be understood by someone with about a Grade 12 education or slightly more). Individually, the students assigned average scores of between 1.6 (for the student with the lowest average text scores) and 2.2 (for the two students with the highest average text scores).[55] The average score assigned by each student is shown below, in Chart 3.3.

Chart 3.3 – Student Average Scores

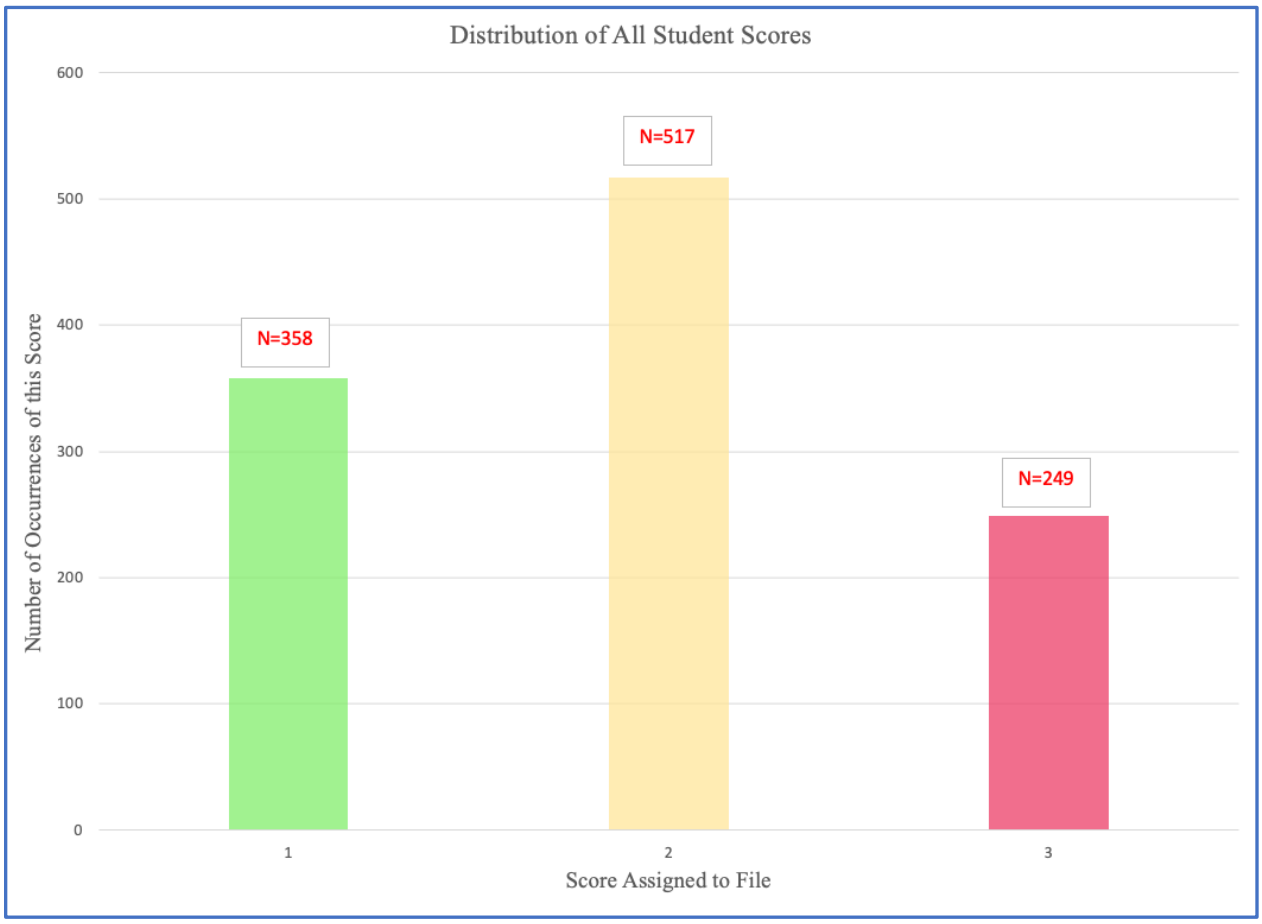


[54] A spreadsheet containing individual and average student scores for all 375 files is publicly available: Mike Madden, "MADRS_MaddenDissertation_Chapter3_375FilesStudentScores" (30 May 2024), online: <https://doi.org/10.7910/DVN/GJYYRB>, Harvard Dataverse, V.

[55] In ascending order, the students' individual average assigned scores were as follows: 1.6 / 1.7 / 1.7 / 1.8 / 1.9 / 1.9 / 1.9 / 1.9 / 1.9 / 1.9 / 1.9 / 2.0 / 2.0 / 2.1 / 2.1 / 2.2 / 2.2.

The full distribution of occurrences of each score assigned by the students is shown below in Chart 3.4.

Chart 3.4 – Distribution of All Student Scores



In order to compare the relationship between expert and lay raters' scores, the Spearman rank-order correlation (Spearman's rho) was calculated. This statistic measures the direction and strength of any association between two variables (in this case, expert and lay raters' scores), but it does so based on the relative ranks of the different scores contained within the dataset, rather than on the raw scores.[56] A rank-order correlation assessment like Spearman's rho allows for the identification of significant relationships between different datasets that are not necessarily using

[56] Michael O Finkelstein & Bruce Levin, *Statistics for Lawyers*, 3rd ed (New York: Springer, 2015) at 366.

the same scale (for instance, where one rater tends to use the lower half of a scoring scale and another uses the upper half of the scale; or, as in this case, where one set of scorers uses a 7-point Likert scale, and the other uses a 3-point Likert scale).[57] Because Spearman's rho uses rank-order scores instead of raw numerical scores, the measurement "trades a slight loss of efficiency relative to the ordinary sample correlation coefficient […] in exchange for applicability to a much broader array of problems for which rank correlation alone is valid, since it does not require specific distributional assumptions."[58] In other words, the statistic tends to report slightly reduced correlation strengths, but it is a valid measure of correlation between variables across a wider spectrum of situations than other types of correlation measures.

Specifically, I calculated the average expert rater score (based on the 6 expert scores) and the average lay rater score (based on 3 student scores) for each file within the 375-file corpus. I then calculated the Spearman rho for these two variables using SPSS software. Preliminary analysis showed the relationship between these two variables to be monotonic, as assessed by visual inspection of a scatterplot. There was a statistically significant, positive correlation between lay and expert raters' average scores, Spearman's rho = 0.59, $p < 0.001$. Cohen suggests that a correlation strength of 0.5 or more represents a strong effect size in this type of research context, and notes that correlations of approximately 0.5 tend to be produced in studies looking at the relationship between IQ or aptitude tests, and student grades in school – where some of the strongest correlations in social science research are seen.[59] For the purposes of Undertaking #1,

[57] Kevin A Hallgren, "Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial" (2012) 8:1 Tutorials in Quantitative Methods for Psychology 23 at 30.
[58] Finkelstein & Levin, *Statistics for Lawyers*, *supra* note 56 at 366.
[59] Cohen, *Statistical Power Analysis for the Behavioral Sciences*, *supra* note 39 at 80-81.

this strong correlation provides an additional measure of the validity and reliability of the expert raters' scores.

3.3.2.3 Step 1 - Gold Standard Scores Established

With confirmation from these different statistical tests about the reliability and validity of my expert raters' scores, I determined that it was acceptable to proceed to the next major step of my Formula-Building Project's method, having established human-generated "gold standard" scores for each of the 375 files in my corpus, based on the average of the 6 experts' scores for the respective files.

3.4 Step 2 – Derive a New Law-Specific Readability Formula

3.4.1 Overview of Method – Derive a New Law-Specific Readability Formula

In order to derive a new law-specific readability formula, I followed a broad method that in many ways mimics the methods used by almost all other scholars who have developed past readability formulas: I use correlations and regression analysis to identify which linguistic variables, with which weight, are best capable of predicting the "gold standard" scores that my experts assigned to a text.[60] More specifically, my method for deriving a law-specific readability formula lines up closely with the method that was recently used by Crossley, Skalicki & Dascalu in creating their general-purpose CAREC readability formula.[61]

3.4.2 Method Step 2A – Identify and Calculate Scores for Relevant Linguistic Variables

To begin with, I split my 375-file corpus into two sub-sets. The first sub-set, consisting of 250 (two-thirds) of the corpus files that were chosen randomly, was earmarked for formula

---

[60] Cunningham, Hiebert & Mesmer, "Investigating the Validity of Two Widely Used Quantitative Text Tools", *supra* note 8 at 814: "Almost all readability formulas, past and present, are regression equations."
[61] Crossley, Skalicky & Dascalu, "Moving Beyond Classic Readability Formulas", *supra* note 2.

development. The second sub-set, consisting of the remaining 125 (one-third) files from the corpus, was earmarked for formula validation and testing. As Snee has suggested, it is important to create or preserve independent data for testing the performance of any predictive regression model, but it can be cumbersome to collect entirely new data for this purpose; consequently, Snee endorses the type of "data-splitting" that I have conducted in this element of my research as a "reasonable way" to proceed that achieves the desired effect without re-starting the data collection process.[62] Crossley, Skalicki & Dascalu used a similar data-splitting technique when validating their CAREC readability formula.[63]

Next, I identified and measured over one thousand different linguistic properties of each file within my 250-file 'formula-building' sub-set. Below, I briefly describe the large classes of linguistic variables that I measured, explain my rationale for choosing to measure these types of variables because of their potential effects on text readability, and offer some examples of each type of variable.

3.4.2.1 Lexical Sophistication

Lexical or semantic sophistication refers to linguistic features at the word or phrase level: "[r]eflecting the importance of vocabulary in readability, lexico-semantic features capture attributes associated with the difficulty or unfamiliarity of vocabulary, i.e. specific words or phrases in a text."[64] There is little doubt amongst readability and language scholars that features such as the average word difficulty for words in a text will tend to affect a reader's ability to

[62] Ronald D Snee, "Validation of Regression Models: Methods and Examples" (1977) 19:4 Technometrics 415, at 416-420.
[63] Crossley, Skalicky & Dascalu, "Moving Beyond Classic Readability Formulas", *supra* note 2.
[64] Kevyn Collins-Thompson, "Computational Assessment of Text Readability: A Survey of Current and Future Research" (2014) 165:2 International Journal of Applied Linguistics 97, at 106 ["Computational Assessment of Text Readability"].

understand the text.[65] Although older language models tended to measure word difficulty rather simplistically as a function of word length,[66] a now "widely-used feature of lexical difficulty for a word is the relative frequency of that word in everyday usage, as measured by its relative frequency in a large representative corpus."[67] As Kyle has noted, "[h]ighly frequent lexical items seem to be learned first and are therefore considered less sophisticated, while less frequent words are learned later (if at all), and are therefore considered more sophisticated."[68] A slightly different measurement considers the breadth of different documents within which one might encounter a word or phrase – again as a potential means of quantifying readers' familiarity with the word or phrase for comprehension purposes: "[d]ispersion, or the range of corpus texts an item occurs provides an indication of the likelihood of encountering a particular [word or phrase] in a wide range of texts."[69]

However, other factors beyond just word difficulty measurements may also have impacts on comprehension: psycholinguistic lexical features (like the concreteness of a word) increasingly feature within readability formulas because they capture "cognitive aspects of reading not directly addressed by the surface vocabulary and syntax features of traditional formulas."[70] Additionally, recent research has shown that the use of more common multi-word phrases (or n-grams) generates faster response times from those who strive to understand the phrases[71] – which suggests that a

---

[65] Xiaobin Chen & Detmar Meurers, "Word Frequency and Readability: Predicting the Text-Level Readability with a Lexical-Level Attribute" (2018) 41:3 Journal of Research in Reading 486, at 486-488 ["Word Frequency and Readability"].
[66] Alan Bailin & Ann Grafstein, *Readability: Text and Context* (New York: Palgrave Macmillan, 2016) at 53.
[67] Collins-Thompson, "Computational Assessment of Text Readability", *supra* note 64 at 106.
[68] Kristopher Kyle, *Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication* (unpublished Ph.D. Dissertation, Georgia State University, 2016), online: <http://scholarworks.gsu.edu/alesl_diss/35>, at 1 [*Measuring Syntactic Development in L2 Writing*].
[69] James Garner, Scott Crossley & Kristopher Kyle, "N-Gram Measures and L2 Writing Proficiency" (2019) 80 System 176 at 178.
[70] Collins-Thompson, "Computational Assessment of Text Readability", *supra* note 64 at 107.
[71] Inbal Arnon & Neal Snider, "More Than Words: Frequency Effects for Multi-Word Phrases" (2010) 62 Journal of Memory and Language 67 ["More Than Words"].

reader's familiarity with not just words themselves, but with the orders in which words are arranged together, can have an impact on comprehension.

Because of the well-established connection between semantic- or lexical-level properties of language and our abilities to understand this language, I decided to measure a large cross-section of lexical sophistication variables for each of the 250 texts in my formula-building subset. I processed each file through the *Tool for the Automatic Analysis of Lexical Sophistication* ("TAALES")[72] software. This software measures "over 400 classic and new indices of lexical sophistication, and includes indices related to a wide range of sub-constructs."[73] In order to save processing time, I did not compute scores for all possible variables. Specifically, I elected to run each text against the "spoken" and "academic" sub-databases from the Corpus of Contemporary American English master database, but did not also use the "fiction", "magazine", or "news" databases. I assessed that these two chosen sub-databases adequately captured the spectrum of language complexity that people might encounter when reading, from relatively easier texts that more closely mimic spoken English,[74] to relatively more difficult texts such as those drawn from academic journals and books. I selected every other variable for computation within TAALES. Consequently, TAALES calculated scores for each of my files in relation to 221 discrete variables

[72] Kristopher Kyle, Scott Crossley & Cynthia Berger. "The Tool for the Automatic Analysis of Lexical Sophistication (TAALES): Version 2.0" (2018) 50 Behavior Research Methods 1030.
[73] Online: <https://www.linguisticanalysistools.org/taales.html>.
[74] Chen & Meurers, "Word Frequency and Readability", *supra* note 65 at 490, are critical of corpora that do not include materials consisting of spoken English: these corpora "failed to represent the also important, if not more important, source of spoken language that students are exposed to. The amount of exposure from spoken language is much greater than from written ones. Failure to include the spoken language does great harm to the representativeness of the frequency list and to its predictive power for text readability."

related to lexical or semantic sophistication.[75] The output was saved in spreadsheet format that included 55,250 measurements (i.e.: 221 variables for each of 250 files).[76]

3.4.2.2 Syntactic Sophistication and Complexity

"Syntax refers to the systematic ways in which discrete units (e.g., words) can be combined to create meaningful utterances (e.g., sentences)."[77] Kyle observes that, generally, as language learners develop, they "produce longer and more varied syntactic structures,"[78] which suggests that greater experience and competence with a language is needed in order to understand more complex and sophisticated phrase and sentence structures. Kyle distinguishes between syntactic *complexity* (which refers to the absolute properties of a sample of language, such as the number of words per sentence, the number of words before the main verb in a sentence, or the number of dependent clauses per sentence), and syntactic *sophistication* (which refers to the relative difficulty of understanding a particular language construct largely based on usage data about how frequently one might encounter this type of language construct).[79]

Since it is accepted that "[s]yntactic complexity is known to be associated with longer processing times in comprehension and is a widely-used factor included in automated readability assessment,"[80] I chose to measure a number of different syntactical complexity and sophistication variables for each of the 250 texts in my formula-building subset. I processed each file through the

---

[75] For illustrative purposes, some of the variables included the following: average word frequency and range counts from the Brown, London-Lund, and Lorge corpora, respectively; average word familiarity scores from the Medical Research Council's Psycholinguistic Database; average Brysabaert Concreteness scores for words in the text; and, the average frequency score for Bigrams (two-word phrases) in a text, based on occurrences of the bigrams in the Corpus of Contemporary American English – Spoken – database.
[76] A spreadsheet containing all TAALES results for the 221 linguistic variables is publicly available: Mike Madden, "MADRS_MaddenDissertation_Chapter3_250FilesTAALESResults" (30 May 2024), online: <https://doi.org/10.7910/DVN/OTDPQI>, Harvard Dataverse, V1.
[77] Kyle, *Measuring Syntactic Development in L2 Writing*, *supra* note 68 at 1-2.
[78] Kyle, *Measuring Syntactic Development in L2 Writing*, *supra* note 68 at 2.
[79] Kyle, *Measuring Syntactic Development in L2 Writing*, *supra* note 68 at 8-34.
[80] Collins-Thompson, "Computational Assessment of Text Readability", *supra* note 64 at 108.

Tool for the Automatic Analysis of Syntactic Sophistication and Complexity ("TAASSC") software application,[81] which measures "indices related to syntactic development. Included are classic indices of syntactic complexity (e.g., mean length of T-unit) and fine-grained indices of phrasal (e.g., number of adjectives per noun phrase) and clausal (e.g., number of adverbials per clause) complexity."[82] TAASSC calculated scores for each of my files in relation to 370 discrete variables related to syntactic complexity or sophistication.[83] The output was saved in spreadsheet format that included 92,500 measurements (i.e.: 370 variables for each of 250 files).[84]

3.4.2.3 Sentiment and Cognition Properties

Although classic readability formulas tended to only incorporate measurements of syntactic and lexical complexity, more recent thinking about readability has suggested that other text properties, such as the overall genre or sentiment of a text may also affect a reader's motivation to understand the text, and therefore the reader's overall comprehension of the text.[85] At least one recent readability study found that the use of positive adjectives was strongly and positively correlated with readability results, and relied upon this sentiment-related variable within the reading comprehension formula that the study produced.[86]

With advances in Natural Language Processing technologies, it has become increasingly possible to measure a vast array of different sentiment-related properties of a text through

---

[81] K Kyle & SA Crossley, "Measuring Syntactic Complexity in L2 Writing Using Fine-Grained Clausal and Phrasal Indices" (2018) 102:2 Modern Language Journal 333.
[82] Online: <https://www.linguisticanalysistools.org/taassc.html>.
[83] For illustrative purposes, some of the variables included the following: average number of prepositions per clause; average number of possessives per noun subject; average number of conjunctions per clause; and, average number of verb-phrases per T-unit (essentially, a T-unit is the smallest unit into which text can be divided while still producing a complete and grammatically correct sentence).
[84] A spreadsheet containing all TAASSC results for the 370 linguistic variables is publicly available: Mike Madden, "MADRS_MaddenDissertation_Chapter3_250FilesTAASSCResults" (30 May 2024), online: <https://doi.org/10.7910/DVN/G4Q1DE>, Harvard Dataverse, V1.
[85] Collins-Thompson, "Computational Assessment of Text Readability", *supra* note 64 at 110.
[86] Crossley, Skalicky & Dascalu, "Moving Beyond Classic Readability Formulas", *supra* note 2 at 553 (Table 2).

sentiment analysis, which is "a general term related to extracting subjective information related to human feelings and opinions from natural language texts."[87] Sentiment analysis is generally performed by comparing the language in a text against other word lists or dictionaries that have been tagged to associate particular words with particular emotions (e.g.: anger, joy, surprise, guilt), topics (e.g.: economics, military matters, religion), or cognitive states (e.g.: thinking, knowing, perceiving, evaluating) that may all evoke or relate to feeling and cognition in different contexts.[88]

Because sentiment-related language may have an impact on readability, I opted to measure a large range of sentiment- and cognition-related variables for each of the 250 texts in my formula-building subset. I processed each file through the *Sentiment Analysis and Cognition Engine* ("SEANCE") software,[89] which compares language in a particular text against "254 core indices and 20 component indices based on recent advances in sentiment analysis,"[90] and can report sentiment measurements from these different indices distinctly for any or all of the following parts of speech (based on the user's selections): all words, verbs, nouns, adjectives, and adverbs. In order to save processing time,[91] I processed my 250 files through SEANCE for comparison against only the VADER, EmoLex, and General Inquirer sentiment dictionaries, for all of the available part-of-speech categories. SEANCE calculated scores for each of my files in relation to 441 discrete

---

[87] 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 804 ["SEANCE"].
[88] Crossley, Kyle & McNamara, "SEANCE", *supra* note 87 at 807-08.
[89] Crossley, Kyle & McNamara, "SEANCE", *supra* note 87.
[90] Online: <https://www.linguisticanalysistools.org/seance.html>.
[91] See Crossley, Kyle & McNamara, "SEANCE", *supra* note 87 at 808: "With the potential for each index to report results for all words, nouns, verbs, and adjectives, in addition to each of these having the potential to be negated, the SEANCE tool can report on almost 3,000 indices."

variables related to sentiment or cognition.[92] The output was saved in spreadsheet format that included 110,250 measurements (i.e.: 441 variables for each of 250 files).[93]

3.4.2.4 Cohesion

We know that "[t]ext is more than a series of random sentences: language exhibits higher-level, longer-range structure by virtue of the dependencies and relationships that exist between its elements."[94] And, "[l]ike syntactic complexity and vocabulary difficulty, cohesion is a theoretical construct believed to be involved in determining reading ease or difficulty."[95] McNamara & Graesser have described the importance of cohesion relative to text comprehension as follows:

> Cohesion is the linguistic glue that holds together the events and concepts conveyed within a text. Beyond the words and separate sentences in the text, are the relationships between the sentences and larger units of text. Explicit cues in the text help the reader to process, understand, or infer those relationships. A simple but powerful source of cohesion is referential and semantic overlap of adjacent sentences, pairs of sentences in a paragraph, and adjacent paragraphs. Simply put, when words, concepts, or ideas overlap between sentences, that content forms a link between the sentences.[96]

By making a text more cohesive through the overlapping or reinforcement of words, phrases, and ideas throughout a text, an author would likely also be making the text more readable (even if many older readability formulas do not rely upon measures of cohesion).[97] This is the case because a cohesive text requires less inferential work on the part of a reader:

---

[92] For illustrative purposes, some of the variables included the following: percentage of words that are positive adjectives; percentage of words that are adverbs associated with fear; and, percentage of words that are nouns associated with hostility.

[93] A spreadsheet containing all SEANCE results for the 441 linguistic variables is publicly available: Mike Madden, "MADRS_MaddenDissertation_Chapter3_250FilesSEANCEResults" (30 May 2024), online: <https://doi.org/10.7910/DVN/QVPDBS>, Harvard Dataverse, V1.

[94] Collins-Thompson, "Computational Assessment of Text Readability", *supra* note 64 at 109.

[95] Michael Flor & Beata Beigman Klebanov, "Associative Lexical Cohesion as a Factor in Text Complexity" (2014) 165:2 International Journal of Applied Linguistics 223, at 226.

[96] Danielle McNamara & Arthur C Graesser. "Coh-Metrix: An Automated Tool for Theoretical and Applied Natural Language Processing" in Philip M McCarthy & Chutima Boonthum-Denecke, eds, *Applied Natural Language Processing: Identification, Investigation, and Resolution* (Hershey, PA: IGI Global, 2012) 188 at 190 ["Coh-Metrix"].

[97] Collins-Thompson, "Computational Assessment of Text Readability", *supra* note 64 at 109: "Well-organized, cohesive content should be on average more readable than texts that are not, yet properties like cohesion are not captured by traditional readability formulas."

> [w]hen there are few or no gaps in overlap across sentences, then a text is seamlessly moving from one point of information to another while giving the reader all the help he or she needs to build new knowledge. This type of text is a highly cohesive text: inferences are explicit and the reader does not have to fill many gaps using his or her own knowledge about the topic. However, in a text where less propositional overlap exists, the reader will be required to fill gaps of information with his or her own knowledge. This type of text has low cohesion: inferences are implicit and require more work on the part of a novice reader.[98]

It is perhaps for this reason that cohesion is now commonly studied in relation to text comprehension and readability,[99] and for this reason that measures of a text's cohesiveness have factored into more recent readability formulas.[100]

Where cohesive properties of a text are theoretically related to the readability of the text, and where others have relied upon measures of cohesion within previous readability formulas, I elected to measure a variety of cohesion-related variables for each of the 250 texts in my formula-building subset. I processed each file through the *Tool for the Automatic Analysis of Text Cohesion* ("TAACO")[101] software, which calculates "150 indices of both local and global cohesion, including a number of type-token ratio indices (including specific parts of speech, lemmas, bigrams, trigrams and more), adjacent overlap indices (at both the sentence and paragraph level), and connectives indices."[102] Based on my processing query (which included searches for word-based indicators of cohesion, but not N-gram/multiword phrase-based indicators of cohesion),

---

[98] Rebekah George Benjamin, "Reconstructing Readability: Recent Developments and Recommendations in the Analysis of Text Difficulty" (2012) 24 Educational Psychology Review 63 at 70.
[99] McNamara & Graesser. "Coh-Metrix", *supra* note 96 at 190: "Many studies, across a variety of paradigms and dependent measures, have shown that cohesive cues in text facilitate reading comprehension and help readers construct more coherent mental representations of text content."
[100] See, e.g., Danielle S McNamara, et al, *Automated Evaluation of Text and Discourse with Coh-Metrix* (New York, Cambridge UP, 2014) at 76-77 and 84-95 (describing the authors' Text Ease and Readability Assessor that measures referential and deep cohesion, among other variables); see also, Crossley, Skalicky & Dascalu, "Moving Beyond Classic Readability Formulas", *supra* note 2 at 553 (Table 2) (showing how the authors relied upon the presence of temporal connecting words – an indicator of temporal cohesion within a text – as a variable that influences their readability formula).
[101] SA Crossley, K Kyle & DS McNamara, "The Tool for the Automatic Analysis of Text Cohesion (TAACO): Automatic Assessment of Local, Global, and Text Cohesion." (2016) 48 Behavior Research Methods 1227.
[102] Online: <https://www.linguisticanalysistools.org/taaco.html>.

TAACO calculated scores for each of my files in relation to 83 discrete variables related to cohesion.[103] The output was saved in spreadsheet format that included 20,750 measurements (i.e.: 83 variables for each of 250 files).[104]

3.4.2.5 Conclusion – Identification and Scoring of Linguistic Variables

As the above description of this part of my experimental method suggests, this element of my research was data-heavy. I identified and calculated scores for 1,115 different linguistic variables relating to lexical sophistication, syntactic complexity and sophistication, sentiment and cognition properties, and textual cohesion, for each of the 250 texts in my formula-building subset of files that had been scored by my expert raters. This endeavour produced 278,750 individual data points that communicated something of potential or theoretical relevance about the readability of the texts within this subset of my "gold standard" corpus of adjudicative decision text files.

3.4.3 Method Step 2B – Statistical Analysis to Create a Regression Readability Formula

After collecting data on all of these different linguistic variables, the next step in my method required me to determine which of the 1,115 variables demonstrated the strongest relationships with the readability scores that my experts assigned to each text file in my formula-building subset of files.

Using SPSS software, I calculated Pearson correlation coefficients to assess the relationship between each of the 1,115 variables and the variable of *average expert score*. I then

---

[103] For illustrative purposes, some of the variables included the following: average adjacent sentence overlap – all lemmas (the average number of lemma types – words that share a common root – that occur at least once in the next sentence); average number of synonym overlaps across paragraphs for verbs (showing how words or their synonyms appear in successive paragraphs); percentage of words in the text that are sentence-linking words (like "nonetheless", "therefore", "although"); and, percentage of words in the text that are ordering words or temporal connectives (like "to begin with", "next", "first").

[104] A spreadsheet containing all TAACO results for the 83 linguistic variables is publicly available: Mike Madden, "MADRS_MaddenDissertation_Chapter3_250FilesTAACOResults" (30 May 2024), online: <https://doi.org/10.7910/DVN/BL8TDW>, Harvard Dataverse, V1.

eliminated any variables where there was no significant correlation ($p > 0.05$), and variables where the correlation strength was less than $r = 0.325$ (either positive or negative), to reduce the list to only the variables that had shown the strongest relationship with *average expert score*.

Finally, I eliminated any variables showing multicollinearity (i.e.: situations where two or more variables were correlated with one another at a strength of greater than $r = 0.6$, positive or negative), and kept only the variable that was most strongly correlated with *average expert score* in such cases. This process of identifying the variables that were most strongly, and independently (i.e.: free from multicollinearity), correlated with *average expert score* reduced the list of variables from 1,115 down to 20 variables. The breakdown of these remaining variables by linguistic category was as follows: syntactic sophistication and complexity = 8 variables; lexical sophistication = 7 variables; sentiment and cognition properties = 4 variables; and, textual cohesion = 1 variable. In other words, there was at least one variable that was independently correlated with *average expert score* from each class of linguistic variables that were calculated using the 4 different linguistic software applications that were described above. This result provided some *post-facto* justification or reassurance for the decision to study each category of linguistic variable in the first place.

A hierarchical linear regression analysis was then run within SPSS software, using the selected linguistic properties as the independent variables to analyse which linguistic features predicted *average expert score* for the 250 files in the formula-development subset. Non-significant variables ($p<0.05$) were removed from the equation and the regression was re-run without these variables to obtain the final prediction equations. This analysis yielded a significant

model, $F(6, 243) = 59.243$, $p < 0.001$, $r = 0.771$, $R^2 = 0.594$.[105] Ultimately, six variables were significant predictors of *average expert score*, including variables related to lexical sophistication (concreteness, imageability, age of acquisition, and trigram proportion scores), syntactic complexity (dependents per nominal – no pronouns – standard deviation) and sentiment (legal nouns).[106] The R^2 value from the model above suggests that this model should be capable of explaining 59.4% of the variance in *average expert score* – well above the threshold for what Cohen would classify as a "large effect size."[107]

[105] There was independence of residuals, as assessed by a Durbin-Watson statistic of 1.882. There was homoscedasticity, as assessed by visual inspection of a plot of studentized residuals versus unstandardized predicted values. There were no no leverage values greater than 0.1, and no values for Cook's distance above 0.1. The assumption of normality was met, as assessed by a Q-Q Plot.

[106] A spreadsheet of the linguistic variable scores for each of the 6 variables, and the average expert score, for each of the 250 files that were used to create the MADRS formula, is publicly available: Mike Madden, "MADRS_MaddenDissertation_Chapter3_LinguisticVariableScores250Files" (May 30, 2024) online: <https://doi.org/10.7910/DVN/NCQYJ1>, Harvard Dataverse, V1.

[107] Jacob Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed (New York: Lawrence Erlbaum Associates, 1988) at 80. Cohen suggests that R^2 values of more than 0.25 can be considered as representing large effect sizes.

Results from the regression are included below in Table 3.1.

Table 3.1 – Summary of Multiple Regression Model for *MADRS*

MADRS	***B***	**95% *CI* for *B***		***SE B***	***β***	R^2	ΔR^2
		LL	***UL***				
Model						0.594	0.584
Constant	10.036*	5.081	14.991	2.516			
Kuperman Age of Acquisition (Content Words) / ***AoA_CW***	0.471*	0.202	0.741	0.137	0.209*		
Medical Research Council – Imageability (All Words) / ***MRC_Imag_AW***	-0.023*	-0.032	-0.013	0.005	-0.246*		
General Inquirer Database – Legal (Nouns) / ***GI_Leg_Nouns***	4.513*	3.215	5.811	0.659	0.292*		
COCA – Spoken (Proportion of Trigrams in the top 30K) / ***COCA_Spoken_TriTop30K***	-0.650**	-10.641	-2.365	2.101	-0.151**		
Dependents per nominal (no pronouns, standard deviation) / ***Nominal_Deps_NN_StDev***	1.802*	1.106	2.499	0.354	0.221*		
Medical Research Council – Concreteness (Function Words) / ***MRC_Concr_FW***	-0.018**	-0.029	-0.006	0.006	-0.165**		

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. * $p < 0.001$; ** $p = 0.002$

The *B*-weights and the constant from this regression model are the key statistical elements that translate the regression data into a new law-specific readability formula. For predictive purposes, this type of regression analysis creates a formula where the value of the dependent

variable, X (representing, in this case, the predicted *average expert score*) can be calculated by the following equation:

$$X = B_0 + (B_1V_1) + (B_2V_2) + [\ldots] + (B_6V_6)$$

where B_0 is the regression constant; B_1, B_2, etc, are the regression B-values for each variable; and V_1, V_2, etc., are the document-specific values for the respective linguistic variables (as calculated for any particular text by the different linguistic software applications that I have described above).[108]

Thus, the new law-specific readability formula that I have created, called *Madden's Adjudicative Decision Readability Score* (MADRS) can be represented by the following equation:

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***)

If the values of the relevant 6 variables are known for any particular adjudicative decision text, then the above equation yields a statistical prediction of the score that my human experts would have assigned to that text. In other words, with this formula, a computer can quickly and effectively calculate the readability score of an adjudicative decision without the need for an expert or other human to actually sit down and read the document in its entirety. It makes readability assessments, at scale, into a much more practical possibility. (Further explanations for each of these variables,

[108] See Snee, "Validation of Regression Models: Methods and Examples", *supra* note 62 at 415 (for an explanation of how regression coefficients and constants produce predictive equations in this manner).

and how they can be theoretically justified for inclusion in my readability formula, are included below, at section 3.5).

3.4.4 Results: Validation and Testing of New Formula Against Existing Formulas

Before a regression equation can be used with confidence, one must check the validity of the equation.[109] A preferred approach for validating a regression formula typically involves assessing the predictive abilities of the formula when it is applied to a new dataset that was not involved in the creation of the formula.[110] The most common approach to finding and using new data is called data-splitting, wherein a portion of any initially gathered data is withheld, and reserved for formula validation purposes.[111] When using a data-splitting approach, it is also common to reserve one-third of the data for formula validation, and keep two-thirds of the data for formula development.[112]

Developers of contemporary linguistic readability formulas have adopted this type of data-splitting approach when validating their formulas. In one example, the researchers derived a readability assessment tool from expert judgments about how difficult it was to read medical communications messages.[113] These researchers used the tool to then classify an independent set of medical messages into categories based on their linguistic difficulty levels.[114] The researchers found that the tool correctly classified the messages in 75% of the cases, and based their determination that the tool was statistically valid on this independent test result.[115] In another

---

[109] Snee, "Validation of Regression Models: Methods and Examples", *supra* note 62 at 416.
[110] Kenneth N Berk, "Validating Regression Procedures With New Data" (1984) 26:4 Technometrics 331 at 331.
[111] Joel H Steckel & Wilfried R Vanhonacker, "Cross-Validating Regression Models in Marketing Research" (1993) 12:2 Marketing Science 415 at 416.
[112] C-X J Feng, Z-G S Yu & J-H J Wang, "Validation and Data Splitting in Predictive Regression Modeling of Honing Surface Roughness Data" (2005) 43:8 International Journal of Production Research 1555 at 1557.
[113] Crossley, et al, "Physicians' Secure Messages", *supra* note 10.
[114] Crossley, et al, "Physicians' Secure Messages", *supra* note 10 at 349.
[115] Crossley, et al, "Physicians' Secure Messages", *supra* note 10 at 349.

example, researchers derived a readability formula from crowd-sourced assessments of the relative difficulty levels of pairs of text samples.[116] The researchers then used this formula to measure the readability levels of files from an independent data set, and found that their readability scores were strongly correlated (r = 0.537) with, and explained a substantial amount of the score variance (R^2 =).288), the original human-assessed difficulty levels for the texts.[117] The researchers relied on these independent test results as a measure of the validity of their formula, which is now publicly available for use through both web-based[118] and software[119] tools.

To be clear, not all developers of linguistic readability formulas have taken this approach to validating their formulas. Flesch, for instance, does not describe at all how he determined that his Flesch Reading Ease formula was valid.[120] Dale and Chall (for their revised 1995 formula) asserted their formula's validity based on the formula's strong correlation (r = 0.92) with Cloze score test results for reading comprehension passages and associated grade levels of students who completed the Cloze tests for those passages.[121] However, as Bailin & Grafstein point out, these Cloze test reading comprehension passages were the same ones that Dale and Chall used to create their formula,[122] so their validation measure is not independent of their formula-building dataset. The same problem arises within the validation measures used by other researchers who relied upon identical Cloze test reading comprehension passages for both building and validating their readability formulas.[123] Validation in these cases is somewhat circular, and therefore suspect.

---

[116] Crossley, Skalicky & Dascalu, "Moving Beyond Classic Readability Formulas", *supra* note 2.
[117] Crossley, Skalicky & Dascalu, "Moving Beyond Classic Readability Formulas", *supra* note 2 at 552.
[118] Online: <https://nlp.gsu.edu/text_editor>.
[119] Online: <https://www.linguisticanalysistools.org/arte.html>.
[120] Flesch, "A New Readability Yardstick", *supra* note 2.
[121] Bailin & Grafstein, *Readability: Text and Context*, *supra* note 66 at 40.
[122] Bailin & Grafstein, *Readability: Text and Context*, *supra* note 66 at 40.
[123] Bailin & Grafstein, *Readability: Text and Context*, *supra* note 66 at 52:

> This, however, is but one instance of the more general problem: one approach bases its soundness on the degree to which its results correlate with previous approaches, which are themselves based on problematic

From among the more well-known linguistic readability formulas, McLaughlin's SMOG formula appears to be the only one wherein validation was performed by testing how the formula fared against subsequently gathered human test scores. This approach was perhaps necessary to lend credibility to the formula, which McLaughlin admits to having derived through "trial and error" in at least two respects.[124] Specifically, McLaughlin validated his formula by asking 64 university students to read independent text passages that were each approximately 1000 words long.[125] Each passage had previously been reviewed by a "literacy expert" who had identified the 10 main ideas from within the passage.[126] Immediately after reading a passage, each student was asked to say everything that they could remember about the passage; their comprehension scores were assessed on a scale of 1-10 based on how many of the experts' main ideas they recalled.[127] However, McLaughlin was, for unexplained reasons, concerned that reading speed might unfairly help comprehension, so he adjusted comprehension scores by dividing them by the time a reader spent reading a passage in order to arrive at a global score for each passage.[128] He then compared these global scores with SMOG scores for the same passages, and found that there was a perfect rank order correlation between them.[129]

No other readability formula developer had used anything like McLaughlin's validation approach. The overwhelming tendency (as seen in the examples noted above) when validating

---

> results of other studies. Proponents of cloze, as we have noted, cite its high correlation with various readability formulas. Many of the original readability formulas have based their soundness on their correlations with the McCall–Crabbs Test Lesson rankings, although, as we have seen, the soundness of these has been questioned. The revised Chall Dale formula (1995) uses the cloze rankings of texts as the independent criteria for its regression formula, as do Crossley, Greenfield, and McNamara (2008), although once again it is by no means established that these scores accurately represent readability.

[124] G Harry McLaughlin, "SMOG Grading – A New Readability Formula" (1969) 12 Journal of Reading 639 at 641 and 642 ["SMOG Grading"].
[125] McLaughlin, "SMOG Grading", *supra* note 124 at 644-645.
[126] McLaughlin, "SMOG Grading", *supra* note 124 at 645.
[127] McLaughlin, "SMOG Grading", *supra* note 124 at 645.
[128] McLaughlin, "SMOG Grading", *supra* note 124 at 645.
[129] McLaughlin, "SMOG Grading", *supra* note 124 at 645.

these formulas is to use a new formula to predict the reading difficulty level of a set of texts for which there is already a known (typically human-assigned) difficulty measurement, to see how well the formula corresponds with the other measurement. From a statistical perspective, the formulas that rely on independent data sets for validation would generally be considered to provide more robust forms of validation than formulas that rely, more circularly, on the same data sets that contributed to the formula's development in the first place.[130] Once validation is performed, researchers in all but one case that was discussed above did not conduct further testing on human subjects to re-assess how well the formula performed compared to text difficulty assessments made by real people.

I therefore decided to validate my formula using the approach that was adopted by Crossley, Skalicky, and Dascalu – who recently created the CAREC formula, and who used validation techniques that are consistent with what I perceive to be the best contemporary practices.[131] Having created MADRS from the 250-file subset of my "gold standard" corpus, and having previously split my data to keep the remaining 125 files from this corpus in reserve for the purposes of formula validation and testing, it was possible for me to compare the performance of MADRS with several other major and general-purpose readability formulas, in order to determine which formula would be most appropriate and effective for use in predicting the readability of adjudicative decision texts.

I therefore calculated MADRS scores for each of the 125 files in my 'formula validation and testing' subset (based on values for each of the 6 regression variables described above, where these values were produced for each file using the TAALES, SEANCE, and TAASSC software

---

[130] See generally, Berk, "Validating Regression Procedures With New Data", *supra* note 110.
[131] Crossley, Skalicky & Dascalu, "Moving Beyond Classic Readability Formulas", *supra* note 2 at 552.

applications). I also calculated readability scores using six other readability formulas (Flesch-Kincaid Grade Level / ***FKGL***; Automated Readability Index / ***ARI***; Simple Measure of Gobbledygook / ***SMOG***; New Dale-Chall Readability Formula / ***DaleChall***; Crowdsourced Algorithm of Reading Comprehension-Modified / ***CAREC-M***; and, Coh-Metrix L2 Readability Index / ***CML2RI***). These latter readability formula values were all calculated by processing the 125 files in my 'formula validation and testing' subset through the *Automatic Readability Tool for English*[132] software application, which quickly calculates the relevant readability scores using an intuitive interface.[133] A spreadsheet containing the readability score results for the 125 files is publicly available.[134]

I then calculated Pearson coefficients of correlation using SPSS software to assess the relationships between MADRS and the other 6 general-purpose readability formulas (on the one hand) and the "gold standard" *average expert score* that my experts had assigned to each file in my formula validation and testing subset (on the other hand). The correlation results are shown below in Table 3.2.

Table 3.2 – Correlations Between Readability Formulas and Gold Standard Scores

	MADRS	FKGL	ARI	SMOG	DaleChall	CAREC-M	CML2RI
Pearson Correlation with *Average Expert Score*	**0.756**	0.327	0.299	0.376	0.457	0.456	-0.504
Note: $p < 0.001$ for all correlations; all correlations are statistically significant.							

[132] Joon Suh Choi & Scott A Crossley, "Advances in Readability Research: A New Readability Web App for English" (2022) *Proceedings of the Twenty-second IEEE International Conference on Advanced Learning Technologies (ICALT 2022),* 6, online: < https://ieeexplore-ieee-org.proxy.bib.uottawa.ca/stamp/stamp.jsp?arnumber=9853738>.
[133] Online: <https://www.linguisticanalysistools.org/arte.html>.
[134] Mike Madden, "MADRS_MaddenDissertation_Chapter3_ARTE+MADRSScores125Files" (30 May 2024), online: < https://doi.org/10.7910/DVN/BAIVK4>, Harvard Dataverse, V1.

These results do two things. First, they affirm and validate the MADRS formula's predictive power on an independent set of adjudicative decision text files (i.e.: MADRS scores are strongly correlated with actual *average expert scores* even in relation to files that were not used in generating the MADRS readability formula). The correlation strength of r = 0.756 yields and R^2 value of 0.572; this R^2 value indicates that the MADRS formula is statistically capable of explaining 57.2% of the variance in the average expert scores that human raters assigned to the files in the independent set of adjudicative decision text files.[135] Second, the above correlations demonstrate that MADRS scores have a substantially stronger relationship with *average expert score* than do any of the other existing general-purpose readability scores. Where r = 0.756 for MADRS, and where this correlation coefficient is only -0.504 for the next strongest readability formula (CML2RI), and an even-lower 0.327 for FKGL (perhaps the most widely-used readability formula in existence, and the formula that two Canadian tribunals refer to when discussing the readability of their written materials),[136] the MADRS formula appears to substantially outperform the other tested formulas in terms of predicting the scores that actual experts assigned to adjudicative decision texts.

Correlation strengths are not necessarily intuitive to understand, so it might be helpful to put MADRS' performance into context. The Flesch Reading Ease formula was derived in 1948.[137] It continues to be used as a readability performance measurement within Microsoft Word's

---

[135] See Cohen, *Statistical Power Analysis for the Behavioral Sciences*, *supra* note 39 at 78 (for an explanation of the relationship between correlation strengths, R^2 values, and the proportion of a dependent variable's variance that can be explained by the independent variable(s)).

[136] Social Secuʳity Tribunal of Canada, "An Evaluation of How Easy it is to Read Decisions of the Social Security Tribunal" (last modified 10 May 2022), online: <https://sst-tss.gc.ca/en/our-work-our-people/evaluation-easy-it-read-decisions-social-security-tribunal>; Shannon Salter, "RDO Speak to Shannon Shalter [SIC] About All Things ODR" (November 2019), online: <https://resolvedisputes.online/blog_shannon_salter.html> (containing a transcript of an interview with Shannon Salter, who was the Chair of the BC CRT at the time of the interview).

[137] Flesch, "A New Readability Yardstick", *supra* note 2.

"Editor" function, and Microsoft reports that over 1.2 billion people use its "Office" suite of software that includes Microsoft Word[138] – so the formula is widely accessible today. Additionally, a search of the Heinonline "Law Journal Library" database[139] for the phrase "Flesch Reading Ease" yielded 246 results published in different legal scholarship sources between 1975-2023, suggesting that the formula has been of enduring relevance even to legal researchers. The formula "has also been used in several U.S. state statutes aiming to ensure the readability of specific documents, such as instructions on income tax returns, financial institution forms, and insurance forms."[140] In other words, the Flesch Reading Ease formula has persisted, and remained influential, for over 75 years. This formula's regression correlation coefficient that was calculated at the time when the formula was first created was r = 0.7047.[141] MADRS' statistically significant regression correlation coefficient (see above at Table 3.1) is a notably stronger r = 0.771, and the MADRS formula has also demonstrated a statistically significant Pearson correlation coefficient, in testing on independent data, of r = 0.756. This means that MADRS's predicted and actual performance on legal texts is superior to the comparable predicted performance of the Flesch Reading Ease formula – a formula that has endured for three-quarters of a century.

With this now-validated formula, MADRS can be used with confidence to predict the types of measurements that were at the core of the formula's development. Specifically, this formula can now be used – drawing only on the linguistic information in an adjudicative decision text for each of the 6 contributing MADRS variables – to predict the score that my human experts would assign

---

[138] John Callaham, "There are now 1.2 billion Office Users and 60 Million Office 365 Commercial Customers" (November 29, 2018), Windows Central, online: <https://www.windowscentral.com/there-are-now-12-billion-office-users-60-million-office-365-commercial-customers>.
[139] This search was performed on February 23, 2024.
[140] Uri Benoliel & Shmuel I Becher, "The Duty to Read the Unreadable" (2019) 60:8 Boston College Law Review 2255 at 2276.
[141] Flesch, "A New Readability Yardstick", *supra* note 2.

to that text on a scale of 1 to 7 (where each score level corresponds to a particular education level for a hypothetical reader of the text).

3.5 Discussion: MADRS – A New Law-specific Readability Formula

Beyond the above statistical analyses that support the validity of MADRS, the formula is also sound from a theoretical perspective. In the sections of this dissertation that follow, I discuss each of the variables that contribute to the MADRS formula, as well as additional considerations that might help readers to understand both strengths and limitations of the formula.

Before moving forward with this discussion, however, it is worth recalling at this point that my "second generation" approach to developing the MADRS formula – using computational linguistics tools and statistical regression analysis – was not the only approach that could have been taken. As I noted above (in section 2.5), machine learning approaches are increasingly being used for predictive tasks, like the task of predicting how well readers will understand a text. It would likely have been possible, therefore, for a researcher with the correct skill set to have used my gold standard corpus of expert-rated decisions as a Large Language Model (LLM) for training an algorithmic or machine learning system. Using this approach, a computer may have been capable of replicating or predicting the scores that human experts would have assigned to new decision texts beyond the gold standard corpus – potentially with even greater reliability than the MADRS formula – without ever defining or identifying the variables that contribute to the system's decision-making processes.

Again, however, it must be recalled that such an approach would be almost completely opaque to anyone who tries to understand the reasons why the machine learning system has made the readability classifications that it has made. As an academic researcher, one would therefore be placed in the difficult position where it would be possible to say, "here is a new tool that works

very well at classifying the readability levels of adjudicative decision texts, but I cannot explain how it works, or why it works as well as it does." This position would be particularly difficult for legal scholars, whose training has conditioned them to rely on evidence and argumentation as the core bases for assessing the validity of conclusions. This position would also make it impossible to engage in the types of discussion that follow, wherein the theoretical arguments about why the MADRS formula works well under some conditions (but not necessarily under others) are considered.

3.5.1 Understanding the Theoretical Place of the Included Variables

As I explain in more detail below, the variables that contribute to the MADRS formula – and the direction in which those variables pull readability scores – accord with our theoretical understanding of how different linguistic properties render a text more or less readable.

3.5.1.1 Kuperman Age of Acquisition (Content Words) / AoA_CW

This variable measures the age at which individuals would have understood a particular 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.[142] 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).[143]

TAALES calculates Kuperman's age of acquisition values separately for "all words", "content words", and "function words". Function words are generally considered to be a closed

---

[142] Victor Kuperman, Hans Stadthagen-Gonzalez & Marc Brysbaert, "Age-of-Acquisition Ratings for 30,000 English Words" (2012) 44 Behavior Research Methods 978 at 980.
[143] Kuperman, Stadthagen-Gonzalez & Brysbaert, "Age-of-Acquisition Ratings for 30,000 English Words", *supra* note 142 at 980-81.

class of words consisting of "pronouns, articles, conjunctions, quantifiers, and prepositions" that fill grammatical functions, but that do not tend to carry significant lexical meaning.[144] Content words are considered to be an open or expanding class of more semantically rich words that "comprise nouns, verbs, adjectives, and some adverbs."[145] The MADRS formula relies upon age of acquisition scores for content words (*AoA_CW*). The value of this variable is calculated using the sum of all age of acquisition scores (for all content words in the text that have such scores) divided by the total number of content words in the text with age of acquisition scores.

The regression B-weight for the variable *AoA_CW* is 0.471. The positive value of this B-weight indicates the direction of the relationship between the variable and the MADRS value: higher age of acquisition scores will positively affect (or increase) MADRS values – indicating that a text is harder to read. This relationship makes intuitive sense: a text that is filled with content words that are not learned by readers until older ages will logically be harder to read for most people than a text that is filled with content words that are learned by readers at a younger age. Additionally, the same variable has been independently found (in another study that developed the Crowdsourced Algorithm of Reading Comprehension) to have a positive relationship with readability scores: texts that use words with higher ages of acquisition tended to be more difficult for readers to understand than texts that use words that are acquired at younger ages.[146]

3.5.1.2 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

---

[144] Sidney J Segalowitz & Korri C Lane, "Lexical Access of Function versus Content Words" (2000) 75 Brain & Language 376 at 376.
[145] Helen Bird, Sue Franklin & David Howard, "'Little Words' – Not Really: Function and Content Words in Normal and Aphasic Speech" (2002) 15 Journal of Neurolinguistics 209 at 210.
[146] Crossley, Skalicky & Dascalu, "Moving Beyond Classic Readability Formulas", *supra* note 2 at 553 (Table 2).

of a reader/listener: "some words (e.g., moon) may activate a very specific and vivid perceptual representation (e.g., a visual image of a full moon in the night's sky). […] Other words (e.g., serene) are considered to be abstract and not closely tied to such perceptual and/or action representations, thus not activating such rich perceptual or action representations."[147]

The *MRC_Imag_AW* variable from the MADRS formula draws on imageability scores for 9,240 different words[148] found in Coltheart's Medical Research Council Psycholinguistic database.[149] As Wilson explains, these imageability ratings were not the product of Coltheart's original work, but were an amalgamation of scores from three others studies that all assessed the imageability of different words.[150] Scores are intended to fall on a scale of between 100-700,[151] although the imageability scores for words in the database actually range between 129 ("plenipotentiary") and 667 ("beach"), with an average score of 450 and a standard deviation of 108.[152]

TAALES calculates MRC imageability values separately for "all words", "content words", and "function words". The MADRS formula relies upon imageability scores for all words (*MRC_Imag_AW*). The value of this variable is calculated using the sum of all imageability scores (for all words in the text that have such scores) divided by the total number of words in the text with imageability scores.

---

[147] 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 Memory 622 at 622 ["Reading Aloud Performance"].
[148] Michael Wilson, "MRC Psycholinguistic Database: Machine-Usable Dictionary, Version 2.00" (1988) 20:1 Behavior Research Methods, Instruments, & Computers 6 at 7 ["MRC Psycholinguistic Database"].
[149] Max Coltheart, "The MRC Psycholinguistic Database" (1981) 33A Quarterly Journal of Experimental Psychology 497.
[150] Wilson, "MRC Psycholinguistic Database", *supra* note 148 at 7.
[151] Wilson, "MRC Psycholinguistic Database", *supra* note 148 at 7. Wilson also explains that "The three sets of norms correlated highly and were merged by adjusting both the means and standard deviations before averaging."
[152] The University of Western Australia, School of Psychology. *MRC Machine Usable Dictionary. Version 2.00*, online: <https://websites.psychology.uwa.edu.au/school/MRCDatabase/mrc2.html>.

The regression B-weight for the variable *MRC_Imag_AW* is -0.023. The negative value of this B-weight indicates the direction of the relationship between the variable and the MADRS value: higher imageability scores will negatively affect (or decrease) MADRS values – indicating that a text is easier to read. This relationship is theoretically understandable. Research has shown that "imageability ratings strongly predict the memorability of items" in a way that is consistent with well-known dual coding theory,[153] according to which "one represents words verbally and as mental images."[154] Dual coding theory (DCT) can be distinguished from single code theories, which

> assume that knowledge in memory is basically abstract and amodal, existing in a disembodied state that is associated with no sensory modality. Such theories either do not deal with concreteness effects in reading comprehension (e.g., schema theory) or explain it in different terms (e.g., context availability theory). In contrast, concreteness is a key factor in the DCT account of reading comprehension […] In DCT, the two codes are theorized to be independent and additive in their effects, predicting that concrete language should be about twice as comprehensible and memorable as abstract language, other factors being equal.[155]

Because the human mind has access to two different codes (one semantic and one visual) in the case of more imageable or concrete words, these words are generally easier to process, and should make text more readable. Multiple studies suggest that this hypothesis is borne out in practice,[156] and confirm the theory-based place that the variable *MRC_Imag_AW* holds in the MADRS formula.

3.5.1.3 General Inquirer Database – Legal (Nouns) / GI_Leg_Nouns

---

[153] Khanna & Cortese, "Reading Aloud Performance", *supra* note 147 at 632.
[154] Khanna & Cortese, "Reading Aloud Performance", *supra* note 147 at 628.
[155] Mark Sadoski, Erin M McTigue & Allan Paivio, "A Dual Coding Theoretical Model of Decoding in Reading: Subsuming the Laberge and Samuels Model" (2012) 33:5 Reading Psychology 465 at 490 ["Dual Coding"].
[156] Sadoski, McTigue & Paivio, "Dual Coding", *supra* note 155 at 490-92.

This variable measures the presence of words that are found on a defined list of 192 words associated with legal, judicial, or police matters.[157] 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.[158] Examples of nouns contained within the General Inquirer's legal word list (*GI_Leg_Nouns*) include the following: advocate, auditor, deposition, fugitive, junta, legislator, etc.

SEANCE calculates values for the variable *GI_Leg_Nouns* by dividing the number of times that a noun in a text is contained within the General Inquirer "legal" word list, by the total number of words in the text – thereby calculating the percentage of total words in the text that are listed on the "legal" word list.

The regression B-weight for the variable *GI_Leg_Nouns* is 4.513. The positive value of this B-weight indicates the direction of the relationship between the variable and the MADRS value: where a text has a higher percentage of words that are legal in nature (i.e.: that appear on the "legal" word list), MADRS values will be positively affected – indicating that a text is harder to read. This relationship accords with previous research suggesting that legal language generally, and legal English in particular, is often unintelligible to most non-lawyers.[159] It is therefore not surprising to find that a greater density of legal terms within a text tends to render the text more difficult to understand.

3.5.1.4 COCA – Spoken (Proportion of Trigrams in the top 30K) / COCA_Spoken_TriTop30K

---

[157] Harvard University General Inquirer, "Descriptions of Inquirer Categories and Use of Inquirer Dictionaries", online: <https://inquirer.sites.fas.harvard.edu/homecat.htm>.
[158] Crossley, Kyle & McNamara, "SEANCE", *supra* note 87 at 806.
[159] Brenda Danet, "Language in the Legal Process" (1980) 14 Law & Society Review 445, at 464-469. See also, Robert P Charrow & Veda R Charrow, "Making Legal Language Understandable: A Psycholinguistic Study of Jury Instructions" (1979) 79 Columbia Law Review 1306 ["Making Legal Language Understandable"].

This variable measures how often trigrams (three-word phrases) within a text can be found within a list of the top 30,000 (i.e.: most-frequently used) trigrams contained inside of the Corpus of Contemporary American English – Spoken (COCA – Spoken) sub-corpus. Some further explanation about both trigrams, and the COCA corpus, are perhaps needed here in order to better understand how the variable *COCA_Spoken_TriTop30K* is computed.

Any text consisting of three or more words contains at least one trigram, and the number of trigrams in any text can be calculated by subtracting 2 from the total number of words in the text. For example, the 6-word phrase, "he loves to snowboard in winter", contains 4 unique trigrams.[160] Looking at language through the lens of trigrams allows scholars to consider not just individual words, but also the sequencing of words together within larger packages of language. This type of research has led to conclusions that individuals process frequently used multi-word phrases more easily than uncommon multi-word phrases.[161]

One major database or corpus that can be used for calculating word or N-gram 'range' and 'frequency' scores is the COCA corpus. This corpus was originally developed in 2010, and contained over 400,000 words at that time.[162] However, unlike other major corpora, the COCA corpus was designed to expand continuously, with new texts being added each year to potentially allow researchers to track changes in language use over time,[163] among other things. As of 2024, the corpus contains over "one billion words of text (25+ million words each year 1990-2019) from eight genres: spoken, fiction, popular magazines, newspapers, academic texts, TV and movies

[160] Specifically: he loves to / loves to snowboard / to snowboard in / snowboard in winter.
[161] See generally, Arnon & Snider, "More Than Words", *supra* note 71.
[162] Mark Davies, "The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English" (2010) 25 Literary & Linguistic Computing 447 ["COCA"].
[163] Mark Davies, "COCA", *supra* note 162 at 459-60.

subtitles, blogs, and other web pages."[164] In particular, the 'COCA – Spoken' sub-corpus contains over 127 million words from texts that consist of "[t]ranscripts of unscripted conversation from more than 150 different TV and radio programs (examples: All Things Considered (NPR), Newshour (PBS), Good Morning America (ABC), Today Show (NBC), 60 Minutes (CBS), Hannity and Colmes (Fox), Jerry Springer, etc)."[165] In the notes that explain how the COCA – Spoken sub-corpus was created, the creators argue persuasively that the texts in this sub-corpus faithfully represent actual and unscripted conversations that are different in kind from other, non-media varieties of Spoken American English.[166] In this sense, the COCA – Spoken sub-corpus represents a unique database of naturally-spoken English that has been reduced to writing and that is therefore capable of being studied through Natural Language Processing techniques.

With this context about both trigrams and the COCA – Spoken sub-corpus in mind, the variable *COCA_Spoken_TriTop30K* can now perhaps be more easily understood. TAALES computes scores for this variable by identifying every trigram within a text, and then by determining the number of trigrams within the text that also appear of a list of the 30,000 most common trigrams within the COCA – Spoken sub-corpus. This number is then divided by the total number of trigrams in the text to give a proportion score that signals the percentage of trigrams in the text that are on the top 30,000 list.

The regression B-weight for the variable *COCA_Spoken_TriTop30K* is -0.650. The negative value of this B-weight indicates the direction of the relationship between the variable and the MADRS value: higher trigram proportion scores will negatively affect (or decrease) MADRS

---

[164] Corpus of Contemporary American English Homepage, online: <https://www.english-corpora.org/coca/>.
[165] Corpus of Contemporary American English: Texts, online: <https://www.english-corpora.org/coca/help/texts.asp>.
[166] Corpus of Contemporary American English: Notes Explaining the Spoken Sub-Corpus, online: <https://www.english-corpora.org/coca/help/spoken.asp>.

values – indicating that a text is easier to read. This relationship is theoretically defensible for two reasons. First, as a general matter, more common and familiar multi-word phrases appear to be easier to understand than less common ones.[167] Second, 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:

> Frequency lists such as *The Teacher's Word Book* and the *Carroll–Davies–Richman* (Carroll et al., 1971) list used by early readability formulas were based on written corpora. Although these corpora were carefully constructed from materials that students read in their daily life, they failed to represent the also important, if not more important, source of spoken language that students are exposed to. The amount of exposure from spoken language is much greater than from written ones. Failure to include the spoken language does great harm to the representativeness of the frequency list and to its predictive power for text readability.[168]

It stands to reason, therefore, that the variable *COCA_Spoken_TriTop30K* would be inversely related to MADRS values, since a text that contains trigrams that are frequently found in spoken English would theoretically be easier to understand than a text that contains trigrams that are rarely or never used in spoken English.

3.5.1.5 Dependents per nominal (no pronouns, standard deviation) / Nominal_Deps_NN_StDev

This variable is a measure of the variation in syntactic complexity of language within a text. But, in order to understand what is being measured, some background on different aspects of the variable would likely be helpful.

The variable first considers nominals within a text. Nominals are essentially noun-phrases, that can consist of a single noun (e.g.: dog), or a noun together with other words (coming either before or after the noun) that modify the noun[169] (e.g.: the brown dog with long fur).

---

[167] See generally, Arnon & Snider, "More Than Words", *supra* note 71.
[168] Chen & Meurers, "Word Frequency and Readability", *supra* note 65 at 490.
[169] Seymour Chatman, "Pre-adjectivals in the English Nominal Phrase" (1960) 35:2 American Speech 83 at 83-84.

The variable then considers how many dependent clauses are associated with each nominal. From a grammatical perspective, a sentence must consist of an independent clause, and may also consist of one or more additional dependent or independent clauses.[170] The following three sentences contain different combinations of clauses:

1. There is a dog outside. [independent clause];
2. There is a dog outside *who is barking*. [independent clause and *dependent clause*]; and,
3. There is a dog outside, **and he is wet**. [independent clause and **independent clause**].

A sentence can contain any number of dependent clauses associated with the nominal (or noun-phrase) in the sentence.

TAASSC calculates the variable *Nominal_Deps_NN_StDev* by first computing the number of dependents for each nominal in a text. However,

> [n]oun phrases in English can consist of pronouns, and except in very rare cases, pronouns do not take direct dependents (relative clauses being an exception). Due to the potential for pronouns as phrases to skew counts of dependents, TAASSC includes two versions of each index, one that includes pronoun noun phrases in its counts, and one that does not.[171]

The variable *Nominal_Deps_NN_StDev* does not include consideration of pronoun-based nominals within its measurement. In order to compute the relevant value for this variable, TAASSC then determines the standard deviation for the values of dependents per nominal (excluding pronoun-based nominals) in the text.

The regression B-weight for the variable *Nominal_Deps_NN_StDev* is 1.802. The positive value of this B-weight indicates the direction of the relationship between the variable and the MADRS value: higher standard deviations in measurements of dependents per nominal will positively affect (or increase) MADRS values – indicating that a text is harder to read. This

---

[170] Kathleen Bardovi-Harlig, "A Second Look at T-Unit Analysis: Reconsidering the Sentence" (1992) 26:2 TESOL Quarterly 390.
[171] Kyle, *Measuring Syntactic Development in L2 Writing*, *supra* note 68 at 57.

relationship is theoretically unsurprising. Higher scores on the variable *Nominal_Deps_NN_StDev* means that there is a relatively higher variance of values for the number of dependents per nominal within a text (both above and below the average value). Since the range of possible values for the number of dependents per nominal is bounded by 0 at the low end, and is unbounded at the high end, 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.

Research has shown that individuals generally have a more difficult time in understanding what a sentence means when this sentence has a more complex structure (which would include sentences with relatively more dependent or subordinate clauses), in part because more working memory is needed in order to decode such sentences.[172] And, research in the legal domain, specifically, has shown a very strong negative correlation between the use of embedded or dependent clauses and text comprehension: with more embedded clauses, we see a drop in comprehension.[173]

3.5.1.6 Medical Research Council – Concreteness (Function Words) / MRC_Concr_FW

This variable measures the concreteness of words used in a text. Concreteness is similar to, but conceptually distinct from, imageability (discussed above). Concreteness refers to the extent to which a word represents something that can be experienced by the senses; thus, "[a]ny word that refers to objects, materials, or persons should receive a high concreteness"[174] score. Concreteness is generally contrasted with abstractness, where highly concrete and highly abstract words fall at opposite ends of the rating spectrum.[175]

---

[172] David Kaplan & Gloria S Waters, "Verbal Working Memory and Sentence Comprehension" (1999) 22 Behavioral & Brain Sciences 77 at 78-79.
[173] Charrow & Charrow, "Making Legal Language Understandable", *supra* note 159 at 1327-28.
[174] Khanna & Cortese, "Reading Aloud Performance", *supra* note 147 at 623-4.
[175] Khanna & Cortese, "Reading Aloud Performance", *supra* note 147 at 623-4.

The *MRC_Concr_FW* variable from the MADRS formula draws on concreteness scores for 8,228 different words[176] found in Coltheart's Medical Research Council Psycholinguistic database.[177] As Wilson explains, these concreteness ratings were not the product of Coltheart's original work, but were an amalgamation of scores from three others studies that all assessed the concreteness of different words.[178] Scores are intended to fall on a scale of between 100-700,[179] although concreteness scores for words in the database actually range between 158 ("as") and 670 ("milk"), with an average score of 438 and a standard deviation of 120.[180]

TAALES calculates MRC concreteness values separately for "all words", "content words", and "function words". The MADRS formula relies upon concreteness scores for function words (*MRC_Concr_FW*).[181] The value of this variable is calculated using the sum of all concreteness scores (for all function words in the text that have such scores) divided by the total number of function words in the text with concreteness scores.

The regression B-weight for the variable *MRC_Concr_FW* is -0.018. The negative value of this B-weight indicates the direction of the relationship between the variable and the MADRS value: higher concreteness scores will negatively affect (or decrease) MADRS values – indicating that a text is easier to read. This relationship is supportable in theory. "Words that are rated as highly concrete have been found to be processed more quickly in lexical decision."[182] The same

---

[176] Wilson, "MRC Psycholinguistic Database", *supra* note 148 at 7.
[177] Max Coltheart, "The MRC Psycholinguistic Database" (1981) 33A Quarterly Journal of Experimental Psychology 497.
[178] Wilson, "MRC Psycholinguistic Database", *supra* note 148 at 7at 7.
[179] Wilson, "MRC Psycholinguistic Database", *supra* note 148 at 77. Wilson also explains that "The three sets of norms correlated highly and were merged by adjusting both the means and standard deviations before averaging."
[180] The University of Western Australia, School of Psychology. *MRC Machine Usable Dictionary. Version 2.00*, online: <https://websites.psychology.uwa.edu.au/school/MRCDatabase/mrc2.html>.
[181] There are 190 function words (i.e.: words that are classified as 'other', as opposed to 'noun', 'verb' or 'adjective') with concreteness rating in the MRC Psycholinguistic Database.
[182] Khanna & Cortese, "Reading Aloud Performance", *supra* note 147 at 623-24.

variable has also independently been found (in another study that developed the Crowdsourced Algorithm of Reading Speed) to have a negative relationship with reading speed: texts that use more concrete words tended to be read more quickly than texts that use less concrete words.[183] And, as Bailin & Grafstein have noted, the use of words that reduce a person's reading speed are thought to also reduce comprehension speed and results, on different levels: "The use of vocabulary that is unfamiliar to a reader may not just slow down the reading process (basic fluency) or the speed of the comprehension process (comprehension fluency), but may also affect the reader's very ability to comprehend the text at all."[184]

3.5.2 MADRS is Length-Independent

The preceding discussion explained how each variable that contributes to the MADRS formula has a justifiable theoretical rationale for inclusion within the formula. From a processing perspective, however, there are other reasons why the formula should be theoretically expected to perform well.

To begin with, all of the contributing variables are proportion or ratio variables that measure distinct linguistic properties in ways that should not be affected by the length of a text (at least above a certain extreme minimum). For instance, the variable *GI_Leg_Nouns* measures the proportion of [words in a text that are found on the General Inquirer "Legal Nouns" word list] to [total words in the text] – it does not simply measure the aggregate number of words in a text that can be found on this General Inquirer word list. All of the other MADRS variables are structured similarly. This consideration makes the MADRS formula more robust when applied to texts of different lengths than formulas like the *Crowdsourced Algorithm for Reading Comprehension*

---

[183] Crossley, Skalicky & Dascalu, "Moving Beyond Classic Readability Formulas", *supra* note 2 at 553 (Table 3).
[184] Bailin & Grafstein, *Readability: Text and Context*, *supra* note 66 at 201.

(CAREC), which incorporated the variable "Lemma type count (content words)" into its readability calculations.[185] Unlike every other variable that contributes to the CAREC formula, this variable is a raw score variable that we would expect to manifest in increasingly higher numbers depending on the length of the text that is being studied (for most narrative or expository texts that are non-repetitive) – because longer non-repetitive texts will tend to use a greater variety of words than shorter texts. As a result, CAREC scores cannot be meaningfully compared across text samples of significantly different lengths, since the longer texts will almost inevitably score as less readable due to the inclusion of a raw score variable within the formula. It was for this reason[186] that authors who developed the CAREC formula subsequently developed a modified version of the formula (CAREC-M), which replaces the above raw score variable with a comparable ratio variable: "Number of content word lemma types divided by number of content words."[187]

The MADRS formula is not susceptible to skews like this based on text length because of its exclusive reliance on proportion or ratio component variables. The formula essentially relies on the "density" of different linguistic properties within a text, rather than on raw scores for each of these variables (that would almost inevitably increase for longer texts and decrease for shorter texts in ways that have no meaningful connection to the texts' difficulty levels).

3.5.3 MADRS is minimally vulnerable to NLP Parsing and Tagging Errors

---

[185] Crossley, Skalicky & Dascalu, "Moving Beyond Classic Readability Formulas", *supra* note 2 at 553 (Table 2, Row 11).

[186] See "ARTE: Automatic Readability Tool for English", online: <https://www.linguisticanalysistools.org/arte.html>, in the explanatory notes associated with the release of "Version 1.1", for the authors' explanation as to why they released a modified CAREC formula.

[187] "ARTE Index Description Sheet", online: <https://docs.google.com/spreadsheets/d/1O2SOCTcUYEpKzRGLC_iVRcrRwY4uefDbdpb0V-40rF0>, at Row 40.

The MADRS formula is also resilient, for the most part, against skews that might arise due to inaccurate tokenization, tagging, or parsing[188] of adjudicative decision text files by Natural Language Processing (NLP) software tools. Adjudicative decisions are written in a peculiar style. For instance, they often include inline citations to cases or secondary sources that can use multiple periods (e.g.: to signal where a court name or journal name has been abbreviated, or after the "v" that represents the word "versus" between two party names within a case name). These periods will generally be read as full stops – denoting the end of a sentence – by NLP tokenization tools, since the tools are generally not sophisticated enough to distinguish in all cases between instances where a period is used to represent a full stop, and where it is used to signal an abbreviation.[189]

As a result, a readability formula that relies on sentence length (or other sentence syntax variables) that depend on accurate assessments by NLP tools as to where sentences begin and end will tend to be somewhat inaccurate when used on adjudicative decision texts. The ultimate extent of these inaccuracies will depend on the weighting of the sentence-level variable within the overall formula, and the number or density of inaccuracies within the text. So, an adjudicative decision that 'misuses'[190] periods (e.g.: for abbreviations within citations) in **every** sentence would be processed less accurately by an NLP-dependent readability formula than an adjudicative decision

---

[188] "Tokenizing" is the process of dividing a text into its component words (word tokenization) and sentences (sentence tokenization) – essentially identifying the boundaries between words, and between sentences. "Tagging" is the process of assigning part-of-speech labels to particular words (e.g.: recognizing that "banana" is a noun, and "jumping" is a verb). "Parsing" is the process of identifying how words in a phrase or sentence relate to one another grammatically or syntactically. For a simple overview of these processes, see Dev Shah, "Breaking Down Natural Language Processing" (27 July 2022), online: <https://devshahs.medium.com/breaking-down-natural-language-processing-7bd8b8b02f95>.

[189] This is true of the TAASSC software tool that I have used for the purposes of computing a MADRS variable score, based on my experimentation with the software. By changing the content of a text file to add or delete periods that denote abbreviations (while leaving all other file content the same), for instance, I created files that scored slightly differently on many of the syntax measures computed by TAASSC.

[190] In this context, when I say 'misuses', I mean, 'uses for a purpose other than to denote a full stop'.

that ‘misuses’ periods only in every 20th sentence, which in turn would be processed less accurately than an adjudicative decision that only uses periods to denote full stops.

Other style properties within judicial decisions may create similar problems at the paragraph level. For instance, where a decision quotes from a statute in ways that include multiple line breaks (e.g.: quoting a statute that contains a list with each item separated by a line break), NLP tools will often read each line break as a new paragraph. However, this counting of paragraphs will be inaccurate in cases where the quoted statute (or other block quote) is embedded within a narrative paragraph, and should properly be treated for readability purposes as forming part of that paragraph. Any readability formula that depends on accurate assessments by NLP tools as to where paragraphs begin and end (for instance, to assess overlap of words or phrases across paragraphs as a measure of thematic cohesion) may therefore also be somewhat inaccurate when used on adjudicative decision texts.

To be clear, the MADRS formula does include one sentence-syntax variable that is vulnerable to the type of skewing that I have noted above (*Dependents per nominal (no pronouns, standard deviation)*). It does not include any paragraph-level variables. And the remaining 5 variables that contribute to the MADRS formula are not dependent on sentence- or paragraph-level segmentation of a text by NLP tools, because they relate to words or trigrams that are likely to be processed by NLP tools as accurately for adjudicative decisions as for any other texts.

Thus, while the MADRS formula is not wholly immune to inaccuracies such as those based on poor NLP tokenization, tagging, or parsing of a text, the formula is certainly less vulnerable to these skews when compared to, for instance, the two Flesch readability formulas that both rely on sentence length as one of the (only) two variables within the formulas. MADRS would similarly be less vulnerable than the Coh-Metrix L2 Reading Index formula, that includes “[p]roportion of

noun and pronoun lemma types that occur in the next two sentences for all sentences" as one of only three variables.[191] In both of these examples, the other formulas are much more heavily dependent on accurate NLP segmentation of the text into sentences than the MADRS formula, based on the ratio of sentence to non-sentence variables in the respective formulas.

However, it is important to understand the limitations of the MADRS formula. The formula will perform most accurately when NLP tools have tokenized, tagged, and parsed an adjudicative decision in ways that best reflect how a reader would perceive the decision. Most readers will logically recognize the difference between when a period is used to denote an abbreviation, and when it is used to denote a full stop. If NLP tools do not recognize periods in the same way, then the MADRS formula will provide a less accurate prediction of how likely the text is to be understood by readers. This practical and processing-based limitation of the MADRS formula offers one explanation as to why there is a less-than-perfect (but still very strong) correlation between MADRS scores and actual expert-assigned scores.

3.5.4 The Perhaps Surprising Exclusion of Certain Variables from the MADRS Formula

As explained above, the MADRS formula includes variables relating to lexical (word) difficulty, trigram frequency, sentiment and cognition, and syntax. But the formula does not include any variables relating to cohesion, and it is not necessarily obvious from a theoretical perspective why certain variables were included within the formula instead of others that might make more sense. As I explain below, however, the statistical integrity of the MADRS formula is sound. And the soundness of the formula would be weakened if linguistic variables were added to

[191] The variables that contribute to this formula, as calculated by the ARTE software tool, are explained within the "ARTE Index Description Sheet", online: <https://docs.google.com/spreadsheets/d/1O2SOCTcUYEpKzRGLC_iVRcrRwY4uefDbdpb0V-40rF0>, at Rows 54-57.

or subtracted from the formula based on theoretical hunches that statistical analyses do not substantiate. Furthermore, upon closer examination, there are theoretical explanations available to us as to why certain variables were not included within the formula.

3.5.4.1 Cohesion Variables do not Feature Within MADRS

There are good reasons for us to believe that cohesion has an impact on readability. I described these reasons above (within section 3.4.2.4 of this dissertation) when I explained why I chose to measure cohesion-related properties of the 250 files that were used to create the MADRS formula. It is therefore somewhat surprising that MADRS does not ultimately incorporate any cohesion measures within the formula.

One explanation for the absence of cohesion variables within MADRS is grounded in statistics. On initial review of the correlations between all 1,115 linguistic variables for which I computed scores and the average score assigned by human experts, I found that two cohesion-related variables had statistically significant correlation strengths above the r = 0.325 cutoff that I had applied: (1) pronoun density (r = -0.385); and, (2) pronoun-noun ratio (r = -0.402). However, these two cohesion variables showed strong multicollinearity (r = 0.973), which suggests that they were likely both measuring the same or closely related properties within the text. I therefore only included the second variable within my hierarchical regression modeling calculations, since the second variable was more strongly correlated with my average human expert scores.

During the hierarchical regression modeling calculations, however, this cohesion variable was not found to add to the model in a statistically significant way. The variable was therefore eliminated from the ultimate model that now describes the MADRS formula – as were 14 other variables that were eliminated for the same reason. In other words, the statistical modeling process that was used to create the MADRS formula (and that is the generally accepted process for creating

linguistic readability formulas over the last two centuries)[192] caused the cohesion variable to be excluded from the MADRS formula not because the variable is unrelated to readability, but because there were simply other variables that contributed to the model in more powerful and statistically significant ways. I make this point because it remains likely that cohesion variables play an important role in determining how well a person will understand a text – even if there are no such variables within the MADRS formula for reasons that relate to the statistical development of the formula.

But I also hypothesize that there is a more fundamental reason why cohesion variables do not feature within the MADRS formula. It must be recalled that the text samples that were used to build the MADRS formula were all only (approximately) 600 words in length. These samples would be considered relatively short within the broader world of adjudicative decisions. They are perhaps so short that readers might not depend heavily on internal cohesion features within the text samples to help the readers understand the texts. One does not necessarily need linked repetition of key words and phrases across sentences and paragraphs, or multiple time-sequencing words, to help one understand such a short text, since the ideas from the beginning of the text can likely still be retained within a reader's working memory as the reader finishes the text. The same cannot necessarily be said for much longer texts: the ideas from the beginning of a 10,000-word text would likely be forgotten much more easily (without the help of cohesion-based writing tactics) in the mind of a reader who is reaching the end of that document – particularly if the document was not read all in one sitting. So, it is possible that the presence or absence of cohesive

[192] Cunningham, Hiebert & Mesmer, "Investigating the Validity of Two Widely Used Quantitative Text Tools", *supra* note 8 at 814.

features within the 250 files that were used to build the MADRS formula did not have a major effect on the way in which my expert raters scored these files.

If this hypothesis is accurate, then the MADRS formula likely under-represents the importance of cohesion in helping readers to understand longer texts. Future research might explore this hypothesis by following a method similar to the method that was used in this chapter to derive a new law-specific readability formula – but one that asks expert raters to score the readability of much longer text samples. If, for instance, the average adjudicative decision that one wishes to assess for readability is approximately 3000 words in length, then a method that relies upon 3000-word samples that have been human-scored to build the formula might produce a more robust formula than MADRS, and it might include one or more cohesion variables.

There are, however, reasons to question the validity of my hypothesis. As an initial observation, I note that the CAREC and CAREC-M readability formulas both use 5 different cohesion-related variables[193] within the formulas' sets of 13 contributing variables. And the text samples that were scored by humans in order to form the basis of these readability formulas were roughly 200-word Wikipedia articles[194] – so the relatively short text lengths used for those recent formulas did not preclude cohesion variables from being incorporated into the formulas (which were developed through regression analysis using the same statistical techniques that I have used in developing MADRS).[195] This example suggests that a factor other than text length (such as, perhaps, the law-specific nature and topics of the adjudicative decision texts as compared to the Wikipedia texts) must explain the exclusion of cohesion variables from MADRS.

---

[193] These variables are all listed (at Rows 22-26 and 36-40, respectively) in "ARTE Index Description Sheet", online: <https://docs.google.com/spreadsheets/d/1O2SOCTcUYEpKzRGLC_iVRcrRwY4uefDbdpb0V-40rF0>.
[194] Crossley, Skalicky & Dascalu, "Moving Beyond Classic Readability Formulas", *supra* note 2 at 546-547.
[195] Crossley, Skalicky & Dascalu, "Moving Beyond Classic Readability Formulas", *supra* note 2 at 551-552.

However, other research into cohesion has also shown that cohesion sometimes helps, and sometimes hinders, a reader's understanding of a text.[196] The difference appears to depend in part on the reader's prior knowledge: "Low-knowledge readers gain greatly from added cohesion whereas more knowledgeable readers (but not necessarily experts) can gain from lower cohesion. This phenomenon has been referred to as the *reverse cohesion effect*."[197] While the reverse cohesion effect may seem counter-intuitive, it can be theoretically explained in the following way: "the high-knowledge readers […] were able to gain from low-cohesion text because it forced them to generate inferences, and that inferencing resulted in a better, or deeper, understanding of the text."[198] So, while it is possible that MADRS fails to accurately account for cohesion as a readability variable in adjudicative decisions (perhaps because of the relatively short text extracts that were used to develop MADRS), it is also possible that cohesion does not have a predictable or uniform effect on reading comprehension – so the exclusion of cohesion variables is prudent and theoretically sound.

In any case, no readability scholar would suggest that cohesion is the only, or the determinative variable that affects readability. It is now widely accepted that readability is a function of many different types of language variables. The MADRS formula relies on several different types of linguistic variables relating to syntax, lexical, N-gram, and sentiment/cognition properties. It was developed in conformity with well-entrenched methods for creating readability formulas. And it has shown itself to be the strongest predictor of the readability scores that human experts assigned to independent adjudicative text files, when compared to 6 other general-purpose readability formulas. These considerations suggest that one can apply the MADRS formula with

---

[196] McNamara, et al, *Automated Evaluation of Text and Discourse with Coh-Metrix*, *supra* note 100 at 27-36.
[197] McNamara, et al, *Automated Evaluation of Text and Discourse with Coh-Metrix*, *supra* note 100 at 27.
[198] McNamara, et al, *Automated Evaluation of Text and Discourse with Coh-Metrix*, *supra* note 100 at 32.

confidence – but a healthy awareness of the fact that the formula does not present a complete picture of all the different factors that influence reading comprehension – because of the formula's wide base in different language variables that are theoretically linked to reading comprehension.

3.5.4.2 Considering Other Excluded Variables

Apart from cohesion, we might intuitively question why certain variables were included or excluded from the final MADRS formula. For instance, it is relatively easy to see why a variable that measures the concreteness of words would feature within a readability formula based on my explanation of the theory behind concrete language, above. But if the variable is influential, why does the MADRS formula only measure its impact in relation to function words (*Medical Research Council – Concreteness (Function Words) / MRC_Concr_FW*), as opposed to content words, or all words? Recalling that function words are "pronouns, articles, conjunctions, quantifiers, and prepositions" that fill grammatical functions, but do not tend to carry significant lexical meaning,[199] it is perhaps surprising that the concreteness of this class of words had a sufficient connection to average expert scores so as to be included within the MADRS formula.

Again, the explanation for inclusion of variables like this is grounded in the accepted statistical methods that were used to derive the formula, but can also draw upon other relevant considerations. The simple answer as to why this variable was included comes from my statistical method: the correlations between average expert score (on the one hand) and MRC concreteness scores for all words / content words / function words (on the other hand) were r = -0.502; r = -0.202; and, r = -0.560, respectively. So, the concreteness scores for content words were below the r = 0.325 cutoff for further analysis. And, the concreteness variable for function words showed

[199] Sidney J Segalowitz & Korri C Lane, "Lexical Access of Function versus Content Words" (2000) 75 Brain & Language 376 at 376.

strong multicollinearity (r = 0.931) with the variable *Medical Research Council – Imageability (All Words) / MRC_Imag_AW*, which again suggests that these two variables were measuring the same or similar language effects on readability. The latter variable was more strongly correlated with average expert score, so only the latter variable was included within my hierarchical regression analysis (and ultimately included within the MADRS formula because it contributed in a significant way to the regression model).

This explanation might seem persuasive to a statistician, but people who study language are probably left unsatisfied by the explanation. Looking at the concreteness scores for actual function words might help to better illustrate why the *Medical Research Council – Concreteness (Function Words) / MRC_Concr_FW* variable was included in MADRS. Consider the following table of the 10 highest- and lowest-scoring function words in the MRC's concreteness database:[200]

Table 3.3 – Highest and Lowest Scoring Function Words – Concreteness

Word	**Concreteness Score (higher is more concrete)**
as	158
of	180
however	186
than	188
then	190
also	194
how	195
so	195
because	196
from	197
[…]	[…]
him	368
you	370
one	379

[200] All concreteness scores in this table are from the Medical Research Council's Concreteness ratings within its Psycholinguistic Database. Online: <https://websites.psychology.uwa.edu.au/school/MRCDatabase/uwa_mrc.htm>.

us	397
he	399
she	406
her	419
mine	452
I	483
me	511

Note how the most concrete function words tend to be personal pronouns, while the least concrete function words tend to be prepositions and conjunctions.

This observation, together with my earlier explanation about how two different cohesion variables related to pronoun density and ratios within a text both correlated strongly with average expert score, perhaps offer us a new explanation for why the variable *MRC_Concr_FW* is included within MADRS. Perhaps the presence of personal pronouns within a text at high densities really does have a positive effect on readability (and a negative effect on MADRS values). The high density of these pronouns may add cohesion to a text that makes it easier to understand (perhaps by signifying very clearly who is performing actions within a text in ways that readers can easily follow). And, perhaps the variable *MRC_Concr_FW* (in light of what we can see in the above table) is just a better predictor of the effect that a high density of personal pronouns will have on a reader's understanding of a text than other related measures of how pronouns are used within a text.

In reality, no one has the ability to see behind somewhat opaque statistical regression analysis techniques in order to understand why variable X contributes significantly to a regression model but variable Y does not, when these two variables might well be measuring very similar or related properties of a text. In this sense, regression analysis represents a somewhat unfulfilling (at least intellectually, or theoretically) tool for devising readability formulas, because the statistical technique stands on its own, and is justified only on its own terms. What I mean is that I can

conceive of persuasive theoretical justifications as to why certain variables were included or excluded, but these justifications never actually flow directly from the statistical technique. Regression does not offer its own justifications from beyond the realm of statistics as to why one variable was chosen and another was omitted.

3.6 Methods Revisited: the Interplay Between Statistics and Theory

In light of what has just been said, it would be natural to wonder whether this statistical approach to creating a readability formula is putting a form of blind trust in the method, since the specific and individual variables that contribute to a formula are ultimately selected for statistical reasons rather than theoretical ones, and are only considered (or reconsidered, as I explain below) from a theoretical perspective after they have already been selected through regression analysis. There is no correct answer to this subjective question, and different people will have different levels of trust in the statistical methods that were used in creating MADRS and other similar readability formulas over the years. However, by reflecting below in more detail on the methodological choices that were made in this dissertation, my hope is that this dissertation will instill an informed level of trust in the methods, rather than skepticism or a worrying form of blind trust.

To begin with, it must be noted that the MADRS variables were not selected for inclusion in the formula based only on statistical reasons. Each broad class of variables (i.e.: lexical sophistication, syntactic sophistication and complexity, sentiment and cognition properties, and cohesion) was measured and assessed for correlations with average expert scores within this chapter because there were sound theoretical reasons for thinking that variables within each group

are related to ultimate comprehensibility for any given text.[201] So, while it is true that the individual linguistic variables that were drawn from the above broad groups and were included within MADRS were 'selected' for purely statistical reasons (specifically, because these variables were strongly correlated with human-assigned expert scores; were free of multicollinearity concerns in relation to one another; and, contributed in statistically significant ways to the MADRS regression model), I had made a prior and informed theoretical decision to assess these variables in the first place because of their likely connection to the comprehensibility of a text. To express this same idea in simpler terms, theory initially directed me to study a group of variables that were thought to be related to comprehension; statistical analysis simply told me which of these variables was most strongly connected with comprehension, and which weighted combination of them produced the best predictive equation for determining the readability scores of other adjudicative texts. In this sense, the role of statistical techniques – which I have described in substantial detail in this chapter because these techniques are unfamiliar to many people whose work is focused on law – to derive the MADRS formula is not disproportionate.

Furthermore, despite the fact that I have offered theoretically sound (albeit *post facto*) reasons that are capable of supporting the inclusion of each MADRS variable within the formula, the provision of such reasons should not be viewed as a precondition that must necessarily be satisfied before one can accept the validity of the MADRS formula. To be sure, one would probably need to explain or hypothesize why variable X or Y is included within a readability formula in a particular (e.g.: positive or negative) way, when theory suggests that the variable should have the opposite, or no, effect on the comprehensibility of a text. But the fact that a variable within a

[201] These variable groups, and the theoretical reasons for studying the different groups within the formula-building process, are discussed above at sections 3.4.2.1 to 3.4.2.4.

regression equation operates in ways that contradict an element of theory is not grounds for determining that the variable has been inappropriately included within the formula, or that the formula is invalid. To the contrary, the statistical analysis could just as easily help to demonstrate – in an empirical and quantitative way – how a theory might be misplaced. Flesch, for instance, theorized that longer words were more difficult to understand than shorter words.[202] However, the statistical techniques used by Dale & Chall,[203] and more recently by Chen & Meurers,[204] both suggest that Flesch's theory was flawed: comprehension is affected far more by a word's frequency measures (that is, how commonly the word is used in natural language) than it is by simple word length. As this example shows, theory can inform research design, but the statistical outputs from empirical studies can also inform, update, and refute past theories.

As I hope the above discussion has shown, there are sound statistical justifications (related to correlation strengths, multicollinearity, and significance of contributions to hierarchical regression models) for the inclusion and exclusion of variables from the MADRS formula. And, there are *post facto* theoretical hypotheses and/or justifications that are at least available to us to explain the same inclusions and exclusions of different variables. While this leaves us in a position where we cannot necessarily know with certainty why MADRS outperforms other tested formulas in predicting the average readability score that human experts would assign to a text, it does leave us with a new law-specific readability formula that is capable of predicting 57.2% of the variance in human-assessed readability scores[205] – which is itself a useful outcome for anyone who strives to understand the readability levels of Canadian adjudicative decisions.

---

[202] Flesch, "A New Readability Yardstick", *supra* note 2 at 223.
[203] Jeanne S Chall & Edgar Dale. *Readability Revisited: The New Dale-Chall Readability Formula*. (Cambridge, MA: Brookline Books, 1995).
[204] Chen & Meurers, "Word Frequency and Readability", *supra* note 65.
[205] See above, at section 3.4.4.

3.7 Conclusion – MADRS: Law's Comprehensive Readability Formula

Within this chapter, I have identified the most appropriate tool for assessing the readability level of Canadian adjudicative decisions: *Madden's Adjudicative Decision Readability Score*, or MADRS. This new readability formula was created from a reference corpus of law-specific texts that are similar in quality and kind to the types of texts that the formula is primarily intended to score. The formula was designed for its ability to predict the "gold standard" scores that human experts assigned to texts, where these expert scores had been independently validated through their strong correlation with scores that grade 12 students who are not lawyers assigned to the same texts.

The MADRS formula draws on different variables that quantify linguistic properties related to sentiment and cognition (*General Inquirer Database – Legal (Nouns)*), lexical sophistication (*Kuperman Age of Acquisition (Content Words*); *Medical Research Council – Imageability (All Words)* and *Concreteness (Function Words)*), N-gram sophistication (*COCA – Spoken (Proportion of Trigrams in the top 30K)*) and syntactic complexity (*Dependents per nominal (no pronouns, standard deviation)*). And, the variables that influence the formula all have solid theoretical justifications for inclusion, based on research about the effects that these types of linguistic properties can have on the processing of language and on reading comprehension. Thus, while this new formula, like all quantitative readability formulas that have been created and used over more than the last 100 years, focuses exclusively on the quantifiable properties of a text, it does so in a way that accords with our theoretical understandings about how readers cognitively

process texts. The formula is therefore far more robust to the types of criticisms that Bailin & Grafstein have leveled at other formulas in the past.[206]

From an effectiveness perspective, the MADRS formula outperforms the major existing general-purpose readability formulas that were tested on adjudicative decision texts, as shown by its much stronger correlation to the "gold standard" expert scores within my formula validation and testing subset of 125 files. The correlation statistics of different readability formulas (in particular, the lower correlation strengths of general-purpose readability formulas on law-specific texts) that were reported above in Table 3.2 perhaps provide further evidence of the subject/domain limitations of any readability formula, and should serve as a caution to anyone who hopes to use the MADRS formula on texts that are substantially different from adjudicative decision texts – such as texts from fields outside of law. By highlighting how relatively less effective other general-purpose readability formulas are on law-specific texts, this chapter reinforces the idea that a readability formula's effectiveness will be less than optimal as one applies it further and further from the context within which it was derived.

Having accomplished this prerequisite task of identifying a suitable tool for measuring the readability of Canadian adjudicative decisions from among those that currently exist (or that were created within this dissertation), it is possible to proceed onward to the next task of calculating and analyzing readability scores for Canadian adjudicative decisions. In the following chapter, I apply the MADRS formula to multiple different corpora of Canadian court and tribunal decisions from three distinct contexts in order to illustrate the readability levels of these decisions.

---

[206] Bailin & Grafstein, *Readability: Text and Context*, *supra* note 66 at 15-52.

4. General Canadian Decision Readability Studies

4.1 Overview

In the previous chapter, I created *Madden's Adjudicative Decision Readability Score* (MADRS), and assessed that it outperformed other well-known general-purpose readability formulas in terms of predicting human experts' readability scores for adjudicative decision texts. Where the MADRS formula appears to offer the best predictive approximation for human experts' scores when applied to this type of text, I have elected to use the MADRS formula within this chapter and the ensuing chapters in order to compute readability scores for large sets of adjudicative decisions. The resulting MADRS values allow me to start answering elements of my research questions that have not been answered in the previous chapters. Specifically, in this chapter, I focus on answering my subsidiary research question #2 ("how readable are court and administrative tribunal decisions, generally?") by using the MADRS formula to calculate and analyze the readability levels of hundreds of adjudicative decisions that were released by Canadian courts and tribunals in 2022.

More specifically, this chapter contains three discrete studies that attempt to illustrate the general readability levels of Canadian court and tribunal decisions in useful ways. The first study looks at differences between readability scores across different court levels (apex, appeal, superior, and inferior courts). The second study looks at readability differences across different legal subject areas (family, criminal, and tax law). Finally, the third study examines the differences between readability scores for courts and tribunals, respectively, when these two entities are dealing with cases involving the identical legal subject matter (in the contexts of human rights and social security law decisions). Collectively, these analyses provide the data and insights that are needed to make a broad assessment of the readability levels of Canadian court and tribunal decisions.

4.2 Readability Scores Across Different Court Levels

There are compelling reasons why, on the one hand, it might be desirable for court decisions to be more readable when they come from courts that are situated higher within the judicial hierarchy. Decisions of the Supreme Court of Canada (SCC), for instance, deal with cases that emerge from all parts of the country (that is, they have wide geographical scope),[1] and they are binding on all other Canadian courts as a matter of *stare decisis* (that is, they have the broadest and strongest precedential force of any Canadian court decisions).[2] Furthermore, because the court generally grants leave to appeal only in cases that raise questions of public importance,[3] Canada's apex court tends to deal with cases that should theoretically be of greater interest to the Canadian population than other courts. In this sense, the world of potential audience members for SCC decisions is likely to be larger than for a lower court's decisions, so the court may need to express itself in a more readable way if it hopes to reach many of these audience members.

On the other hand, there are also compelling reasons why courts at lower levels in Canada's judicial hierarchy may want to produce more readable decisions. These courts dispense more immediate justice (since it takes time for cases to make their way up to appeal courts and the SCC after a trial court decision is rendered) and may therefore capture more public attention. The evidence also shows that self-represented litigants (SRLs) appear far more often at trial courts than

---

[1] *Supreme Court Act*, RSC 1985, c S-26, s 35: "The Court shall have and exercise an appellate, civil and criminal jurisdiction within and throughout Canada."
[2] *R v Sullivan*, 2022 SCC 19, at para 59.
[3] *Supreme Court Act*, RSC 1985, c S-26, s 40(1): the court can only grant leave to appeal (in cases where an appeal does not exist as of right) where "the Supreme Court is of the opinion that any question involved therein is, by reason of its public importance or the importance of any issue of law or any issue of mixed law and fact involved in that question, one that ought to be decided by the Supreme Court or is, for any other reason, of such a nature or significance as to warrant decision by it."

at appeal courts and the SCC.[4] Because of this presence of SRLs, trial court judges may have a stronger obligation to communicate their decisions readably so that the parties can understand the decisions, as compared to higher court judges who preside in cases where the parties are more often represented by lawyers.

In short, various competing hypotheses about where decisions might be expected to be more readable, and various normative arguments about where decisions should be more readable, can all be advanced and defended in different ways. In reality, however, we have no current (or even historical) data about decision readability levels for trial, appeal, and apex courts in Canada. This information would go a long way toward answering one of my key research questions about general readability levels for Canadian courts and tribunals, so the analysis that follows attempts to provide and explain the data on this topic.

4.2.1 Method – Assessing Readability Scores Across Court Levels

In order to assess readability levels of Canadian court decisions from different levels of courts, I first compiled separate corpora of decisions for each court level. That is, I generated a database of text files that comprised either the entire population of decisions being studied (in the case of the SCC), or an appropriate number of randomly identified decisions from the relevant court level (for appeal courts, superior courts, and inferior courts), so that I could measure the readability levels of decisions from each level of court.

Since there was not a high number of SCC decisions in 2022, I used the full population of cases within my corpus, subject to the exclusion criteria described below. For all other levels of court, it was not necessary for me to study the whole population of decisions, since the population

---

[4] Julie Macfarlane & Charlotte Sullivan, "Tracking the Trends of the Self-Represented Litigant Phenomenon: Data from the National Self-Represented Litigants Project, 2019-2021" (2021), at 12, online: <https://scholar.uwindsor.ca/lawnsrlppubs/12>.

was large enough that I could reliably study a random sample of the population (of a particular size), with the statistical expectation that this sample would be representative of the larger population. This technique of sampling from the population significantly reduced the time and processing effort required to collect the necessary data – as will be seen below in my description of the steps involved in collecting and computing readability scores for decisions.[5] And, for most purposes, it is generally not necessary to have data about each individual element of a population in order to make use of the data because, if one uses a random sampling method, it is possible to statistically estimate the relevant characteristics of the population as a whole: "[s]ampling allows the estimation of the characteristics of a population by directly observing a portion of the entire population."[6] In other words, one can statistically infer characteristics of the entire population based on characteristics of the random sample.

In determining the appropriate sample size for each level of court other than the SCC, I used a 95% confidence interval, and a 7% margin of error. Using these two parameters, and the number of total cases within the population of 2022 decisions from each level of court,[7] I then calculated the required sample size using a free online statistical calculator.[8] My random sample sizes were N=183 (appeal court); N=193 (superior court); and, N=174 (inferior court). Based on these parameters and sample sizes, it is statistically expected that, 19 times out of 20, the true average readability score for the entire population of cases from which my random samples are

---

[5] This resource constraint in relation to studying an entire population is well-recognized in quantitative research methodological literature. See, e.g., Patrick Dattalo, *Determining Sample Size: Balancing Power, Precision, and Practicality* (New York: Oxford UP, 2008) at 8 [*Determining Sample Size*]: "the cost of studying an entire population usually is prohibitive to both researchers and those being studied in terms of privacy, time, and money. Consequently, a subset or sample of a given population must be selected."

[6] Statistics Canada. *Statistics: Power from Data!* (Government of Canada: Ottawa, September 2, 2021), online: https://www150.statcan.gc.ca/n1/edu/power-pouvoir/toc-tdm/5214718-eng.htm, at section 3.2.1.

[7] The total populations consisted of 2709 appeal court, 9665 superior court, and 1490 inferior court decisions, from which my smaller random samples were respectively drawn.

[8] Online: <https://www.qualtrics.com/blog/calculating-sample-size/>.

drawn will fall within +/- 7% of the sample's average readability score. In the case of the SCC corpus, since I used all opinions within the population, no statistical inference was required: the average readability score of the opinions that were studied is the average readability score of the entire population of 2022 decisions that met the inclusion criteria.

4.2.1.1 Case Selection

I applied the following selection approach when building each corpus of decision files. Except in the case of SCC decisions (for my Apex Court corpus), my searches for cases and my extraction of text from cases used the CanLII database (www.canlii.org), because this database is open and free for all users (unlike other subscription-based legal databases, like the Westlaw and Lexis Advance Quicklaw databases). I therefore assessed that the CanLII database was probably the most widely used database by the full spectrum of audiences for court and tribunal decisions, so I relied on this database for my purposes.

I used the SCC's LexUM website (https://scc-csc.lexum.com/scc-csc/en/nav.do) to find and extract text from SCC cases, since this website was the first result in my Google search for "Supreme Court of Canada decisions." I therefore assessed that this website was probably the most widely used site by the full spectrum of audiences for SCC decisions, so I relied on this database for my purposes. Although I have explained my choices of databases here, I have no reason to believe that the results of my analysis would have been different if different databases had been used – as will become clearer momentarily.

For appeal, superior, and inferior court searches within CanLII, I began with a full text search of the database for documents containing the word "a" – a word that I expect would appear in all cases. I then filtered the search period to only results from 2022. I then filtered to only "cases" (excluding secondary sources and legislation). For appeal courts, I used CanLII's filter "all appeal

courts" to narrow the search. For both inferior[9] and superior courts,[10] I used the manual filter "select specific courts" to include the relevant courts.

For all corpora, decisions from Quebec courts were excluded. These decisions were overwhelmingly written in French, and my readability formula is based on the English language. The lexical, N-gram, and sentiment/cognition variables that contribute to the formula all rely on measurements about uniquely English words or phrases. The C*OCA_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 3-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 problem exists for other variables: the concreteness and imageability measures that psycholinguists developed for English words (and that MADRS relies upon) have not been created in an identical manner from French words, and even if they did exist, it is not obvious that imageability or concreteness variables would quantitatively have the same effect on comprehension in French as they do in English. It is perhaps because of these language-specific aspects of readability formulas that only a few of the world's 6000+ languages have associated

[9] For "inferior courts", I included the following trial courts with non-federally-appointed judges: Provincial Court of British Columbia; Alberta Court of Justice; Provincial Court of Saskatchewan; Provincial Court of Manitoba; Ontario Court of Justice; Provincial Court (New Brunswick); Provincial Court of Nova Scotia; Provincial Court of Newfoundland and Labrador; Territorial Court of Yukon; Territorial Court of the Northwest Territories.

[10] For "superior courts", I included the following trial courts with federally-appointed judges: Federal Court of Canada; Tax Court of Canada; Supreme Court of British Columbia; Court of King's Bench of Alberta; Court of King's Bench for Saskatchewan; Superior Court of Justice (Ontario); Court of King's Bench of New Brunswick; Supreme Court of Nova Scotia; Supreme Court of Prince Edward Island; Supreme Court of Newfoundland and Labrador; Supreme Court of Yukon; Supreme Court of the Northwest Territories; Nunavut Court of Justice.

readability formulas.[11] Likely for the same reason, scholars have concluded that "[u]nfortunately, existing approaches cannot be easily extended to handle thousands of different languages."[12] 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 dissertation 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). I was concerned that translated decisions might not express the authoring judge's intentions as accurately as decisions that were written first in English, so I excluded these decisions from my corpora where they were identified.

Each text file contains a single judicial opinion, not necessarily the entire decision. Where a decision other than a SCC decision contained more than one opinion, only the majority or controlling opinion was included. All SCC opinions meeting the criteria were included, since even dissenting and concurring opinions from that court tend to be influential, or to capture more widespread interest.[13] This decision to include dissenting and concurring opinions from the SCC but not from other appeal courts could potentially skew cross-court readability comparisons if

---

[11] Jesse Saba Kirchner, Justin Nuger & Yi Zhang, "An Extensible Crosslinguistic Readability Framework" in *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora* (Singapore: ACL Anthology, 2009) 11 at 11.
[12] Kirchner, Nuger & i Zhang, "An Extensible Crosslinguistic Readability Framework", *supra* note 11 at 11.
[13] *See generally*, Claire L'Heureux-Dube, "The Dissenting Opinion: Voice of the Future" (2000) 38:3 Osgoode Hall Law Journal 495 (discussing the importance of dissenting opinions, and the impact that many such opinions have had in Canada and the world). *See also*, Christina M Frohock, "Schrödinger's Dissent: The Hybrid Authority of a Dissenting Opinion" (2024) 107:4 Marquette Law Review 963.

there are appreciable readability score differences relating to opinion types. I return to this point below, at section 4.2.2.2, when I discuss readability scores across different court levels.

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

Random selection of cases (for appeal, superior, and inferior courts) was achieved in the following manner. First, a full list of cases within the population was generated using CanLII (which reports results in screens of 25 cases at a time, up to a maximum of 500 results). For each corpus, it was therefore necessary to subdivide the population chronologically, and to then generate 'partial' results lists for each smaller sub-period within the year 2022. This approach ensured that the searches would yield less than 500 cases for each sub-period, until a complete list of cases for the entire year was extracted. For instance, the inferior court population was broken down into 4 sub-periods from 2022 (Jan-Mar; Apr-Jun; Jul-Sep; and, Oct-Dec), so that a results list of approximately 375 cases was produced for each sub-period (i.e.: results lists of less than the CanLII maximum of 500 results).

The full list of cases from 2022 was then transferred into a Microsoft Excel spreadsheet, sorted alphabetically by case name, and each case was assigned a number from 1 – [N]. Using a free online research randomizer tool,[14] I then generated a random list of numbers equal to my required sample size for that level of court, such that each case had an equal chance of being selected for inclusion. I then highlighted the randomly identified cases (by number) within my list

[14] Online: <https://www.randomizer.org/>.

of all cases, and began the process (described in more detail below) of downloading these cases for analysis.

In any situation where a randomly selected case did not meet the criteria for inclusion (for instance, the opinion contained less than 400 words, or was a French-language decision), then the next number in the list of all cases that was not already marked for inclusion within the random sample was selected instead. A copy of the Excel spreadsheet, with case names, randomly selected cases, all other cases, and notes on where and why certain cases were excluded and replaced by the next-in-line case is publicly available for each of the appeal court, superior court, and inferior court corpora.[15]

4.2.1.2 Text Extraction and File Preparation

Once a case was identified for inclusion within the corpus for the relevant court level, I navigated to the web-based publication of the decision (on CanLII or LexUM), and manually copied from the first word to the last word of the body of the decision (or opinion, as appropriate) onto my electronic clipboard. 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 that court level.

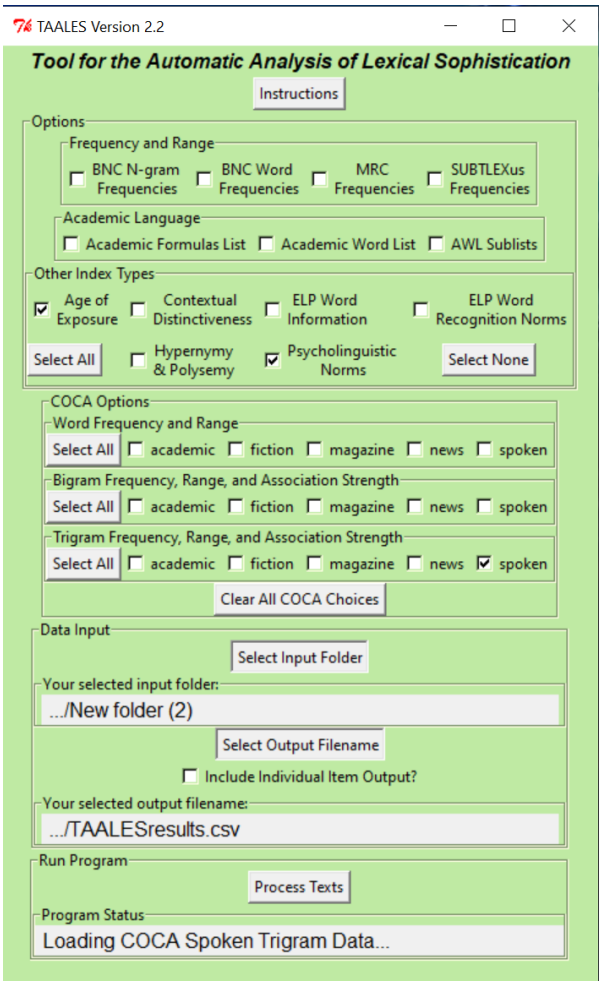
---

[15] Mike Madden, "MADRS_MaddenDissertation_Chapter4_AppealCourtSampled+Population" (2025), online: <https://doi.org/10.7910/DVN/HPFJCP>, Harvard Dataverse, V1 (Appeal Court Corpus). Mike Madden, "MADRS_MaddenDissertation_Chapter4_SupCourtSampled+Population" (2025), online: <https://doi.org/10.7910/DVN/QAHP2R>, Harvard Dataverse, V1 (Superior Court Corpus). Mike Madden, "MADRS_MaddenDissertation_Chapter4_InfCourtSampled+Population" (2025), online: <https://doi.org/10.7910/DVN/F41BKZ>, Harvard Dataverse, V1 (Inferior Court Corpus).

4.2.1.3 Linguistic Processing of Files / MADRS Calculations

I processed all text files containing court decisions/opinions through 3 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.[16] I elected to calculate "age of exposure", "psycholinguistic norms", and "Trigram Frequency, Range, and Association Strength – Spoken" using this program, as shown in the image below.

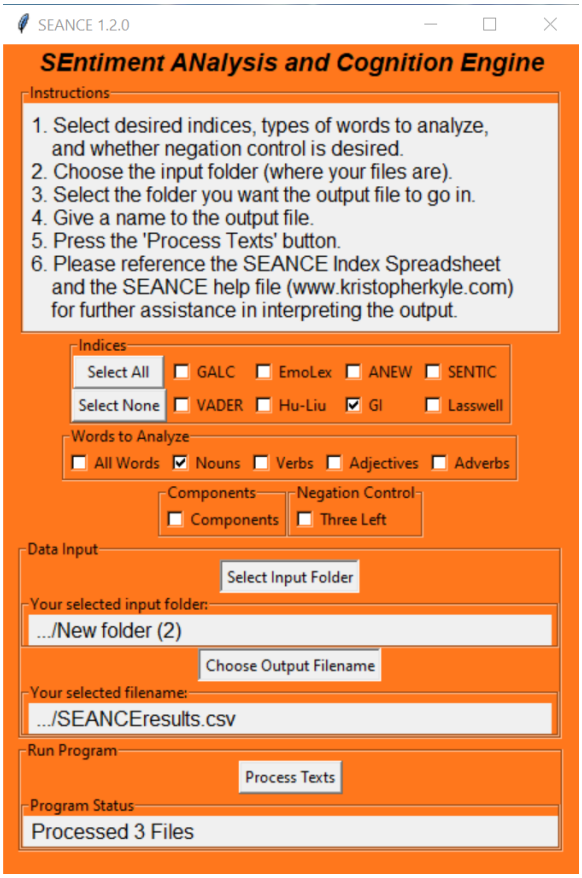


[16] Online: <https://www.linguisticanalysistools.org/taales.html>.

The output generated by TAALES saves as a comma-separated values (CSV) spreadsheet file that is readable using Microsoft Excel. The spreadsheet contains 46 columns of data on different lexical sophistication and complexity variables, including 4 of the variables that contribute to the MADRS formula. Column 5 contains data for the variable *Medical Research Council – Imageability (All Words) / MRC_Imag_AW*. Column 11 contains data for the variable *Medical Research Council – Concreteness (Function Words) / MRC_Conc_FW*. Column 15 contains data for the variable *Kuperman Age of Acquisition (Content Words) / AoA_CW*. And, column 36 contains data for the variable *COCA – Spoken (Proportion of Trigrams in the top 30K) / COCA_Spoken_TriTop30K*. I also relied on Column 2, which lists the number of words contained in every file that has been processed using TAALES. This word count served as a double-check of my initial assessment of text length for each file, and confirmed that only files with 400 or more words were included within my analysis. The remaining columns within the TAALES results spreadsheet contain data on variables that are not relevant to the MADRS formula.

I also processed the files through the *Sentiment Analysis and Cognition Engine* (SEANCE) program.[17] I elected to compute values using the General Inquirer ("GI") index, for "nouns", as shown in the image below.

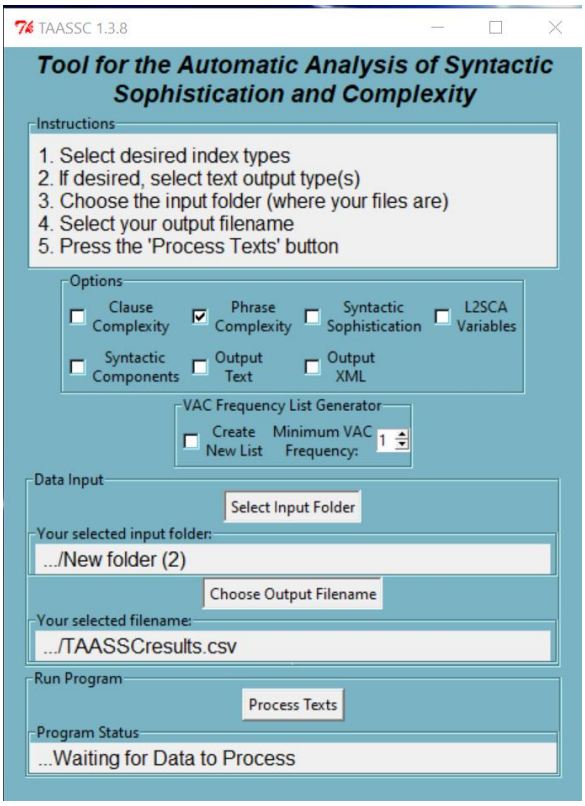
[17] Online: <https://www.linguisticanalysistools.org/seance.html>.



The output generated by SEANCE also saves as a CSV spreadsheet file that is readable using Microsoft Excel. The spreadsheet contains 121 columns of data on different sentiment, cognition, and topic-related variables, including the variable *General Inquirer Database – Legal (Nouns) / GI_Leg_Nouns*, which contributes to the MADRS formula. Data for this variable can be found in Column 30 of the SEANCE results spreadsheet. None of the other variables within the spreadsheet are relevant to the MADRS formula.

Finally, I processed the files through the Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC) program.[18] I chose to compute values only for "phrase complexity", as shown in the image below.

[18] Online: <https://www.linguisticanalysistools.org/taassc.html>.



The output generated by TAASSC also saves as a CSV spreadsheet file that is readable using Microsoft Excel. The spreadsheet contains 133 columns of data on different syntax-related variables, including the variable *Dependents per nominal (no pronouns, standard deviation) / Nominal_Deps_NN_StDev*, which contributes to the MADRS formula. Data for this variable can be found in Column 27 of the TAASSC results spreadsheet. None of the other variables within the spreadsheet are relevant to the MADRS formula. Workbooks containing the output from calculations performed by each of the above linguistic software tools for the relevant corpora are publicly available.[19]

[19] Mike Madden, "MADRS_MaddenDissertation_Chapter4_SCCLinguisticResults" (2025), online: <https://doi.org/10.7910/DVN/XVUB91>, Harvard Dataverse, V1 (Supreme Court of Canada Corpus). Mike Madden, "MADRS_MaddenDissertation_Chapter4_AppealCourtLinguisticResults" (2025), online: <https://doi.org/10.7910/DVN/7C2FTG>, Harvard Dataverse, V1 (Appeal Court Corpus). Mike Madden, "MADRS_MaddenDissertation_Chapter4_SupCourtLinguisticResults" (2025), online: <https://doi.org/10.7910/DVN/BQE66O>, Harvard Dataverse, V1 (Superior Court Corpus). Mike Madden,

Once values had been calculated by these 3 different linguistic software programs for each of the 6 variables that contribute to the MADRS formula, for each file within the different corpora that were being studied, it was possible to compute MADRS values for each file using the regression equation described in the previous chapter. This calculation was relatively straightforward to perform by first consolidating all of the data for the 6 variables on a single spreadsheet. I then entered the regression (MADRS) formula into a blank cell beside the data for the first file (which calculated that file's MADRS value). Finally, using Microsoft Excel's "fill down" feature, Excel replicated the formula and its calculations for every file within the spreadsheet, giving me the calculated MADRS value for every judicial decision/opinion within the different corpora.

4.2.2 Descriptive Results – Assessing Readability Scores Across Court Levels

For the purposes of assessing how readability levels might differ between courts of different levels in Canada, I ultimately calculated readability scores from 4 different corpora – one corpus each for decisions from the SCC, appeal courts, superior courts, and inferior courts. Relevant characteristics of each corpus are shown, below, in Table 4.1.

"MADRS_MaddenDissertation_Chapter4_InfCourtLinguisticResults" (2025), online: <https://doi.org/10.7910/DVN/WLR4GR>, Harvard Dataverse, V1 (Inferior Court Corpus).

Table 4.1 – Properties of Each Court-Level Corpus

Court Level	Number of Decision/Opinion Files	Number of Words of Text
Apex	75	633,000
Appeal[20]	183	889,000
Superior[21]	193	1,125,000
Inferior[22]	174	1,050,000
Total	625	3,697,000

Spreadsheets of the MADRS results and component variable scores for all the SCC,[23] appeal,[24] superior,[25] and inferior[26] court opinions/decisions within these corpora are publicly available. Each of the appeal, superior, and inferior court spreadsheets also include the full list of all decisions within the population, and (in red-highlighted cells) cases that were randomly selected for inclusion within the present study.

4.2.2.1 Putting MADRS Values into Context

Before presenting the readability results for these different levels of courts, I draw attention here again to the scoring scale, to reiterate what any particular score means in terms of who is likely to understand the text. The scoring scale is reproduced below:

[20] This corpus is available here: Mike Madden, "MADRS_MaddenDissertation_Chapter4_CorpusAppealOpinions" (2024), online: <https://doi.org/10.7910/DVN/WC630E>, Harvard Dataverse, V1.
[21] This corpus is available here: Mike Madden, "MADRS_MaddenDissertation_Chapter4_CorpusSuperiorOpinions" (2024), online: <https://doi.org/10.7910/DVN/KELQZC>, Harvard Dataverse, V1.
[22] This corpus is available here: Mike Madden, "MADRS_MaddenDissertation_Chapter4_CorpusInferiorOpinions" (2024), online: < https://doi.org/10.7910/DVN/XHLKQ2>, Harvard Dataverse, V1.
[23] Mike Madden, "MADRS_MaddenDissertation_Chapter4_SCCMADRSResults" (2024), online: <https://doi.org/10.7910/DVN/TI6QT9>, Harvard Dataverse, V1.
[24] Mike Madden, "MADRS_MaddenDissertation_Chapter4_AppealMADRSResults" (2024), online: <https://doi.org/10.7910/DVN/LE8YAJ>, Harvard Dataverse, V1.
[25] Mike Madden, "MADRS_MaddenDissertation_Chapter4_SuperiorMADRSResults" (2024), online: <https://doi.org/10.7910/DVN/1VWWI6>, Harvard Dataverse, V1.
[26] Mike Madden, "MADRS_MaddenDissertation_Chapter4_InferiorMADRSResults" (2024), online: <https://doi.org/10.7910/DVN/3IC6ZW>, Harvard Dataverse, V1.

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

As an additional means of situating the MADRS results from different court levels (and all other MADRS values that are subsequently reported in this dissertation) within a context that readers might understand, I have also calculated MADRS values for several other judicial texts that may be familiar to most common law lawyers, and for Chapter 1 of this dissertation. These texts, and their corresponding MADRS results, are discussed briefly below.

I begin with the results for a text that may be most fresh in the memory of any reader who is reading this dissertation in its entirety. Chapter 1 of this dissertation scores 3.52 on the MADRS scale. My quick assessment of the component variable scores for this MADRS value, compared with the average values for the component variables for 1621 decisions that are presented later in this dissertation (in Chapter 6),[27] suggest that Chapter 1 seems to use more complex syntax than the average adjudicative decision uses (such as sentences like this one with embedded clauses).[28] This syntax would drive up the MADRS value. But Chapter 1 also seems to include less legalese

[27] These results are publicly available. Mike Madden, "MADRS_MaddenDissertation_Chapter6Results" (30 May 2024), Harvard Dataverse - V1, online: <https://doi.org/10.7910/DVN/NIQ2KJ>.
[28] Chapter 1's score for the variable *Nominal_Deps_NN_StDev* was 1.231; the average for the decisions reported on in Chapter 6 was 1.157. Chapter 1 used phrases that have, on average, more dependents per nominal than phrases used within the average decision reported on in Chapter 6.

than the average adjudicative decision[29] (since the introduction, and the dissertation generally, is focused on the **form** of adjudicative decisions rather than on the **substantive content** of the decisions). The minimization of legalese (as measured by the density of words found on the General Inquirer's Legal Nouns word list) would drive down the MADRS value for this chapter. Now, I do not deceive myself into thinking that – despite the reasonably low MADRS score for Chapter 1 – there will be a massive public audience full of people who have the motivation to read this chapter or the dissertation in its entirety. But a reader who has some post-secondary education and who attempts to read Chapter 1 would likely be able to easily understand this chapter, based on the MADRS result for the chapter.

I turn now to MADRS assessments for some well-known judicial texts. One of these texts is drawn from a famous English case,[30] wherein Lord Denning opens his opinion by describing how delightful village cricket is in the summertime (before going on to deal with the law of torts that applies to that case). 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 opinion in this decision[31] (in isolation), and for Lord Denning's entire

---

[29] Chapter 1's score for the variable *GI_Leg_Nouns* was 0.044; the average for the decisions reported on in Chapter 6 was 0.114. Chapter 1 used a lower density of words found on the General Inquirer's Legal Nouns word list than the average decision reported on in Chapter 6.

[30] *Miller v Jackson*, [1977] EWCA Civ 6 (UK), online: <https://www.bailii.org/ew/cases/EWCA/Civ/1977/6.html>.

[31] The full text of the opening paragraph is 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. The village team play there on Saturdays and Sundays. They belong to a league, competing with the neighbouring villages. On other evenings after work they practice while the light lasts. Yet now after these 70 years a Judge of the High Court has ordered that they must not play there anymore, He has issued an injunction to stop them. He has done it at the instance of a newcomer who is no lover of cricket. This newcomer has built, or has had built for him, a house on the edge of the cricket ground which four years ago was a field where cattle grazed. The animals did not mind the cricket. But now this adjoining field has been turned into a housing estate. The newcomer bought one of the houses on the edge of the cricket ground. No doubt the open space was a selling point. Now he complains that, when a batsman hits a six, the ball has been known to land in his

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) 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),[32] and *Donoghue v Stevenson* (involving a woman who felt ill after drinking a ginger beer from a bottle that contained a decomposed snail).[33] I computed MADRS values for Justice Cardozo's majority opinion in the former case, and for the complete decision (consisting of 5 separate opinions) from the latter case. The railway case scored 4.7, and the snail-in-the-bottle case 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.

> garden or on or near his house. His wife has got so upset about it that they always go out at weekends. They do not go into the garden when cricket is being played. They say that this is intolerable. So they asked the Judge to stop the cricket being played. And the Judge, I am sorry to say, feels that the cricket must be stopped: with the consequences, I suppose, that the Lintz cricket-club will disappear. The cricket ground will be turned to some other use. I expect for more houses or a factory. The young men will turn to other things instead of cricket. The whole village will be much the poorer. And all this because of a newcomer who has just bought a house there next to the cricket ground.

[32] 298 NY 339 (App Ct 1928), online: <https://www.nycourts.gov/reporter/archives/palsgraf_lirr.htm>.
[33] [1932] UKHL 100 (UK), online: <https://www.bailii.org/uk/cases/UKHL/1932/100.html>.

Finally, I have examined a Canadian example of a judicial author – Justice Watt, formerly of the Court of Appeal for Ontario, now retired – whose decisions are said to read more like a crime novel, and who "is seen by some as part of a welcome trend away from impenetrable, legal jargon."[34] I calculated MADRS values for one of Justice Watt's notable decisions,[35] and have again produced scores for the opening section of the decision in isolation,[36] 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 Justice Watt discusses things like the four threshold requirements 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).

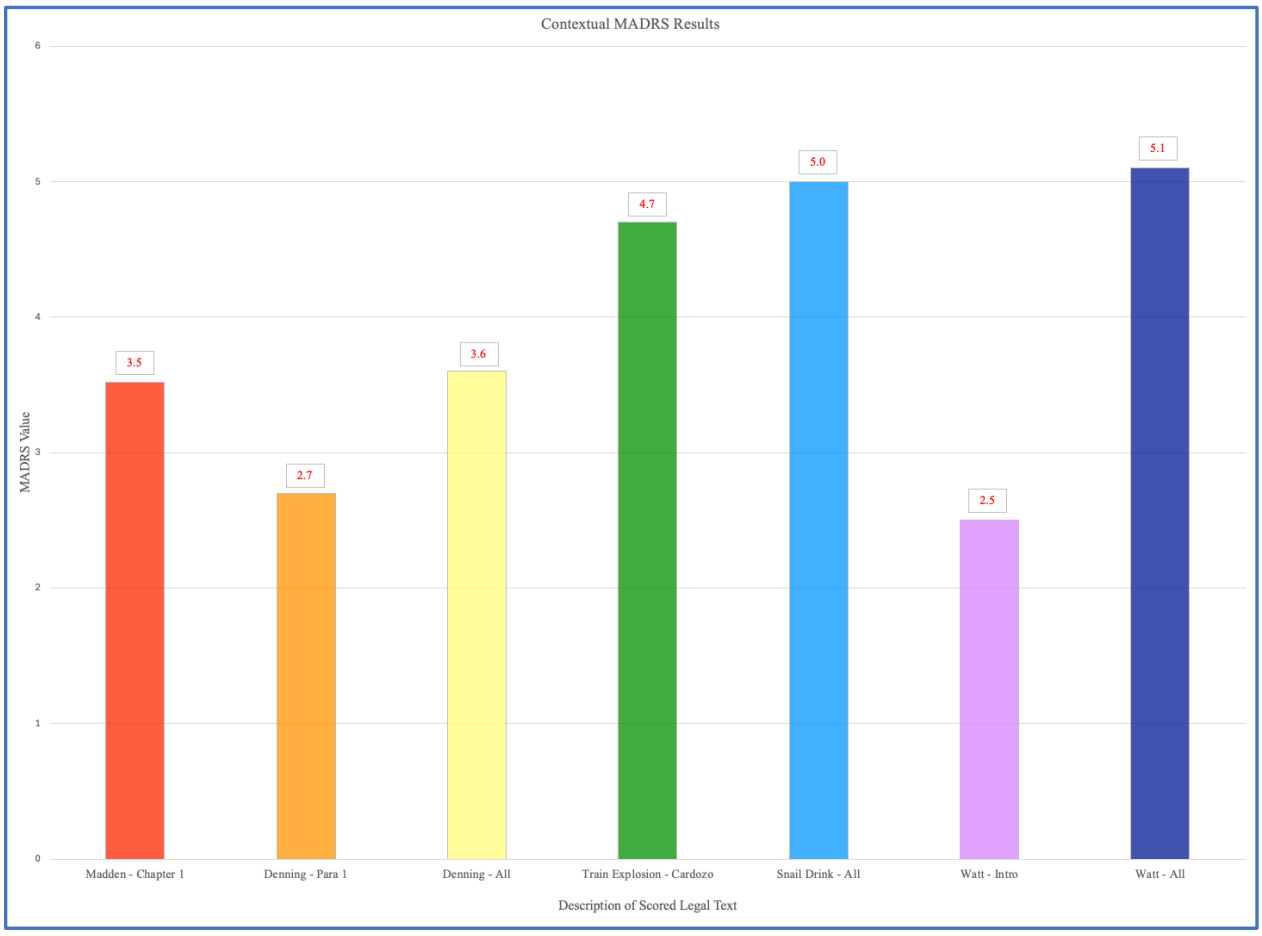
---

[34] Kirk Makin, "The Judge Who Writes Like a Paperback Novelist," *The Globe and Mail* (March 10, 2011), online: <https://www.theglobeandmail.com/news/national/the-judge-who-writes-like-a-paperback-novelist/article570811/>.
[35] *R v Shafia*, 2016 ONCA 812.
[36] The first 8 paragraphs of the decision that make up the opening section read as follows:

> 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.
> And inside the vehicle, a terrible loss of life. Four dead family members. Three young women. One adult.
> About three weeks later, three arrests. Three members of the same family. The father, mother and brother of the young women. Each charged with four counts of first degree murder.
> At the end of a joint trial, each person charged was found guilty of four counts of first degree murder. Each seeks a new trial. One, Hamed Shafia (Hamed), the eldest son of Mohammad Shafia (Shafia) and Tooba Mohammad Yahya (Tooba) says he should not have been tried with the others – his parents – because he was too young to be tried as an adult. Hamed asks us to admit fresh evidence about his age that proves, so he says, that he was tried in the wrong court. All three complain that, among other things, the trial judge wrongly admitted some expert opinion and hearsay evidence, then made matters worse by not properly instructing the jury about how they could use that evidence to help them decide the case.
> As I will explain, I would not give effect to any of the grounds of appeal and would dismiss the appeals.

The MADRS results for each of the above contextual documents are shown below, in Chart 4.1.

Chart 4.1 – Contextual MADRS Results



Ideally, my discussion of these examples and their MADRS values will help to do two things. First, the examples should illustrate through familiar texts how the 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. Second, and perhaps more importantly, some of the examples help to show how different parts of a text may have different readability scores. A particularly readable (or unreadable) portion of a text does not necessarily mean that the text as a whole is consistently written at that same level of readability. This reality becomes important if

one is interested in the readability levels of court and tribunal decisions as wholes, rather than just in the readability levels of the more memorable, or sensational, parts of these decisions. While there is nothing wrong with targeted testing of readability levels for particular portions of a decision, one should not extrapolate from a portion to reach conclusions about the whole. For instance, if a person with only secondary school education were truly interested in understanding Justice Watt's reasoning for upholding murder convictions in the above case, the reader would not necessarily be assisted by the highly readable language that Justice Watt uses in the opening section of the decision (wherein he discusses facts and his conclusion, without any explanation as to how he reaches his conclusion), and might be confounded by the less readable language that the judge uses throughout the analytical portion of his opinion.

4.2.2.2 Overall MADRS Values by Court Level

A workbook containing the full results for decisions in each corpus (separated by tabs), including component scores for each contributing variable, is publicly available.[37] A graphical representation of MADRS results by court level is shown below in Chart 4.2, in the form of boxplots. For each court level, the "box" represents the interquartile range[38] for MADRS values at that court level. 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 for that court level, 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

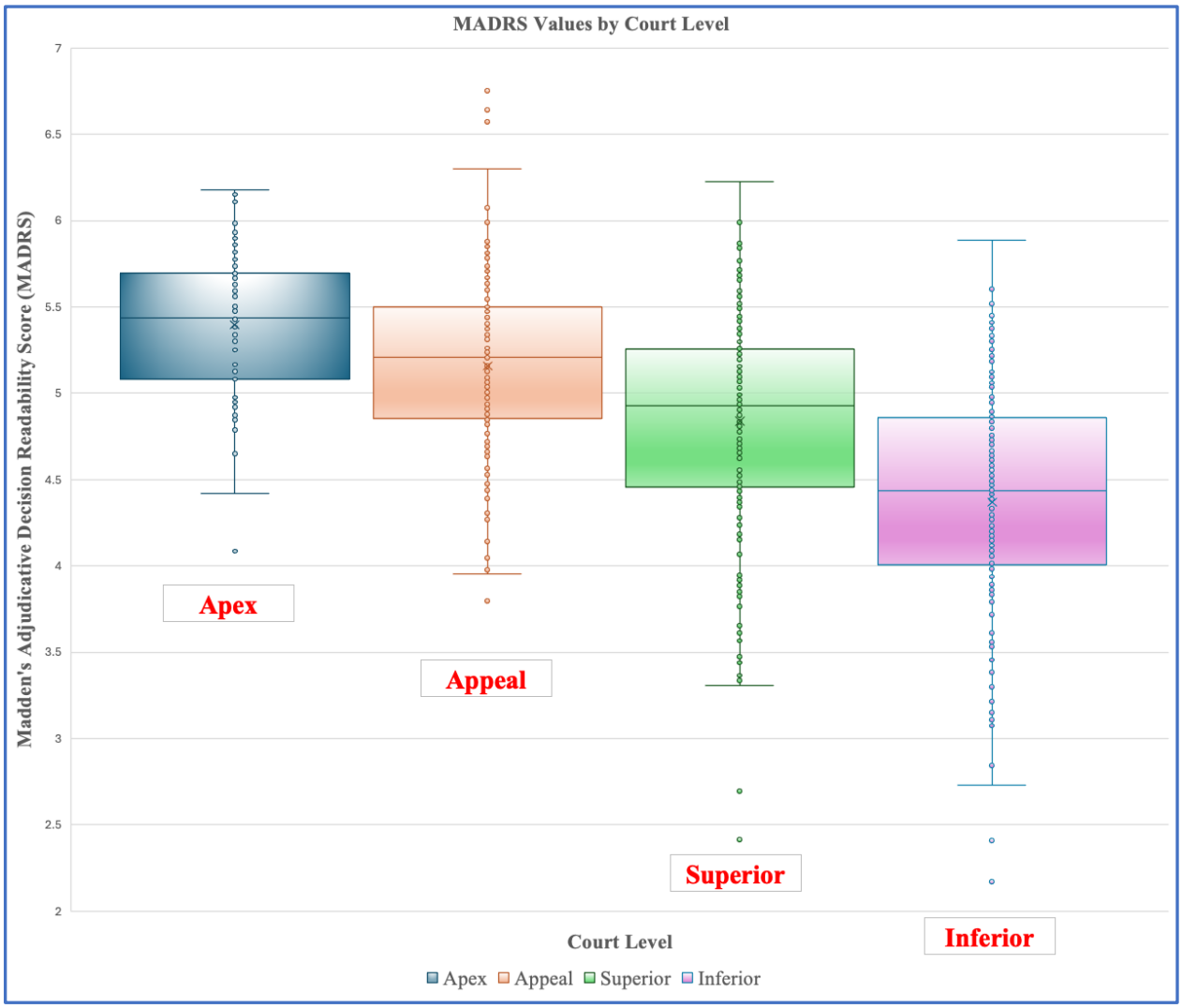
---

[37] Mike Madden, "MADRS_MaddenDissertation_Chapter4_CourtLevelResults" (2025), online: <https://doi.org/10.7910/DVN/CX028J>, Harvard Dataverse, V1.

[38] "Interquartile range" refers to the range between the 25th and 75th percentiles; it offers a view of the middle 50% of the data points, and is less sensitive to outlier points than other dispersion measures: Michael O Finkelstein & Bruce Levin, *Statistics for Lawyers*, 3rd ed (New York: Springer, 2015) at 25.

are noteworthy for their distance away from the central tendency (that is, the average) for that court level. Similar boxplots are shown for various other data in subsequent Charts (below), and should be read in the same manner as the current boxplot.

Chart 4.2 – MADRS Values by Court Level



As this chart shows, decisions decrease in readability (that is, they have higher MADRS values) for courts that are higher in the judicial hierarchy – with the SCC having the least readable decisions, and inferior courts having the most readable decisions. The average scores for each level of court were the following: Apex – 5.40; Appeal – 5.16; Superior – 4.84; and, Inferior – 4.37. As

I noted earlier, at section 4.2.1 when I described my methods for this study, the Apex court database of decisions included all opinions types from the court (dissenting, concurring, majority, and unanimous), whereas only majority or unanimous (i.e.: controlling) opinions were assessed for all other court levels. This difference in corpus design for the different courts' corpora of decisions could have skewed the above results.

On closer examination, however, the skew appears to have been minimal, and does not change the relative readability hierarchy that is shown above. Specifically, there were 6 concurring SCC opinions in the corpus with an average score of 5.58 for these decisions; 20 dissenting opinions with an average score of 5.41; and, 49 majority or unanimous opinions with an average score of 5.37. In other words, even if only the controlling SCC opinions were considered for the purposes of comparing that apex court's opinions against the opinions of courts from other levels in Canada, the results would be almost the same (but with an average of 5.37 for the SCC, instead of the 5.40 that is represented in the above chart).

The most readable decision from any of these courts was *R v Sivapragasam*, 2022 ONCJ 443,[39] which scored 2.17. This was a criminal case (involving 5 assault-related charges) that was written by Justice West. The decision spanned 83 paragraphs and consisted of over 12000 words. Another 3 of the next 4 most readable cases were also criminal cases, while one was a family law case. These cases scored 2.41, 2.41, 2.42, and 2.42, and were from the Ontario Court of Justice (again, a decision of Justice West),[40] New Brunswick Court of King's Bench,[41] Provincial Court of British Columbia,[42] and Ontario Court of Justice,[43] respectively.

---

[39] Online: <https://www.canlii.org/en/on/oncj/doc/2022/2022oncj443/2022oncj443.html>.
[40] *R v Boyle*, 2022 ONCJ 177.
[41] *RV v KM*, 2022 NBQB 105.
[42] *R v MS*, 2022 BCPC 129.
[43] *R v Knight*, 2022 ONCJ 211.

The least readable decision was *Miller v Canada (National Revenue)*, 2022 FCA 183,[44] which scored 6.75. The decision was a tax appeal case, written for a unanimous court. It was 86 paragraphs (11129 words) long. My sense is that this decision quotes far more extensively from a bilingual statute than one would expect from a typical decision, so it is possible that the inclusion of such extensive statutory language, and French language, within the decision, may have contributed to the readability score. The second-least readable decision was a criminal sentencing appeal from the Court of Appeal for Ontario,[45] that was collectively authored by 3 judges, and consisted of only 755 words. Despite the short length of the decision, it contains a high density of language that is more particular to the legal profession than to everyday English.[46]

4.2.3 Discussion and Analysis – Assessing Readability Scores Across Court Levels

The above results offer a useful picture of the general readability levels of Canadian court decisions, broken down in accordance with the levels of courts that issue the decisions. The results also raise both questions and observations that are worth exploring.

4.2.3.1 The Big Picture: How Readable are Court Decisions?

To begin with, there is a strong chance that most Canadian court decisions could not be easily understood by large segments of the Canadian population. Within the 625 decisions / opinions that were studied here from across all levels of courts, only 11 had MADRS values of less than 3.0 – which is to say, only 11 of 625 texts could be easily understood by an adult who has

---

[44] Online: <https://www.canlii.org/en/ca/fca/doc/2022/2022fca183/2022fca183.html>.
[45] *R v Martin*, 2022 ONCA 157.
[46] For instance, the longest paragraph in the 11-paragraph decision reads as follows (at para 5):

> The appellant argues that the sentencing judge erred by failing to give effect to the parity principle when imposing the sentence. While the appellant acknowledges that the sentencing judge properly adverted to and described the parity principle, he contends that, on the face of the sentences imposed and the purported gulf between them, there is a clear misapplication of that principle. Given their almost identical roles in the crime, the appellant maintains that his sentence should not have been almost twice that of his co-accused, particularly since his co-accused has a criminal record and he does not. Therefore, he asks this court to set aside his sentence and impose a sentence of 15-months in duration. We decline to do so.

less than a completed secondary school education. Statistics Canada census data from 2021 indicates that 11.6% of individuals in Canada between the ages of 25-64 do not have a "high school diploma or equivalency certificate."[47] For these individuals, the overwhelming majority of court decisions could not be easily understood. And, there is no compelling reason to believe that this 11.6% of the Canadian population do not find themselves in family court, or criminal court, or in other situations where they may have a need to understand court decisions.

A perhaps even more concerning reality is that none of the decisions or opinions in the present study scored less than 2.0 on the MADRS scale – which means that none of these decisions are likely to be easily understood by people who have no secondary school education, or who are less than fluent in English. Statistics Canada census data from 2021 indicates that 6.8% of the population does not regularly speak French or English at home.[48] This statistic can be contrasted with a related statistic indicating that 15.5% of the population regularly speaks a language that is not English or French at home, while also regularly speaking English or French at home.[49]

We might infer that most of the people described by the latter statistic have enough competence in English or French such that their fluency in one or the other of these languages, alone, is not likely to represent a major barrier to understanding adjudicative decisions written in the relevant language. However, we might be far less willing to draw the same inference about people described by the former statistic. Although we do not have clear data about the English or French fluency levels of people who do not speak English or French at home regularly, one could

---

[47] Statistics Canada. "Table 98-10-0385-01 – High school completion by census year: Canada, provinces and territories, census metropolitan areas and census agglomerations" (Oct 4, 2023), online: <https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=9810038501>.

[48] Statistics Canada. "Table 2 – Language spoken at least regularly at home, provinces and territories, 2021" (Aug 17, 2022), online: < https://www150.statcan.gc.ca/n1/daily-quotidien/220817/t002a-eng.htm>.

[49] Statistics Canada. "Table 2 – Language spoken at least regularly at home, provinces and territories, 2021" (Aug 17, 2022), online: < https://www150.statcan.gc.ca/n1/daily-quotidien/220817/t002a-eng.htm>.

infer that many of these people do not regularly speak English or French at home because they are not sufficiently competent in these languages to easily speak and understand them. To be clear, many of these people may simply choose, for any number of reasons, to speak other languages at home to the exclusion of English and French – despite their fluency in the latter languages. The point, however, is that there may be a non-negligible percentage of Canada's population (potentially amounting to millions of people) whose limited fluency in English, apart from any other contributing factor, prevents them from understanding adjudicative decisions easily. These people would likely be unable to understand any of the decisions in the present study.

We do not have data about how many of the above people (with less than a completed high school education, or less than fluent abilities in English) actually want to read court decisions, or have tried and failed in their efforts to understand such decisions. However, the above discussion at least suggests that there may be a communication problem here that requires the attention of our courts. Practically speaking, if someone with less than a completed high school education (or less than fluent abilities in English) ever wanted to read and understand a Canadian court decision, the odds are almost certain that they would be unable to do so. Where judges purport, or hope, to be writing in ways that make their decisions understandable to ordinary Canadians (as their own words suggest that they do),[50] then the above results raise a red flag: potentially millions of people in Canada do not have the educational or literacy skills that appear to be required to understand approximately 98% (614 of 625) of the decisions released by courts of various levels in 2022.

---

[50] See, e.g., Canada, Department of Justice. *The Honourable Sheilah Martin's Questionnaire (Questionnaire for the Supreme Court of Canada Judicial Appointment Process)* (21 December 2021), at Question 4. Online: <https://www.fja.gc.ca/scc-csc/2017-SheilahMartin/nominee-candidat-eng.html>; Canada, Department of Justice. *The Honourable Nicholas Kasirer's Questionnaire (Questionnaire for the Supreme Court of Canada Judicial Appointment Process)* (18 April 2019), at Question 4. Online: <https://www.fja.gc.ca/scc-csc/2019/nominee-candidat-eng.html>; and, Canada, Department of Justice. *The Honourable Mahmud Jamal's Questionnaire (Questionnaire for the Supreme Court of Canada Judicial Appointment Process)* (17 Jun 2021), at Question 4. Online: <https://fja-cmf.gc.ca/scc-csc/2021/nominee-candidat-eng.html>.

Potentially millions more people would similarly struggle because of their limited fluency in English.

Furthermore, 286 of the 625 studied decisions / opinions scored higher than 5 on the MADRS scale, indicating that just over 45% of all the decisions in this sample could only be easily understood by a person with "an undergraduate degree or higher level of non-legal education." However, Statistics Canada data from the same 2021 census indicates that only 32.9% of the Canadian population between the ages of 25-64 have a "Bachelor's degree or higher."[51] The disconnect between population capacity to understand a text, and the level at which the text is written, is even more clear here: more than two-thirds of the population would be unable to easily understand almost half of Canada's court decisions.

From a practical perspective, it is impossible to determine whether this disconnect creates a problem that needs to be solved. In order to make that determination, one would probably need more information about whether actual readers (that is, people who are actively trying to understand court decisions) are unable to understand what they are reading. If the evidence (that does not currently exist, and that could not easily be obtained) suggested that this were not the case, then perhaps the disconnect is inconsequential. After all, does it really matter if someone who has no interest in reading a court decision cannot actually understand the content of that decision? Perhaps courts are already tailoring their writing to match the needs of their audiences most of the time, such that the 45% of court decisions that score above 5 on the MADRS scale are probably only of interest to people with a "Bachelor's degree or higher" level of education in any case.

---

[51] Statistics Canada. "Table 98-10-0384-01: Highest level of education by census year: Canada, provinces and territories, census metropolitan areas and census agglomerations" (December 9, 2022), online: <https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=9810038401>.

This hypothesis seems tremendously implausible, since legal issues (and particularly public law legal issues) cut across all elements of the population, regardless of one's education level. People without university degrees still pay taxes, get divorced, commit crimes, invent new products that require patents, get fired unjustly from their jobs, and generally live their lives in wide varieties of ways that might require them to interact with Canada's legal and court systems. In other words, these people might need to read court cases – either to understand their legal rights and obligations as a starting point (to assess things like whether they need a lawyer, could proceed to court without a lawyer, or have no legal case worth pursuing), or to understand how their situation has been dealt with by the court after the fact (if they find themselves before a court in defence of their legal rights). As Semple has observed on this point, when "there are no recent decisions explaining the law, parties will be unable to understand and negotiate for their own rights even if they intend and can afford to. Reported decisions create a 'shadow of the law' that lets people understand how the law applies to their own situations, which in turn lets them avoid or settle disputes on fair terms."[52] But people need to understand the decisions in order for the texts to have this beneficial effect.

Regardless of how implausible the above hypothesis might be about courts already calibrating the readability levels of their decisions to meet the needs of people who are actually reading the decisions, there is arguably a causation question here that also needs to be considered. Are there large numbers of ordinary Canadians who are uninterested in reading, or who are currently not trying to read, Canadian court decisions because there is no inherent need for them to ever read these decisions? Or is it possible that many Canadians are currently uninterested in

---

[52] Noel Semple, "Tribunals for Access to Justice in Canada", 2024 CanLIIDocs 2368, online: <https://canlii.ca/t/7ng33>, at 19.

reading such decisions because these people are incapable (or believe that they are incapable) of understanding them? Would more Canadians read court decisions if they were written more readably? Before drawing any conclusions about whether Canadian court decisions are appropriately readable right now that are grounded in arguments about who is actually interested in reading these decisions, one would likely need to answer the above causation questions – ideally from an empirical perspective. Depending on the answers to these questions, it might be possible to conclude (in a way that is somewhat circular and unhelpful) that Canadian court decisions are currently readable to the people who are reading them, but to also conclude that these decisions are not sufficiently readable to be understood by some or many people who want to read them, but are currently making no effort to read them because they are incomprehensible (or are perceived to be incomprehensible) to these people. My intuition suggests that many court decisions are already not understandable to large groups of people who are trying, or who have an interest in trying, to read and understand the decisions.

Even setting aside the empirical question about how many people who want to read and understand court decisions are actually unable to do so, I believe that there is a strong normative argument in favour of making almost all court decisions as widely readable as possible. If it is important for the people of Canada to be capable of seeing and understanding the work that their courts are performing as a way of supervising or ensuring the public accountability of the courts,[53] then surely these rights belong to all citizens, and not only the more educated and more literate citizens. And if parties who are affected by court decisions should be capable of understanding the reasons for disposition of their cases by the courts (out of fairness and respect to the parties), then,

---

[53] Joseph Goldstein, *The Intelligible Constitution: The Supreme Court's Obligation to Maintain the Constitution as Something We the People Can Understand* (New York: Oxford UP, 1992) at 19.

again, surely this measure of fairness and respect should extend to all parties, not just the ones who read at higher competency levels.

This normative argument for readable court decisions is completely distinct from the empirical argument about how readable court decisions ought to be. The normative argument suggests that people have an equal right to be able to understand court decisions, and this right is undermined when it can only be accessed by individuals along a sliding scale, with higher levels of access afforded to those among us who are more educated, and lower levels of access (or, realistically, no meaningful access) for those among us who have much less formal education.

I accept that all normative arguments must be bounded by practical considerations at a certain point, and that there is likely a limit after which efforts to make decisions more readable to more people becomes unhelpful. For instance, there is likely a point after which further efforts to refine language will result in court decisions that are so linguistically simple that they lose important legal specificity, thereby creating harmful uncertainty in the law. Similarly, there is probably a level of time increment after which any additional time spent by an author in editing and revising a document to make it more readable becomes disproportionate to the readability gain that could be achieved by the author: the extra time cost would not be worth the benefit of making the document more understandable to a small number of additional people.

The above MADRS results, however, suggest that most courts and judges have probably not yet reached these points of diminishing or negative returns yet. Inferior court judges have demonstrated that it is possible to write decisions that are, on average, more than a full point more readable on the MADRS scale than the SCC's decisions. There could be a variety of explanations for the differences in scores (and I will discuss some of these potential explanations below). But the example set by inferior courts provides persuasive evidence that it is possible to perform the

function of judging (as these courts do) in ways that do not appear to attract any greater criticism than when this function is performed by higher levels of courts, while still writing decisions that are relatively more readable. So, this first set of MADRS results provides one building block for the foundation of an argument that many Canadian courts can and should produce more readable opinions than they are currently producing.

4.2.3.2 Assessing the Differences Across Court Levels

As I noted above, there are good reasons why we might expect or want decisions to be more readable from courts that are higher on the hierarchy, but also good reasons to expect or want the same thing from courts that are lower on the hierarchy. The results contained in Chart 4.2, above, show a clear decline in MADRS scores (or increase in readability levels) as one moves down the hierarchy from apex, to appeal, to superior, to inferior courts. It is worth considering why decisions are more readable from lower courts, and also whether this tendency for higher courts to write less readable decisions is desirable or acceptable from a theoretical perspective.

One explanation for the readability score differences may come from the individualized nature of a trial court judge's work, in contrast with the collegial nature of an appeal or apex court judge's work. For decisions whose results are included in Chart 4.2, above, all of the inferior court decisions were authored by single judges – as one would expect from this level of court where individual judges preside over trials. And, although at least one superior court (Ontario's Superior Court of Justice) cross-appoints its judges to a sort of *ad hoc* appellate court (Ontario's Divisional Court, which "hears statutory appeals from administrative tribunals"[54] that are "usually heard and

[54] Superior Court of Justice (Ontario), "Divisional Court", online: <https://www.ontariocourts.ca/scj/divisional-court/>.

decided by a panel of three judges, but may be heard by a single judge in some circumstances"),[55] none of these superior-appellate Divisional Court decisions were included within the sample represented within Chart 4.2, above. However, all of the apex court decisions within the sample were written by SCC judges who presided as part of panels, and the smallest panels within the sample were panels of 5 SCC judges. It is similarly the general practice of all Canadian courts that I have classified as appeal courts to determine decisions on the merits of an appeal in panels of 3 judges (although some of these courts use panels of 5 judges in a very small number of cases). These courts each also assign a single judge, presiding alone, to decide motions. Although I did not keep statistics on the proportion of decisions within my appeal courts corpus that were decided by panels rather than single judges, there were at least some decisions that were decided by single judges, although many or most were decided by panels of judges.

Based on this description of how courts at different levels assign judges to decide cases, it would be accurate to say that all of the inferior and superior court decisions within the above set of results were decided by single judges, working alone. The average number of judges who participated in the decisions within the "appeal court" corpus was necessarily greater than 1 but less than 5 judges (although I would estimate that this average was likely close to 2.5 judges). The average number of SCC judges who participated in decisions within the apex court corpus was greater than 5 but less than 9 judges. In other words, more judges tended to be involved in the decision-making process – although not necessarily in the judgement-writing process – at the appeal court level than at the superior and inferior court levels, and even more still were involved at the apex court level.

---

[55] Superior Court of Justice (Ontario), "Divisional Court", online: <https://www.ontariocourts.ca/scj/divisional-court/>.

The number of judges who were involved in a decision may be a relevant readability factor. In a previous study, I measured readability scores (using a general-purpose readability formula) for judicial opinions from apex court judges across different English-speaking countries (including Canada).[56] The results of this previous study demonstrated that there was a low strength and statistically significant correlation ($r = 0.133$, $p < 0.001$) between reading complexity scores, and the percentage of judges on a panel who agreed with the opinion.[57] In other words, in cases where more judges signed on to opinions, the opinions tended to be less readable than in cases where fewer judges endorsed the opinions. The same study, which reported on results of 939 opinions from the different countries' courts, also found that the average readability score for individually-authored opinions (N=746) was substantially better than for collectively-authored opinions (N=193).[58] Taken together, these two previous results might suggest that the process of authoring a decision in isolation – without any co-authors, and without any possibility of attracting other judges to endorse one's opinion – may be associated with better readability scores.[59] Again, however, the association does not signal causation. We cannot know from this research whether judges write more linguistically complex decisions because they know the decisions will be

---

[56] Mike Madden, "Judging Readability: A Study of Opinions by Apex-Court Judges from Australia, Canada, South Africa, the United Kingdom, and the United States" (2024) 21 The Scribes Journal of Legal Writing 15 ["Judging Readability"].
[57] Madden, "Judging Readability", *supra* note 56 at 46.
[58] Madden, "Judging Readability", *supra* note 56 at 39.
[59] I am indebted here to the judge of the Court of Appeal for Ontario (who I will not name to protect their anonymity) who suggested this hypothesis to me when I presented my findings from this previous research at Osgoode Hall Law School's *Refugee Law Lab Online Seminar Series*: Mike Madden, "Judging Readability: A Study of Opinions Written by Apex Court Judges from Australia, Canada, South Africa, the United Kingdom, and the United States" (October 20, 2022). I had expressed surprise during my presentation that collectively-authored opinions were less readable than individually-authored opinions. The judge observed that, in the judge's experience, the process of compromise involved in writing an opinion either with, or for the agreement, of other judges results in trade-offs that he felt would foreseeably reduce an opinion's readability. This judge's perspective accords with comments from appeal court judges elsewhere. *See, e.g.*, Patricia M Wald, "How I Write" (1993) 4 Scribes Journal of Legal Writing 55 at 58: "persistent colleagues have a wearying habit of pressing for the excision of a particular word, phrase, sentence, or paragraph in the middle of an opinion; and often in our desire to get approval and publish, we will comply. The result is a choppy or erratic transition or no transition at all. Dissents are literarily liberating in that respect."

circulated to their peers on a panel for comment or endorsement (i.e.: the primary audience for a panel decision becomes a judge's fellow panel members), whether the decisions only become more complex after other judges have offered their comments and a multi-perspective opinion takes shape, or whether the readability differences between isolated-judge and panel-judge opinions are simply different for coincidental or other as-yet-unexplained reasons.

It may also be the case, however, that lower courts develop an ability to write more readably because the legal subject matter of their decisions is more restricted than higher courts. This hypothesis provides a plausible explanation as to why inferior court decisions are generally more readable than higher court decisions: inferior courts deal mostly with criminal law matters, and in some jurisdictions with family law and/or small claims matters.[60] Perhaps judges of these inferior courts become proficient at writing decisions in simple terms because the different legal concepts that they need to explain are more finite and narrow in comparison to the range of legal concepts that higher courts encounter. Or, perhaps courts with narrower jurisdiction happen to only deal with cases that fall into legal subject areas that are objectively less complex to describe in English: perhaps criminal and family law, for instance, are inherently easier topics for readers to understand than other legal topics.[61] The problem, however, is that this hypothesis that links narrower jurisdiction to improved readability scores does not explain the readability level differences between superior and appellate courts. These two levels of courts tend to deal with identically broad ranges of legal subject areas within their decisions, but superior courts have noticeably lower MADRS scores than those of appeal courts. The breadth of legal subject areas that are dealt with

[60] The limited jurisdictional scope of inferior courts' authorities can be seen in the various statutes that empower each inferior court, but was seen even more specifically in the very narrow class of subjects that were discussed within the corpus of inferior court decisions that formed part of this study.

[61] This hypothesis will be examined specifically within a subsequent section of the dissertation, below, wherein family law, tax law, and criminal law MADRS results are contrasted with one another.

at any particular court level therefore offers, at best, an incomplete explanation for readability differences across court levels.

Even without comparing the readability results for apex court decisions to those of other courts, the SCC's results are somewhat disappointing. Theoretically, the audience sizes for SCC decisions are likely to be larger than the audience sizes for decisions from other levels of courts, since SCC decisions apply across all of Canada, and tend to deal with cases of greater public importance. Where the MADRS average for SCC opinions (5.37 or 5.4, depending on whether dissenting and concurring opinions are excluded or included, respectively) indicates that something between a completed undergraduate degree and a law degree would be required to understand the opinions, and where even the most readable opinion from the court in 2022 was scored as 4.1 (signaling that at least some post-secondary education is needed to understand the decision), it seems clear that this court is currently incapable of reaching a large portion of Canada's population through its decisions.

This reality seems to be at odds with the descriptions of the intended audiences for SCC decisions that have been identified by recently appointed judges of that court.[62] Perhaps the aspiring judges who offered their statements about who they see as the audiences for SCC

---

[62] See, e.g., Canada, Department of Justice. *The Honourable Mary Moreau's Questionnaire (Questionnaire for the Supreme Court of Canada Judicial Appointment Process)* (26 Oct 2023) (online: <https://fja-cmf.gc.ca/scc-csc/2023/nominee-candidate-eng.html>):

> The audience for Supreme Court of Canada decisions is broad and diverse because its decisions have a significant impact on all sectors of Canadian society and all aspects of life. […] The Supreme Court's decisions also speak to jurists, law students and the general public, who make up not only today's audience, but also tomorrow's. When addressing this audience, the Court must explain the legislation, case law, facts, values and social context so that its reasons will be clear and intelligible not only at the precise moment the judgment is issued, but also for years into the future.

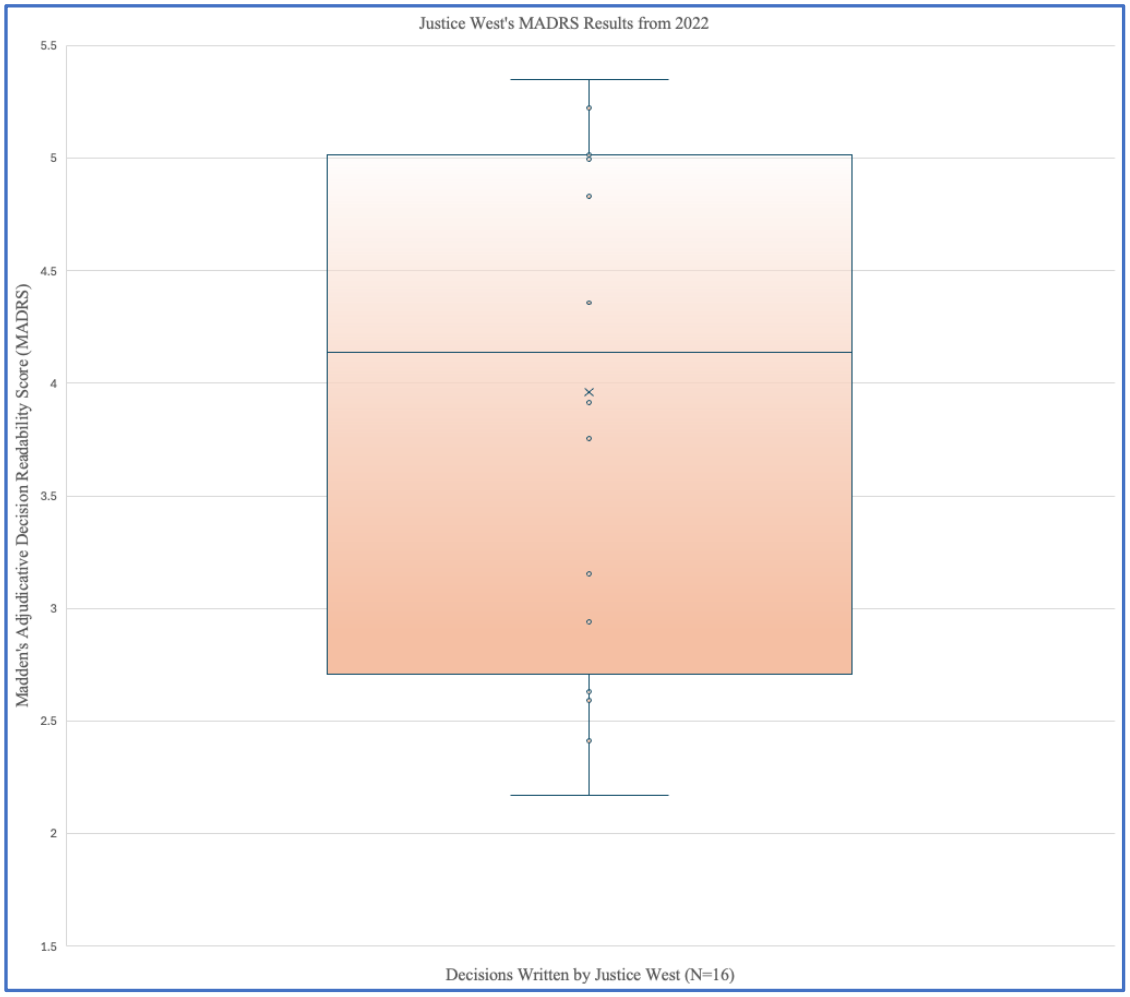
See also, Canada, Department of Justice. *The Honourable Michelle O'Bonsawin's Questionnaire (Questionnaire for the Supreme Court of Canada Judicial Appointment Process)* (22 Aug 2022) (online: <https://fja-cmf.gc.ca/scc-csc/2022/nominee-candidat-eng.html>): "Because the audience of a decision is varied, a judge must write in language that is clear, concise and easy to understand. […] The public must be able to easily understand the reasoning process behind the judge's decision and to conclude that the judicial proceeding was fair. A judge should keep the attention of the audience during the entire reading of the decision."

decisions realized after appointment that the job simply does not allow them to write for all of the audiences who they wanted to reach when they expressed their candidacy for the SCC. Or, notwithstanding their statements within their questionnaires that they complete as part of their candidacy – which they now likely know will be made public, and which therefore must be politically palatable to public audiences – the judges have always understood their primary audience to be members of the profession of law: practicing lawyers and other subordinate judges. Perhaps the nod to members of the public, and the parties, is simply lip service. But this hypothesis is inconsistent with my own (impressionistic, but extensive) experiences interacting with judges across a wide range of circumstances over the last 15 years, since I first entered law school. My sense is that judges, almost without exception, are people with integrity, strong work ethics, and a concern for those who are both immediately and more remotely affected by their decisions. So when I point out that there seems to be a disconnect between judges' descriptions of their intended audiences, and their MADRS-predicted ability to actually be understood by many members of these audiences, I attribute the disconnect to a failed objective, rather than a false objective. My hope is that through studies like the present one, judges who realize that they may not be reaching their audiences will take steps to change their writing in ways that increase the likelihood that they will reach these audiences in future decisions. I return to this idea later, in Chapter 7, when I discuss my recommendations that flow from this dissertation.

Even though there are signs that some judges may not be writing as readably (based on the MADRS formula) as they might like, there are also causes for optimism within the results. Justice West produced two decisions that would likely be easily understandable by adults who have not completed secondary school education. These results from Justice West on decisions that had been randomly included within my sample caused me to wonder whether all of his decisions might be

more readable than the decisions of other judges, so I calculated MADRS values for all of his reported decisions from 2022.[63] The average MADRS result was 3.96 (with a range of between 2.17 and 5.35). This average was 0.4 points lower than the average readability score for inferior court judges and was 1.44 points lower than the average opinion from SCC judges in 2022. A boxplot of Justice West's MADRS results for his 16 reported decisions from 2022 is shown below, at Chart 4.3.

Chart 4.3 – Justice West's MADRS Results from 2022



[63] There were 16 reported decisions by Justice West in the CanLII database from 2022.

I have no reason to believe that Justice West is writing readable decisions that are wrong in law, or otherwise lacking in some other dimension of writing or judicial decision-making quality. For instance, none of his 2022 decisions were the subject of appeals to the Court of Appeal for Ontario that are reported in the CanLII database. Furthermore, between 2020-2023, only 4 of his decisions have been reported in appeal judgements from the Court of Appeal[64] – and none of these decisions were overturned (although the terms of an internet prohibition order that was issued by Justice West were varied slightly in one appeal).[65]

Similarly, I have no reason to suspect that Justice West is sacrificing productivity in order to make his decisions more readable: his total of 16 reported decisions in 2022 seems to fall in the middle of the range of between 9-22 reported decisions that I observed from a small sample of other judges from the Ontario Court of Justice that I searched for in CanLII (although it must be understood that the number of a judge's decisions that are reported is not necessarily representative of the total number of decisions that a judge makes in a year, for a variety of reasons relating to the ways in which inferior court decisions are reported). My point is not to confirm that Justice West is as productive as his peers, but simply to note that there do not appear to be reasons to question his productivity, or the quality of his judicial decisions on dimensions outside of readability. From the above discussion, Justice West therefore appears to be a leader among his peers in terms of writing readable decisions, at least based on MADRS results that predict how human experts would assess the readability of his decisions.

---

[64] *R v Reid*, 2021 ONCA 206; *R v Vong*, 2020 ONCA 745; *R v RH*, 2021 ONCA 236; and, *R v Sillars*, 20222 ONCA 510.

[65] *R v RH*, 2021 ONCA 236, at paras 14-15. Despite narrowing the wording of this prohibition order, the Court of Appeal upheld Justice West's global sentence of 6 years imprisonment for the relevant offences.

I draw attention to Justice West's readability scores for two reasons. First, these results and his example establish how it is possible to produce judicial decisions that would likely be understandable to almost the entire population of Canada. The evidence that such readability results can be achieved without apparent compromises in different dimensions of judicial decision-making will ideally motivate other judges to strive to achieve similar results, in cases where these results would be appropriate.

Second, the results offer useful quantitative (non-impressionistic) data about a judge who appears to have unlocked some of the keys to writing readable decisions. I had never heard of Justice West before I saw his MADRS results. His biography that formed part of a news release announcing his appointment to the bench in 2008 notes that he "is the author of numerous articles on criminal law theory and practice, and has also served as a part-time Crown Attorney. Mr. Justice West taught part-time at Osgoode Hall Law School, and has been a frequent speaker at legal educational events."[66] But nothing in this biography suggests, *a priori*, that Justice West is likely to be someone who writes decisions that are more easily understood than the decisions of many other judges. If this dissertation were to be followed up with other qualitative work that explores in more detail the causative side of readability – that explores the biographical or other background factors that are likely to cause a judge to write more readable decisions – then one might begin by interviewing people like Justice West. Perhaps their perspectives on how to write readable decisions should be sought out, if we hope to take the next logical step forward in this line of research by determining not just the "who" of readability, but also the "why" and "how." With data like the data that is revealed through a judge's MADRS results, the task of identifying people who

[66] Government of Ontario, "Attorney General Announces Four Judicial Appointments" (January 18, 2008), online: <https://news.ontario.ca/en/release/87966/attorney-general-announces-four-judicial-appointments>.

we might want to interview to learn more about their processes and perspectives on readability potentially becomes easier.

4.2.3.3 Conclusion on Overall Readability Across Court Levels

The results of this first analysis that are set out above provide, for the first time, a comprehensive look at the readability levels of decisions from different levels of courts in Canada. As a general matter, all levels of courts are producing decisions that are unlikely to be understandable to millions of people in Canada. To the extent that some of these people may actually be trying, and failing, to read the court decisions, a practical problem relating to mismatches between text complexity and reader competency levels exists – although we cannot know with certainty what the scale of this problem might be.

But even if everyone who tries to read a court decision today can succeed in that task (perhaps because people with lower reading competencies self-select out of the audience for these decisions), there is still a theoretical problem that requires attention. Court decisions are the major, and usually the only, way for ordinary people to look into how our courts function. The privilege of (or responsibility for) assessing how one feels about the work that courts do, so that one can comment and act politically on this subject as one chooses, is not something that should be reserved to those within society who have more advanced levels of education or reading competencies. There is simply no principled reason why only strong readers, or more educated people, should bear a disproportionate burden of the public responsibility for keeping an eye on the business of the courts.

And, in any event, the need to interact with our court system is one that often arises for people from across all of the different education- and literacy-related tiers within our society. While it is possible that judges are already tailoring the readability levels of their court decisions to the

needs of the parties before them, the high average MADRS results for all court levels suggests that this is likely not the case. Thus, the results described above indicate that elements of the various audiences for judicial decisions, including both parties to a case and other members of the population-at-large, may be incapable of easily understanding what Canadian courts are writing in their decisions.

4.2.3.4 Formula Selection Revisited: Comparing MADRS and Other Formulas' Results

Before proceeding to a discussion of the next study that examines readability levels across different legal subject areas, I return briefly here to a more concrete discussion about the MADRS formula and its performance in comparison with other known readability formulas. In Chapter 3 (at Table 3.2), I assessed that the MADRS formula was more strongly correlated with actual expert scores for adjudicative decision text files than 6 other general-purpose readability formulas. Based on this correlation, and other research which suggested that general-purpose readability formulas may not perform well when applied to technical or domain-specific language samples,[67] I argued that the MADRS formula was most suitable for measuring the readability levels of Canadian court and tribunal decisions.

However, there are costs and potential drawbacks associated with the development and use of any new or highly specialized readability formula. To begin with, it takes significant time and energy – both from the supervising researcher(s) and from any human participants who are involved in the experimental process – to create a new readability formula. This reality is probably evident to anyone who read Chapter 3 of this dissertation, and who now understands the many steps and person-hours that went into the development of the MADRS formula. It is perhaps

[67] See generally, Jiaping Zheng & Hong Yu, "Readability Formulas and User Perceptions of Electronic Health Records Difficulty: A Corpus Study" (2017) 19:3 Journal of Medical Internet Research.

imprudent, then, for researchers to perpetually develop new readability formulas for narrow purposes – unless there are good reasons to think that these new formulas will outperform existing formulas in ways that actually have a meaningful impact on any particular set of results that the researchers are hoping to obtain. In other words, the cost of developing a new readability formula might not be worthwhile if the new formula tends to produce results that are substantially similar to the results that an existing formula would have produced, even if the new formula has improved slightly on the accuracy levels of these results.

In the case of any new specialized or domain-specific readability formula, there is an additional layer of cost-benefit analysis that ought to be considered. Such a formula will tend to have a narrower scope of application than a general-purpose formula: it will be useful in a narrower set of circumstances than a general-purpose formula that is designed to be used on any or almost any class of texts. So, even if researchers are confident that a new specialized formula would outperform existing formulas in ways that would have a meaningful impact on the results of any particular data set, these researchers should arguably still consider whether the cost of developing the new specialized formula is worth the increase in accuracy or predictive power that the formula will yield, given that the formula has a limited sphere of application. Is the specialized field of study so important that it needs its own readability formula, when substantial time and energy is involved in developing the formula?

In order to answer these questions, I compared how the results of the above study of readability levels across court levels would look, depending on one's choice of readability formula. Specifically, I calculated readability levels based on the *Flesch-Kincaid Grade Level* (FKGL),[68]

[68] J P Kincaid et al, *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel* (Memphis, TN: US Naval Technical Training Command - Research Branch, 1975).

Simple Measure of Gobbledygook (SMOG),[69] and the *New Dale-Chall* readability formula (Dale-Chall),[70] for each file within the SCC, Appeal, Superior, and Inferior court corpora. I relied on the Automatic Readability Tool for English software application to compute these scores.[71]

The average readability scores for each level of court, and each readability formula (including the MADRS formula), are shown below in Table 4.2. The relevant readability score is shown to 2 decimal places, and in the same cell, the rank for each court level is shown in parentheses for the relevant readability formula (with a rank of (1) indicating the highest or least readable score, and a rank of (4) indicating the lowest or most readable score, using that particular formula).

Table 4.2 – Average Readability Scores by Court Level and Readability Formula

Court Level	**MADRS**	**FKGL**	**SMOG**	**Dale-Chall**
SCC	5.40 (1)	11.43 (3)	13.21 (3)	10.96 (1)
Appeal	5.16 (2)	12.16 (1)	13.54 (2)	10.66 (3)
Superior	4.84 (3)	12.11 (2)	13.60 (1)	10.68 (2)
Inferior	4.37 (4)	10.73 (4)	12.86 (4)	9.98 (4)

As the results within this table show, each readability formula produces a different rank order for the average readability scores of each court level. Interestingly, all 4 formulas show that inferior courts produce the most readable decisions; however, the ranks for other court levels differ depending on the formula that one applies. This information, alone, suggests that readability results

[69] G Harry McLaughlin, "SMOG Grading – A New Readability Formula" (1969) 12 Journal of Reading 639.
[70] Jeanne S Chall & Edgar Dale. *Readability Revisited: The new Dale-Chall Readability Formula*. (Cambridge, MA: Brookline Books, 1995).
[71] "ARTE: Automatic Readability Tool for English", online: <https://www.linguisticanalysistools.org/arte.html>.

will be meaningfully different (not just slightly more or less accurate) depending on the readability formula that is used to compute the results.

The results within this table, however, also suggest that the 3 general-purpose readability formulas also produce inaccurate results. Within the "formula validation" subset of 125 files that were not involved in the process of creating the MADRS formula, and that from part of the gold standard database of expert-scored texts,[72] the average score assigned by human experts to these adjudicative decision texts was 4.6. This score means that a reader would need more than a completed high school level of education (that is, more than grade 12 education) in order to easily understand the average text file within the gold-standard database that was described in Chapter 3. More specifically, a score of 4.6 means that a reader would need more than "some post-secondary education" (4.0), but less than "an undergraduate degree, or higher level of non-legal education" (5.0), in order to easily understand the average file within the gold standard database. Although it is difficult to equate this 4.6 score with a numeric grade level, one might conceptualize a score of 4.6 as being approximately equal to a grade level of grade 14: that is, approximately 2 years of post-secondary education. None of the general-purpose formulas in Table 4.2, above, report average scores for any court level that are at, or higher, than grade 14. This fact suggests that the general-purpose readability formulas are, on average, under-reporting the reading difficulty levels of the adjudicative decision files in the court level corpora.

It might also be helpful to look at score distributions using different formulas, not just at the average scores using these formulas. The 125-file "formula validation" subset of files within the gold standard database included 9 of 125 files (7.2%) that had an average expert score of 6.0 or higher – indicating that these files could only easily be understood by someone with a law

[72] See above, at section 3.3.

degree. In Canadian common law jurisdictions, a graduate of a law program overwhelmingly tends to have completed a 4-year undergraduate degree prior to starting their law degree. In other words, an average expert score of 6.0 would be conceptually analogous to a grade level score of approximately grade 19 (that is, 12 grades of primary and secondary school, plus 4 "grades" of an undergraduate degree, plus 3 "grades" of a law degree).

Additionally, the same "formula validation" subset within the gold standard database included 57 of 125 files (45.6%) with an average expert score of 5.0 or higher – indicating that the files could be easily understood by someone with an undergraduate degree. In Canada, an undergraduate degree typically represents at least 3 years of education. So, an average expert score of 5.0 would be conceptually analogous to a grade level score of at least grade 15 (that is, 12 grades of primary and secondary school, plus 3 "grades" of an undergraduate degree).

Based on this distribution of scores within the gold standard database, one would expect to see something close to **7.2%** of the files from the different court level corpora that were scored using 3 general-purpose (grade level) readability formulas at the grade 19 level or higher, and something close to **45.6%** of these files at the grade 15 level or higher. The actual number and percentage of the 625 files within the 4 corpora of decisions from different court levels that scored greater than or equal to grades 15 and 19, respectively, are shown below in Table 4.3.

Table 4.3 – General-Purpose Readability Formula Scores Greater than Grades 15 and 19

	FKGL ≥ 15	**SMOG ≥ 15**	**MADRS ≥ 5**	**FKGL ≥ 19**	**SMOG ≥ 19**	**MADRS ≥ 6**
# of files	19	74	286	2	1	10
% of files	3.04%	11.84%	45.76%	0.30%	0.16%	1.60%

None of the 625 files scored higher than grade 14 using the Dale-Chall readability formula, so this formula is not represented at all in Table 4.3. The data in Table 4.3 suggests the 3 general-purpose readability formulas that have been compared in this section to the MADRS formula for scoring the readability of decisions based on court level all seriously under-report the grade levels of decisions at the higher end of the reading complexity spectrum. The distribution of scores greater than or equal to 6.0 (conceptually analogous to grade 19) in the gold standard database was 7.2%, but scores greater than or equal to 19 account for less than 0.3% of both the FKGL and SMOG formulas' scores (and none of the Dale-Chall formula's scores) when these formulas are applied to the 625 files in the different court level corpora. To be fair, the MADRS formula also seems to under-report the extent of scores that one would expect to see at the 6.0 / grade 19 level, but the MADRS formula – by a large margin – comes closest to matching the expected gold-standard percentage of scores at this level.

Similarly, the distribution of gold standard scores greater than or equal to 5.0 (conceptually analogous to grade 15) was 45.6%, but scores greater than or equal to grade 15 account for less than 12% of the results when any of the above 3 general-purpose readability formulas are applied to the 625 files in the court level corpora. In contrast, the MADRS formula matches this gold standard percentage of expected scores that are at or higher than the 5.0 / grade 15 level, almost exactly: the MADRS results over-report the expected amount by a very small 0.16%.

Of the 3 general-purpose formulas, the SMOG formula comes closest to achieving the kinds of scores that are contained in the gold standard database. But even this formula underestimates the reading difficulty levels of court decisions at all court levels in comparison with the reading difficulty levels of the average texts within the gold standard database, and in comparison with the more difficult texts within the gold standard database. If one accepts that the

gold standard database is the most relevant and accurate human-generated measurement standard for assessing the reading difficulty level of adjudicative decision text files, then even the SMOG formula appears to be less accurate than the MADRS formula, and produces results that are materially different from the results produced by the MADRS formula, when applied to databases of adjudicative decisions from different levels of Canadian courts. And since the gold standard database within Chapter 3 of this dissertation is the only readability-related database in existence that relies on quantitative and expert human judgements about how well readers will understand the English-language texts of Canadian court and tribunal decisions, there are no obvious reasons why one would not accept the reliability and validity of this database for the purposes of generating and assessing the performance of readability formulas.

As the above discussion suggests, the work that has gone into creating a domain-specific readability formula that can be used on Canadian court and tribunal decisions was worthwhile. The formula generates results that are materially different from the results that some of the most common and accessible general-purpose readability formulas generate. Furthermore, the results that these general-purpose formulas produce are not as accurate as the results that the MADRS formula produces when compared to the average human expert score for an independent data set of adjudicative decision text files. In other words, this is not a case where a researcher could have hoped to use an existing readability formula in order to achieve substantially similar results to those that MADRS has produced. The peculiarities of law's unique language seem to demand a unique readability formula for the purpose of reliably and accurately measuring the readability of Canadian court and tribunal decisions.

4.3 Readability Scores Across Different Legal Subject Areas

The previous study demonstrated that inferior courts, which have relatively narrow jurisdiction over criminal law – and in some locations, also family law and/or small claims – tend to produce decisions that are substantially more readable than decisions from other levels of courts. It may be the case that judges of inferior courts are simply better at writing readable decisions than other judges. However, it is also possible that the legal subject area of any particular decision may be related to the readability level of the decision. For instance, if family law and criminal law decisions are generally easier to read (perhaps because these subjects are inherently less complex, or perhaps because judges take more pains to communicate their decisions readably when dealing with these subjects), then it would stand to reason that inferior courts have lower average MADRS results than other levels of courts that deal more frequently with subjects beyond just family law and criminal law.

There are other reasons why it might be useful to understand how readable decisions are within different classes of legal subject areas. We know that in cases involving family law, self-represented litigants are estimated to be involved in up to 80% of cases.[73] Under these circumstances, one might expect family law decisions to be generally more readable than decisions in other areas of law where parties are more frequently represented by lawyers. Additionally, judges seem to think that criminal cases are of broad interest to the public, and these judges suggest that they write their criminal law decisions in part so that members of the public can understand the decisions[74] – so one might also expect these decisions to be more readable when compared to

---

[73] Julie Macfarlane, *The National Self-Represented Litigants Project: Identifying and Meeting the Needs of Self-Represented Litigants – Final Report* (May 2013), online: <https://representingyourselfcanada.com/wp-content/uploads/2016/09/srlreportfinal.pdf>, at 15.

[74] Canada, Department of Justice. *The Honourable Warren B Milman's Questionnaire (Questionnaire for the British Columbia Supreme Court Judicial Appointment Process)* (3 Nov 2017), online:

other legal subject areas. In other areas of law, however, the expectation for readable decisions might be reduced. In tax law cases, for instance, the legal subject matter may be more complex[75] so as to defy explanations using simple language within judicial decisions. For years, there have been calls for tax simplification in Canada, although "neither federal nor provincial governments have reacted to the various pleas for simplicity – or complaints about complexity."[76] If the legislation that governs a particular subject is complex and difficult to understand, then perhaps the case law about that subject will be similarly difficult to understand.

In order to assess how Canadian court decisions might differ in their readability levels across various subject areas, I used the MADRS formula to analyze readability levels for samples of court decisions from each of the above three legal subject areas: family law, criminal law, and

---

<https://www.canada.ca/en/department-justice/news/2017/11/the_honourable_warrenbmilmansquestionnaire.html>: "The public has, in that sense, an interest in the disposition of every case that comes before the courts. **This is especially true in criminal cases, where the public is also, in effect, the initiating litigant. The denunciatory element in criminal judgments is, by definition, directed specifically at the public**" (emphasis added). See also, Canada, Department of Justice. *The Honourable Cynthia Petersen's Questionnaire (Questionnaire for the Ontario Superior Court of Justice Judicial Appointment Process)* (last modified on 30 May 2022), online: <https://www.canada.ca/en/department-justice/news/2022/05/the-honourable-cynthia-petersens-questionnaire.html>: "In matters of criminal law, the main audience of the court's decision typically consists of the accused and the victim (and/or victim's family), though in many cases the public at large will also have a substantial interest in either the outcome and/or reasoning of the court, and may therefore also be part of the primary audience". And see, Canada, Department of Justice. *The Honourable Anne Louise Kirker's Questionnaire (Questionnaire for the Alberta Court of Queen's Bench Judicial Appointment Process)* (13 Nov 2018), online: <https://www.canada.ca/en/department-justice/news/2018/11/the-honourable-anne-louise-kirkers-questionnaire.html>:

> The reasons for decisions in criminal cases must be understood by the accused and those affected by the crime, but are often of interest to the media and the public at large. In this area of law, perhaps more than in others, court proceedings and decisions are the subject of media reports and social media commentary, with people sharing their thoughts and perspectives about the court process and outcomes. With social media revolutionizing the way people communicate and understand the world around them comes an increasingly wide audience of people from all walks of life attempting to comprehend what the courts are doing in some cases […].

[75] Tamer Budak, Simon James & Adrian Sawyer, "The Complexity of Tax Simplification: Experiences From Around the World" in Tamer Budak, Simon James & Adrian Sawyer (eds), *The Complexity of Tax Simplification: Experiences From Around the World* (New York: Palgrave Macmillan, 2016), 1 at 3, suggest that the following factors all contribute to the complexity of income tax legislation: "the number of exemptions, the length and readability of the legislation, the number of times it has been changed, the number of taxpayers and their average ability and the risk of tax avoidance."

[76] François Vaillancourt & Richard Bird, "Tax Simplification in Canada: A Journey Not Yet Mapped" in Tamer Budak, Simon James & Adrian Sawyer (eds), *The Complexity of Tax Simplification: Experiences From Around the World* (New York: Palgrave Macmillan, 2016), 70 at 73.

tax law. In the following sections of this dissertation, I describe my methods for conducting this analysis, and discuss the readability results.

4.3.1 Method – Assessing Readability Scores for Family, Criminal, and Tax Law Decisions

My research method for this study substantially resembled my method for the previous study that looked at readability variances across court levels. I began by assembling three separate corpora of judicial decisions/opinions – one each for family law, criminal law, and tax law cases. For each of these subject areas, the full population of cases from 2022 was large enough to merit the use of a random-sampling approach.[77] I again used a 95% confidence interval, and a 7% margin of error. Using these two parameters, and the number of total cases within the population of 2022 decisions from each subject area,[78] I then calculated the required sample size using a free online statistical calculator.[79]

My random sample sizes were N=177 (family law); N=184 (criminal law); and, N=106 (tax law). Statistically, one can infer that results from these sample sizes would be representative of results (within + / - 7%) for the full population of cases from which the samples are drawn, 19 times out of 20.

In terms of case selection, I applied the following selection approach when building each of corpus of decision files. I searched for cases using the CanLII database. For all three subject areas, I began with a full text search of the database for documents containing the word "a" – a word that I expect would appear in all cases. I then filtered the search period to only results from 2022. I then filtered to only "cases" (excluding secondary sources and legislation), from "all

---

[77] Dattalo, *Determining Sample Size*, *supra* note 5 at 8.
[78] The total populations consisted of 1814 family law, 2788 criminal law, and 227 tax law decisions, from which my smaller random samples were respectively drawn.
[79] Online: <https://www.qualtrics.com/blog/calculating-sample-size/>.

courts" (excluding tribunals). I manually selected all jurisdictions except for Quebec. For the family law and tax law cases, I used CanLII's subject filter, set to "family" and "taxation", respectively. For criminal law cases, I used CanLII's subject filter, set to "criminal or statutory infractions." However, since I was primarily interested in true criminal cases, but not in cases involving regulatory offences, I added the phrase "criminal code" as an additional term within my full text search.

All other exclusion criteria that applied to cases (described above) for my first study of readability scores across court levels were also applied here, for all of the same reasons. Randomization was also achieved in the same manner as in the previous study. Text was extracted from cases, inserted into files, and linguistically analyzed all in the same way as for the previous study. Following these steps allowed me to compute MADRS values for decisions/opinions from each of the 3 legal subject areas. Spreadsheets containing the full population list of cases for each subject area, and the randomly selected cases for each subject area, are publicly available.[80]

4.3.2 Discussion & Analysis – Readability Scores for Family, Criminal, & Tax Law Decisions

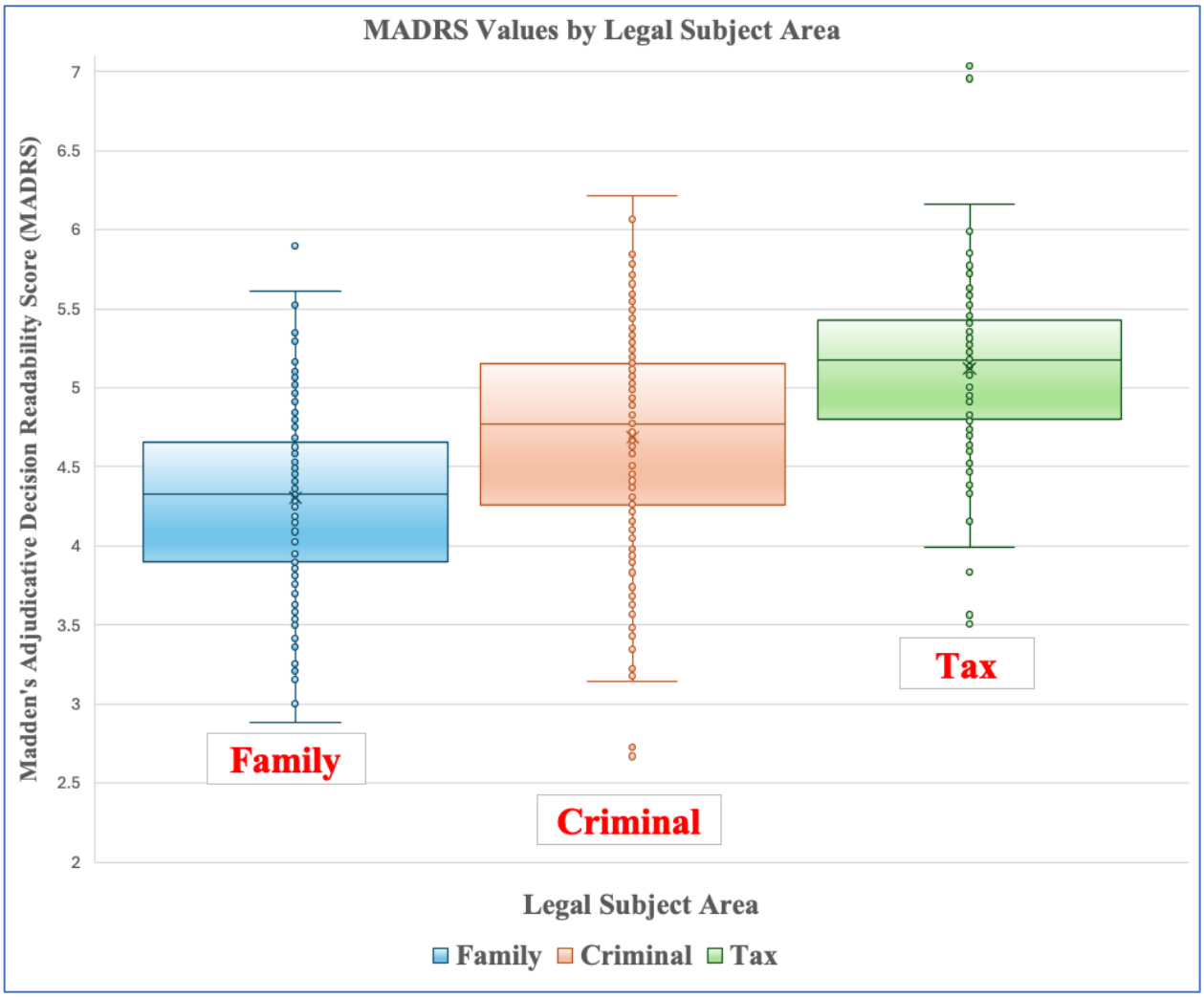
In total, I calculated results for 467 decisions/opinions, containing over 2.8 million words of text. A workbook containing the full results for decisions in each corpus (separated by tabs), including component scores for each contributing variable, is publicly available.[81] A graphical representation of MADRS results by legal subject area is shown below in Chart 4.4.

---

[80] Mike Madden, "MADRS_MaddenDissertation_Chapter4_CrimLawSampled+Population" (2025), online: <https://doi.org/10.7910/DVN/PJPYER>, Harvard Dataverse, V1 (Criminal Law Corpus). Mike Madden, "MADRS_MaddenDissertation_Chapter4_FamLawSampled+Population" (2025), online: <https://doi.org/10.7910/DVN/MUQPOE>, Harvard Dataverse, V1 (Family Law Corpus). Mike Madden, "MADRS_MaddenDissertation_Chapter4_TaxLawSampled+Population" (2025), online: <https://doi.org/10.7910/DVN/PRKKRM>, Harvard Dataverse, V1 (Tax Law Corpus).

[81] Mike Madden, "MADRS_MaddenDissertation_Chapter4_FamilyCrimTaxResults" (2025), online: <https://doi.org/10.7910/DVN/N5OIQD>, Harvard Dataverse, V1.

Chart 4.4 – MADRS Values by Legal Subject Area



The results on this chart show that family law decisions tend to be more readable than criminal law decisions, which in turn are more readable than tax law decisions. The average scores for each subject area were the following: family law – 4.30; criminal law – 4.69; and, tax law – 5.13. However, there are two outlier tax law decisions that fall close to the 7.0 level on the MADRS scale.[82] Both of these decisions have values for the variable *Dependents per nominal (no pronouns,*

[82] *Soulliere v Canada*, 2022 FCA 126, which scored 7.02; and, *National R&D Inc v Canada*, 2022 FCA 72, which scored 6.96.

standard deviation) / *Nominal_Deps_NN_StDev* that are unusually high, and that fall outside of the range for this variable that existed within the files that were used to derive the MADRS formula.[83] As I noted previously, the use of regression formulas (like the MADRS formula) on data that falls outside of the scope of the formula's original data is called extrapolation, and a regression formula's results may be biased or imprecise when extrapolating.[84] Consequently, MADRS results for these two decisions are not necessarily valid predictions of the scores that expert human readers would assign to the decisions.

If these two decisions are excluded from the average tax decision readability score, then the average score for decisions within the "tax" legal subject area falls to 5.09 – which is still a substantially higher average than what can be seen for both family and criminal law decisions. Having reviewed the texts of these two outlier decisions, I noticed that they both quoted extensively from legislation that used relatively deep and complex syntactic structures.[85] I attribute the uncharacteristically high scores for the variable *Nominal_Deps_NN_StDev* in these two outlier decisions to the high density of statutory excerpts in each decision.

---

[83] These two decisions had values for the variable of 2.91 and 1.96, respectively, whereas the files that were used to derive the MADRS formula had ranges for the variable value of between 0.83 and 1.75.

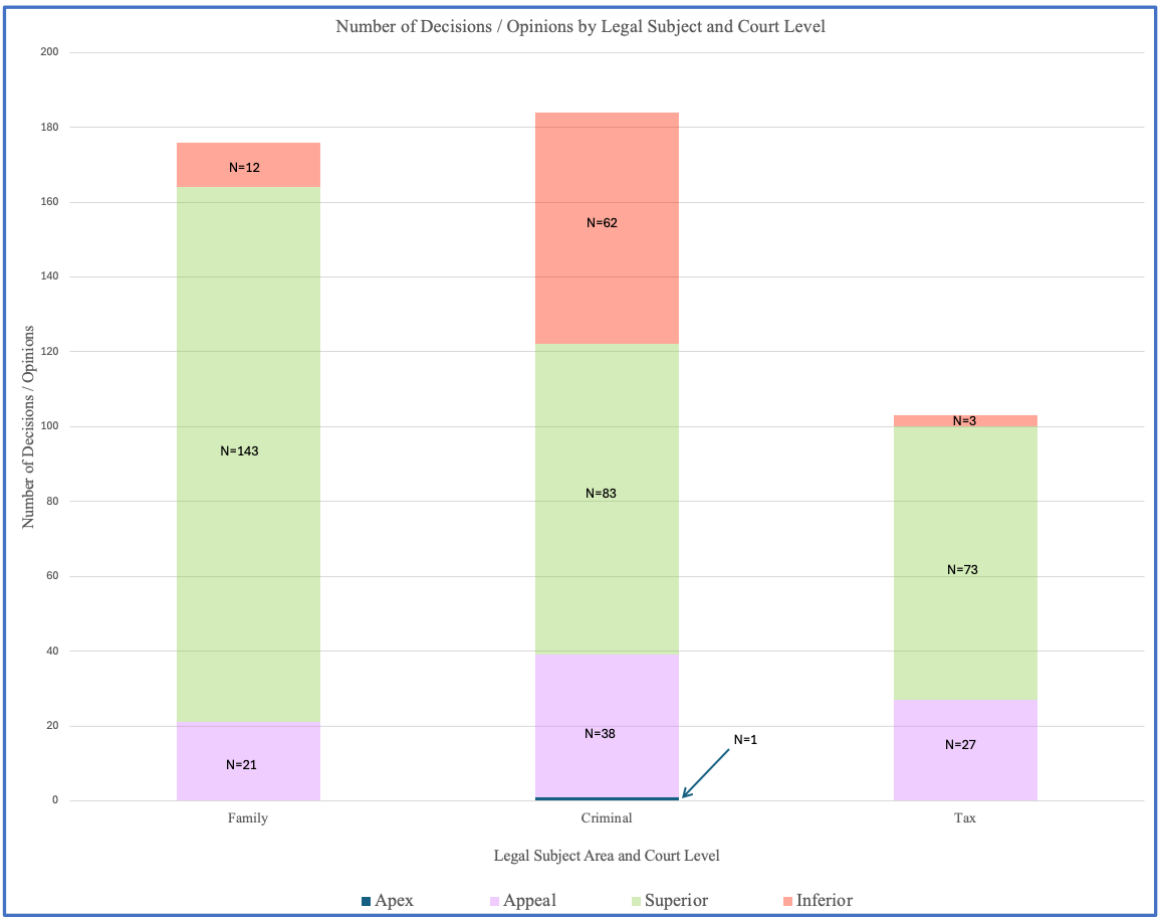
[84] *See, generally*, Gerald J Hahn, "The Hazards of Extrapolation in Regression Analysis" (1977) 9 Journal of Quarterly Technology 159.

[85] Regarding the concept of **syntactic complexity**, *see* Tim R Samples, Katherine Ireland & Caroline Kraczon, "TL;DR: The Law and Linguistics of Social Platform Terms-of-Use" (2024) 39 Berkeley Technology Law Journal 47 at 75: "syntactic complexity reflects the range and complexity of language forms in a given text. Compound sentences, embedded structures, and modifying clauses, for instance, make sentences more syntactically complex." Regarding the concept of **syntactic depth**, *see* Hélène Delage & Ulrich Hans Frauenfelder, "Syntax and Working Memory in Typically-Developing Children: Focus on Syntactic Complexity" (2019) 10:2 Language, Interaction and Acquisition 141 at 154 (illustrating how clauses can be embedded in successively deeper layers within sentences, making the meaning of the sentence more difficult to discern). Delage & Frauenfelder (at 154) offer the following three examples of 1-, 2-, and 3-level embedded clauses, respectively:

1. The teacher is looking at the boy [who is reading a Christmas book]
2. He thinks [that his son likes the teacher [who gives good marks]]
3. He believes [that she says [that the boy hates the girl [who is crying]]]

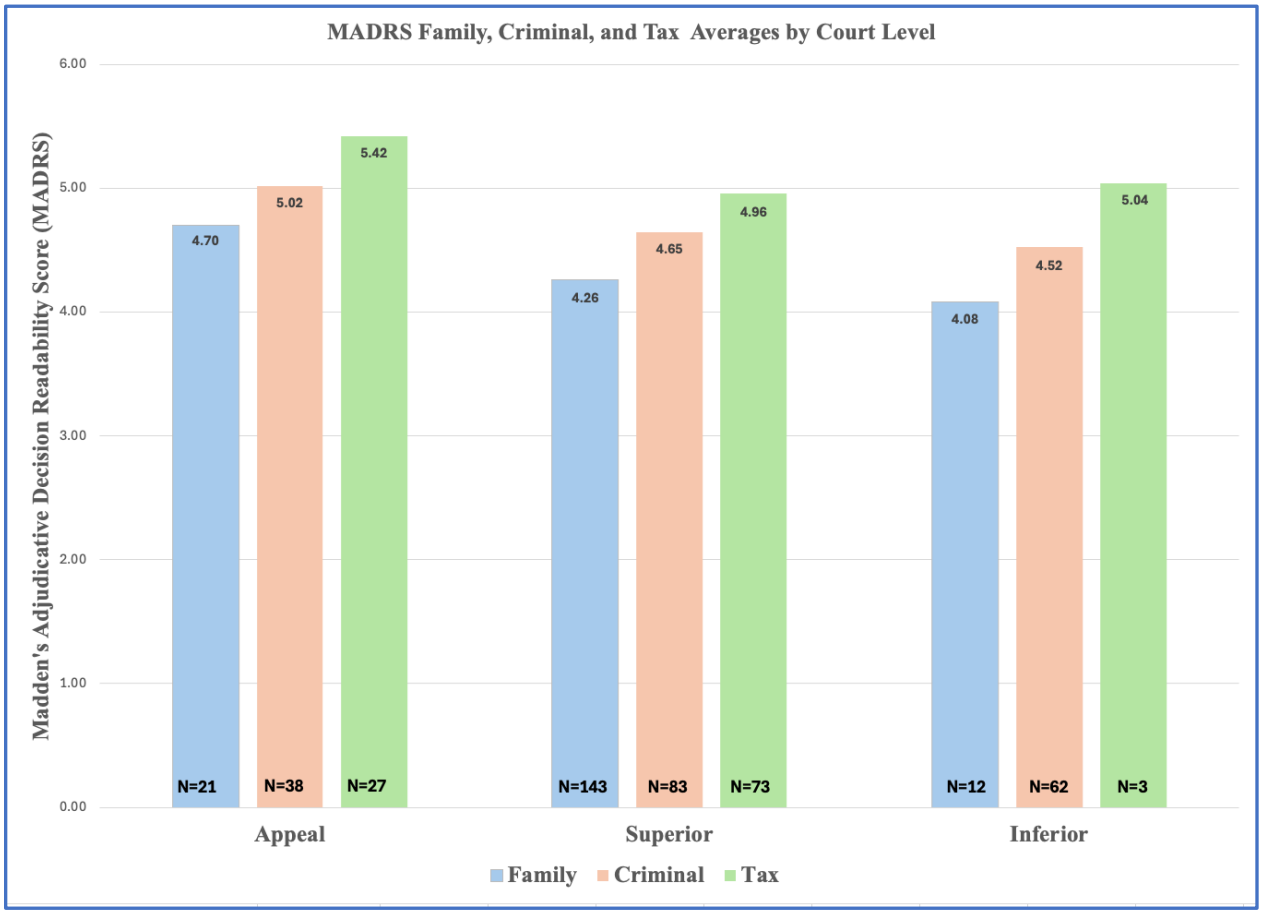
Although the above chart and MADRS average results show readability levels for family, criminal, and tax law decisions, the cases that contribute to these results were drawn from all levels of court. That is, the process of randomly selecting decisions from the populations of **all** decisions within each subject area naturally leads to the inclusion of cases from each different level of court that writes decisions on that subject (usually on a proportionate basis). So, the family law sample contained 21 appeal court, 143 superior court, and 12 inferior court decisions. The criminal law sample contained 1 apex court, 38 appeal court, 83 superior court, and 62 inferior court decisions. The tax law sample (excluding the two outlier cases mentioned above) contained 27 appeal court, 73 superior court, and 3 inferior court decisions. The breakdown of decisions by court level and subject area is shown graphically, below, in Chart 4.5.

Chart 4.5 – Number of Decisions / Opinions By Legal Subject and Court Level



Knowing from the previous analysis that decisions tend to have different readability levels depending on the level of court that produces the decisions, it could be helpful to think about how family, criminal, and tax law decision readability levels vary from one another in general, but also about how they vary from one another at each level of court. In Chart 4.6, below, the average MADRS value for decisions on each legal subject, at each court level, are shown graphically.

Chart 4.6 – MADRS Family, Criminal, and Tax Averages by Court Level



Two different trends are arguably visible within this chart. First, within each court level, the hierarchy of most readable (lowest score) to least readable (highest score) across subject areas is identical: family law is most readable, then criminal, then tax law. Second, for each individual subject area, the hierarchy of most readable to least readable court level is identical – with one exception. That is, for both family law and criminal law, inferior courts are the most readable, then

superior courts, then appeal courts. For tax law, there is a subtle variation of the pattern. Appeal courts are the least readable. But inferior courts occupy the middle ground in tax law cases, where their decisions are slightly less readable than those of superior courts. There was a very small sample size of these inferior court tax law decisions (only 3), so it would perhaps be unwise to generalize about the population of inferior court tax law decisions on the basis of this small sample.

The trends that manifest in Chart 4.6 reinforce my conclusion from the previous analysis about the relationship between MADRS values and court levels (higher MADRS values are associated with higher court levels, and vice-versa), even when controlling for different legal subject areas. And, if one excludes the 3 inferior court tax law decisions from consideration, then the trends also support my conclusion that family law decisions are more readable than criminal ones, which are more readable than tax law decisions – even when controlling for the level of court that produces the decisions.

Setting aside how family, criminal, and tax law decisions score from a relative perspective, it is also worth asking whether these classes of decisions are as readable as we might expect or want them to be. Family law decisions are the most readable of any subgroup that has been studied so far – and as we shall see from the results of subsequent subgroups that are assessed later in this dissertation, these family law decisions are the second-most readable of any class for which results are reported in this dissertation. However, 123 of 177 family law decisions scored greater than 4 on the MADRS scale (indicating that at least some post-secondary education would be needed to understand the decisions), and only one decision scored lower than 3 on the scale.[86] Since judges seem to be aware of both the high proportion of self-represented litigants who appear in courts on

[86] *Abdulmalik v Cassidy*, 2022 ONSC 4730, a decision wherein the applicant mother sought the court's permission to take her children to Ethiopia for summer vacation, against the respondent father's wishes. This superior court decision has a MADRS score of 2.88.

family law matters, and the tendency for self-represented litigants to do their own legal caselaw research, the average MADRS value for family law decisions is perhaps unnecessarily high. In other words, while judges seem to be doing a better job of writing at a level that larger groups of people will understand in family law cases when compared to other areas of law, there is still room for improvement.

To be clear, my suggestion here is not that every family law decision needs to be written at an equally easy readability level. Consider, for instance, a case at a trial court where both parties are highly educated and represented by counsel, where the law is already settled and well-stated within other previous decisions, and where there are no other public interest factors within the case that would attract a wide audience for the court's decision. In such cases, even though the decisions would deal with family law, there is not necessarily a good reason for judges to spend extra time and energy in making their decisions more readable than the parties require them to be. The parties are likely the only audiences for the decisions, so meeting their readability needs is the only real concern. The difficulty with this hypothetical scenario is that it is not necessarily easy to predict in advance when a case will attract only a limited audience, so a judge would be taking a risk of potentially excluding interested people from being able to read a decision if the judge assumed at the outset that the audience for the decision was narrow – and was wrong in that assumption.

But surely there are adults in Canada whose education level consists of only a completed secondary school diploma (or less), but who nonetheless still need to read court decisions about family law (e.g.: because they are contemplating divorce, or want greater access to their children after divorce, but they cannot afford to hire a lawyer to represent them). It seems highly unlikely that people in this situation would only need to understand 1 of 177 family law decisions, or 0.5% of all family law decisions from 2022 (assuming that my random sample is representative of the

population of family law decisions from that year), that are actually written at a level that these people are capable of understanding.

In other words, while not every family law decision necessarily needs to be written so that a person with only a high school education can understand the decision, it would be ideal – at least from the perspective of a self-represented family law litigant – if a sufficient number of important family law decisions were written at this level. And since it is likely often difficult to ascertain in the moment of writing the decision whether the decision will be important, it would also be ideal if judges erred on the side of writing to a larger rather than narrower audience in any case of doubt as to the probable audience size and composition. The above results from Chart 4.4 suggest that there is currently not a critical mass of family law decisions that people on the lower end of the educational or literacy spectrum could look to for guidance, since these decisions are overwhelmingly written at a post-secondary reading level.

The readability results for criminal law cases are also somewhat surprising. Criminal law is another area of law that some judges have specifically identified as being of interest to the broad public, and these judges suggest that they write their criminal law decisions so that the ordinary public can understand the decisions. And the offenders in these cases would often be expected to have low levels of education, since "there is a strong negative correlation between educational attainment and various measures of crime,"[87] and since a majority of incarcerated offenders in some places have not completed their high school education.[88] Admittedly, most criminally-accused persons in Canada seem to be represented by counsel (who would presumably have a role

[87] Lance Lochner, "Education and Crime" in Penelope Peterson, Eva Baker & Barry McGaw (eds), *International Encyclopedia of Education* (3rd ed) (London: Elsevier Science, 2010), 239 at 241.
[88] Lochner, "Education and Crime", *supra* note 87 at 241.

to play in explaining court decisions to their clients).[89] However, the MADRS average of 4.69 (between some post-secondary education and a completed undergraduate degree would be needed to easily understand the decisions) suggests that judges are not necessarily writing their criminal law decisions in ways that could be directly understood by many of the accused persons or offenders who are affected by the decisions.

Perhaps some individuals would be satisfied with having their court decision explained to them by a lawyer – in which case, the readability levels of criminal law decisions would not be a true concern for these people. But the above readability results show that the element of choice has essentially been taken away from these and other individuals with lower literacy levels: regardless of whether they would prefer to read a decision personally, this option will be out of reach for many individuals, since they will be unable to understand the decision if they try to read it without assistance.

The tax law MADRS values are perhaps not surprising, considering how tax law is generally viewed as being complex. And, if one pictures the typical tax law litigant as a wealthy person who is represented by a lawyer, and who has a high level of education – or as a corporation led by sophisticated individuals – then perhaps the tax law readability scores are appropriate. But this is not an accurate picture of every tax law litigant. Consider, for instance, the *Zazula v Canada (Attorney General)*[90] decision, which scored 5.27 on the MADRS scale – a score that is higher than the average score for tax law decisions. This decision involved **a self-represented litigant** who sought judicial review of a Canada Revenue Agency decision to tax the person's accidental over-contributions to his tax-free savings account. In other words, it was a case involving a

[89] Within my corpus of criminal law decisions, there were no cases involving a self-represented litigant.
[90] 2022 FC 1156.

personal (not corporate) taxation issue, where the applicant was not being assisted by a lawyer. In cases like this, one might expect a court to take greater pains to ensure that its decision can be easily understood by the litigant, and by others who might find themselves in similar situations, instead of writing at a level that requires a person to have somewhere between a completed undergraduate degree and a law degree in order to easily understand the decision.

The above analysis of readability differences across legal subject areas begins to paint a more nuanced picture of the readability landscape for Canadian court decisions. It helps us to see that both court level and legal subject are associated with different readability levels, but also that these associations operate independently from one another. It affirms that courts already seem to be writing more readable decisions in family law cases (where the proportion of self-represented litigants is notoriously high) than in other areas of law, but it also shows that courts are not writing particularly readable decisions in criminal law, despite their professed desire to reach ordinary people with these decisions. But the analysis reiterates the points that I made earlier when discussing readability scores across court levels: even the most readable subgroup of decisions (family law decisions, in this analysis) are still generally written in ways that make comprehension by large percentages of the population unlikely. Authors of decisions across all of these subject areas may wish to consider whether there is more that they can do to improve their decision readability levels.

4.4 Readability Scores For Courts and Tribunals – Same Subject

Adjudication within Canada's justice system does not happen exclusively in courts. At least in part because of the way in which courts are perceived to be slow, expensive, and complicated venues for dispute resolution, administrative tribunals are often conceptualized as an alternative mechanism for adjudication: "[v]irtually all tribunals are justified as desirable models of dispute

resolution because they can provide enhanced accessibility relative to courts and other alternatives. Tribunals are often justified as providing less expensive, less formal, and more specialized dispute resolution."[91] One might therefore expect decisions from administrative tribunals to be more readable – for their language to be more accessible – to their audiences than court decisions.

In the analysis that follows, I test this hypothesis within two specific contexts, while trying to control for the effects of other variables to the extent possible. Specifically, I examine readability levels for decisions of courts and tribunals that are dealing with the same legal subject matter, to see whether tribunals write their decisions more readably than courts. As I outline below, I use two case studies from the Canadian federal jurisdiction, in the fields of human rights law and social security law, respectively. Because there are federal administrative tribunals (the Canadian Human Rights Tribunal and the Social Security Tribunal of Canada) with jurisdiction to adjudicate claims and disputes in each of these subject areas in the first instance, and federal courts to deal with judicial reviews and appeals of tribunal decisions in these same matters, these case studies offer a somewhat unique opportunity to contrast the readability levels of tribunal and court decisions on the same topics.

4.4.1 Method – Assessing Readability Scores For Courts and Tribunals – Same Subject

For this analysis, I again used a research method that was substantially similar to my methods from the two previous analyses. I initially assembled 4 discrete corpora of adjudicative decisions – one for decisions from the Canadian Human Rights Tribunal (CHRT); one for reviews or appeals of these decisions at the Federal Court (FC) or Federal Court of Appeal (FCA); one for

[91] Lorne Sossin, "Designing Administrative Justice" (2017) 24:1 Windsor Yearbook of Access to Justice 87, at 94.

decisions of the Social Security Tribunal of Canada (SST); and, one for reviews or appeals of these decisions at the FC or FCA.

The number of reported SST decisions from 2022 (N=1759) was large enough to merit the use of a random-sampling approach.[92] I again used a 95% confidence interval, and a 7% margin of error. Using these two parameters, and the number of total cases within the population of SST decisions from 2022, I calculated the required sample size so that I could statistically infer characteristics of the entire population based on characteristics of my random sample. The results from this sample of SST decisions would be representative of results (within + / - 7%) for the full population of SST cases from 2022, 19 times out of 20.

The total number of cases from 2022 that met the criteria for inclusion within the other 3 corpora was small enough that I used the full populations of these cases for my analysis.[93] From a statistical perspective, this means that no inferences are required. The results described below represent the results of the full population of cases for the 3 relevant corpora, because each case in the population contributed to the results.

In terms of case selection, file preparation, and computation of results, my methods were similar to previous methods described above in this chapter. I applied the following selection approach when building each corpus of decision files, all using the CanLII database. For the SST corpus, I began by filtering to "federal" jurisdiction, then to "Social Security Tribunal of Canada" documents. I narrowed the year to "2022", and then produced a list of all 1759 cases that formed part of the results by further subdividing the time period into 3-month periods (making up the 4 quarters of 2022), so that CanLII would produce a list of less than 500 results for each sub-period.

---

[92] Dattalo, *Determining Sample Size*, *supra* note 5 at 8.
[93] The full populations for each corpus were as follows: CHRT – 44; Federal Court / Federal Court of Appeal human rights cases – 17; and, Federal Court / Federal Court of Appeal social security law cases – 25.

A spreadsheet showing the full population of decisions and those that were randomly selected for inclusion is publicly available.[94]

For the CHRT corpus, I filtered to only "federal" jurisdiction, then to "Canadian Human Rights Tribunal" documents. I narrowed the year to "2022", and then relied on the full list of 44 results.

For the Federal Court / Federal Court of Appeal Corpus dealing with social security law cases, I used a full text search for the phrase "Social Security Tribunal", selected "all cases", and then manually selected the two relevant federal courts from the "select specific courts and tribunals" field. I narrowed the year to "2022", and then used the complete list of all 25 results.

Similarly, for the Federal Court / Federal Court of Appeal Corpus dealing with human rights law cases, I began with a full text search for the phrase "Canadian Human Rights Tribunal", selected "all cases", and then manually selected the two relevant federal courts from the "select specific courts and tribunals" field. I narrowed the year to "2022", and then used the complete list of all 17 results.

All other exclusion criteria that applied to cases (described above) for my first study of readability scores across court levels were also applied here, for all of the same reasons. Randomization was also achieved (for the SST corpus) in the same manner as in the first study. Text was extracted from cases, inserted into files, and linguistically analyzed all in the same ways as for the previous studies. Following these steps allowed me to compute MADRS values for decisions of both courts and tribunals, in human rights law and social security law cases, respectively.

---

[94] Mike Madden, "MADRS_MaddenDissertation_Chapter4_SSTSampled+Population" (2025), online: <https://doi.org/10.7910/DVN/PXHE8S>, Harvard Dataverse, V1.

4.4.2 Discussion & Analysis – Readability Scores For Courts & Tribunals – Same Subject

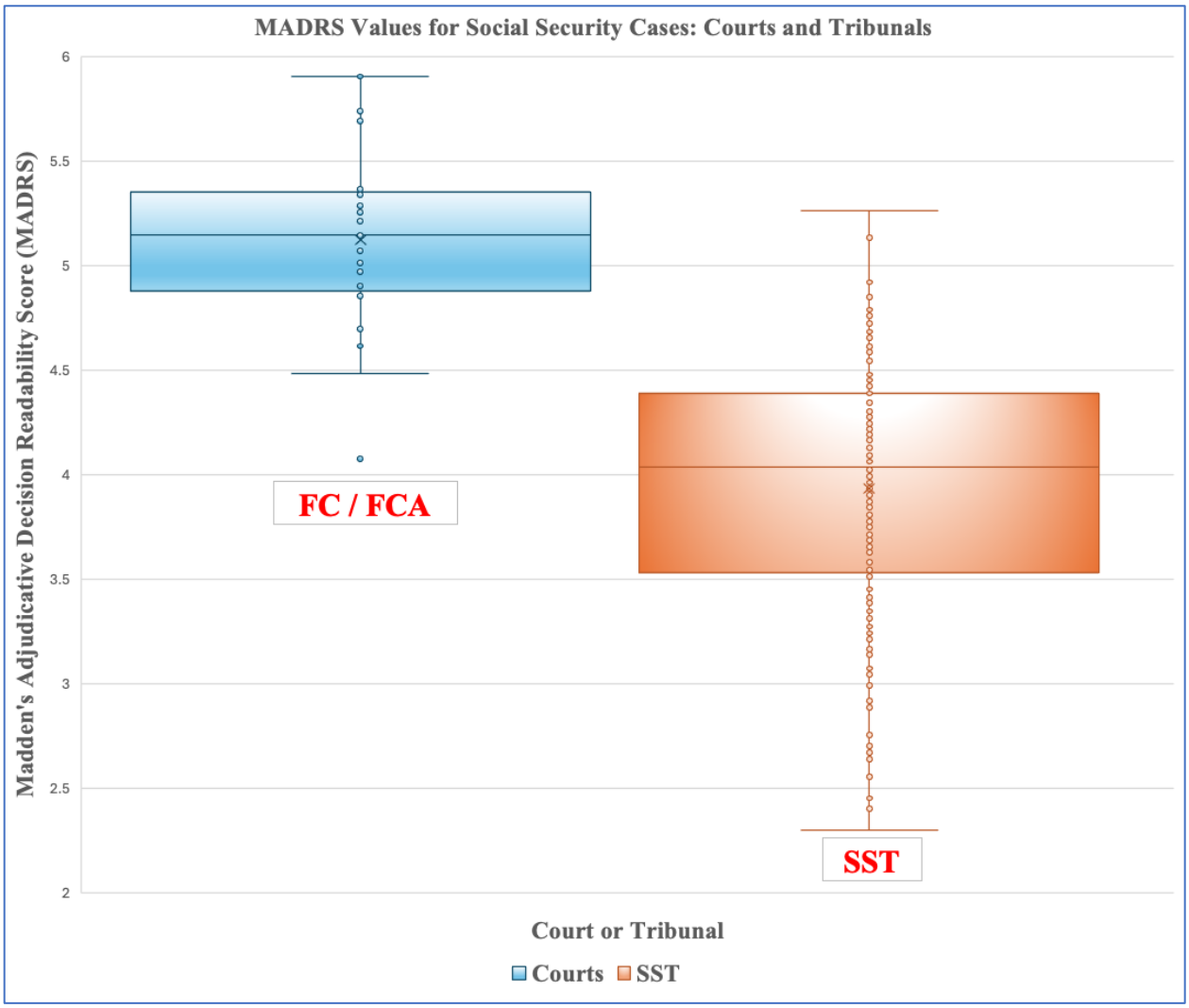
In this section, I will describe and discuss the results first for social security cases, and then do the same afterward for human rights law cases. As the results that follow will show, one cannot easily draw universal conclusions about whether tribunals write more readably than courts.

4.4.2.1 Social Security Law Decisions from Courts & Tribunals: Administrative Justice Success

A workbook containing the full results on separate spreadsheets for decisions from the Federal Court / Federal Court of Appeal, and the Social Security Tribunal, including component scores for each contributing variable, is publicly available.[95] A graphical representation of MADRS results for social security law decisions from courts and tribunals is shown below in Chart 4.7.

[95] Mike Madden, "MADRS_MaddenDissertation_Chapter4_SocSecurityResults" (2025), online: <https://doi.org/10.7910/DVN/ZQCNDL>, Harvard Dataverse, V1.

Chart 4.7 – MADRS Values for Social Security Cases: Courts and Tribunals



These results are extremely revealing. The average MADRS value for court decisions dealing with social security cases is 5.12, whereas the average for SST decisions dealing with these cases is 3.93. Looking at where these "x" marks representing average values are located on the box and whisker plots, I also note the following two observations: first, not a single court decision is as readable as the average tribunal decision (i.e.: the most readable court decision scored 4.07); and, second, with only 2 exceptions from the sample of 177 tribunal cases, every tribunal decision scored more readably than the average court decision. The two least readable SST decisions scored

5.26 and 5.13 (respectively), but every other SST decision scored less than 5, and 13 of these decisions scored less than 3 on the MADRS scale.

The most readable SST decision was *PW v Canada Employment Insurance Commission*,[96] written by Kristen Thompson. The decision considers whether a claimant had just cause for quitting her job, because she would not be entitled to employment insurance benefits if she left her job without just cause. The decision succinctly describes and applies the law of just cause, and of reasonable alternatives to quitting before concluding that the claimant did not have just cause for quitting. It scored 2.30 on the MADRS scale (a person would need some high school education to understand the decision).

The second-most readable decision was *KL v Canada Employment Insurance Commission*,[97] written by Linda Bell. It was also a case about whether a claimant had just cause for quitting a job. For anyone who believes that law is inherently complex, and cannot ever be simplified to the point that ordinary people can easily understand it, this decision offers a compelling counter-argument. For instance, the tribunal explains the law of just cause within this decision as follows:

> The law says that you are disqualified from receiving benefits if you left your job voluntarily and you didn't have just cause. Having a good reason for leaving a job isn't enough to prove just cause.
>
> The law explains what it means by "just cause." The law says that you have just cause to leave if you had no reasonable alternative to quitting your job when you did. It says that you have to consider all the circumstances.
>
> It is up to the Claimant to prove that she had just cause. She has to prove this on a balance of probabilities. This means that she has to show that it is more likely than not that her only reasonable option was to quit.

---

[96] 2022 SST 1449, online: <https://www.canlii.org/en/ca/sst/doc/2022/2022sst1449/2022sst1449.html>.
[97] 2022 SST 574, online: <https://www.canlii.org/en/ca/sst/doc/2022/2022sst574/2022sst574.html>.

> When I decide whether the Claimant had just cause, I have to look at all of the circumstances that existed at the time she quit. Then the Claimant has to show that she had no reasonable alternative to leaving at that time.[98]

As the above example shows, the decision uses familiar and casual language. It describes legal concepts with non-legal explanations. And, it uses simple sentences that do not contain many clauses. The decision scored 2.40 on the MADRS scale, and could therefore also be easily understood by someone with less than a completed secondary school education.

The two least readable decisions were both from the Federal Court,[99] and both scored 5.90 on the MADRS scale (very close to the level at which they could only easily be understood by a person who has a law degree). The language of these decisions is noticeably different from the language of the SST decision quoted above. In the *Mohammed* decision, for instance, the Federal Court stated the following about whether **the self-represented** Applicant could provide the court with evidence of a conversation that he had with a previous lawyer about his circumstances:

> There are some, albeit limited, exceptions to this rule, as enumerated in Access Copyright at paragraph 20, specifically pertaining to that which (a) providing a general background in circumstances where that information might assist the Court in understanding the issues relevant to judicial review; (b) where necessary to bring to the attention of the judicial review court procedural defects that cannot be found in the evidentiary record of the administrative decision-maker, so that the judicial review court can fulfil its role of reviewing for procedural unfairness; or (c) to highlight the complete absence of evidence before the administrative decision-maker when it made a particular finding.[100]

This paragraph-long sentence contains numerous dependent clauses, extensive legal terminology, and relatively uncommon content (e.g.: enumerated, evidentiary, pertaining)[101] and function (e.g.:

---

[98] 2022 SST 574, online: <https://www.canlii.org/en/ca/sst/doc/2022/2022sst574/2022sst574.html>, at paras 20-23.
[99] *Mohammed v Canada (Social Security Tribunal)*, 2022 FC 1124; and, *Keshinro v Canada (Attorney General)*, 2022 FC 643.
[100] *Mohammed v Canada (Social Security Tribunal)*, 2022 FC 1124, at para 19.
[101] In the COCA corpus, the word "enumerate" is ranked #14360 based on frequency of appearance in the corpus. The word "evidentiary" is ranked #24494. The word "pertaining" is ranked #10499. In terms of similar words, the verb "list" (a synonym for "enumerate" in some contexts) is ranked #2043. The preposition "about" (a synonym for "pertaining" in some contexts) is ranked #45 (online: <https://www.english-corpora.org/coca/>).

albeit)[102] words. It is generally representative of the way in which the decision, as a whole is written. And, it was rendered in a case where the Applicant was representing himself without the benefit of a lawyer to help him understand the court's decision.

As the above results and discussion indicate, the field of social security law provides an example of one where a tribunal is writing decisions much more readably than courts are writing decisions on the same topic. In one sense, this result is to be expected, given that administrative tribunals are supposed to be more accessible than courts. And courts are potentially only seeing cases that involve more complex issues that make it worth pursuing the case beyond the level of an administrative tribunal, or may be seeing more complex issues that the tribunal did not see, which might make it more challenging to write decisions in simpler language. But in another sense, the wide divergence in results between courts and tribunals is surprising: the courts are performing judicial reviews (or are sitting in appeal) of the tribunal's decisions in these cases. The courts therefore have the tribunal's decisions as central elements of their records. One might expect that these tribunal decisions would serve as inspiration for the reviewing/appeal courts to write their own decisions in similarly understandable language. After all, it would not be difficult for the courts to endorse and adopt within their own decisions the language that the tribunal has used to dispose of the cases in the first instance, at least in cases where the court is in agreement with the result that the tribunal reached.

In this sense, the above results are indicative of a partial success by the SST in producing readable decisions, and of a likely failure by the courts to consistently reach their audiences when writing decisions on social security law. Where the SST has produced decisions that, on average,

---

[102] In the COCA corpus, the word "albeit" (a conjunction) is ranked #6054 based on frequency of appearance in the corpus. The conjunction "although" is ranked #451, and the conjunction "and" is ranked #3 (online: <https://www.english-corpora.org/coca/>).

score somewhere between requiring a high school diploma and some post-secondary education in order for a reader to easily understand the decisions, the courts can almost certainly do better than producing decisions that, on average, require more than a completed undergraduate degree in order for reader to understand these decisions.

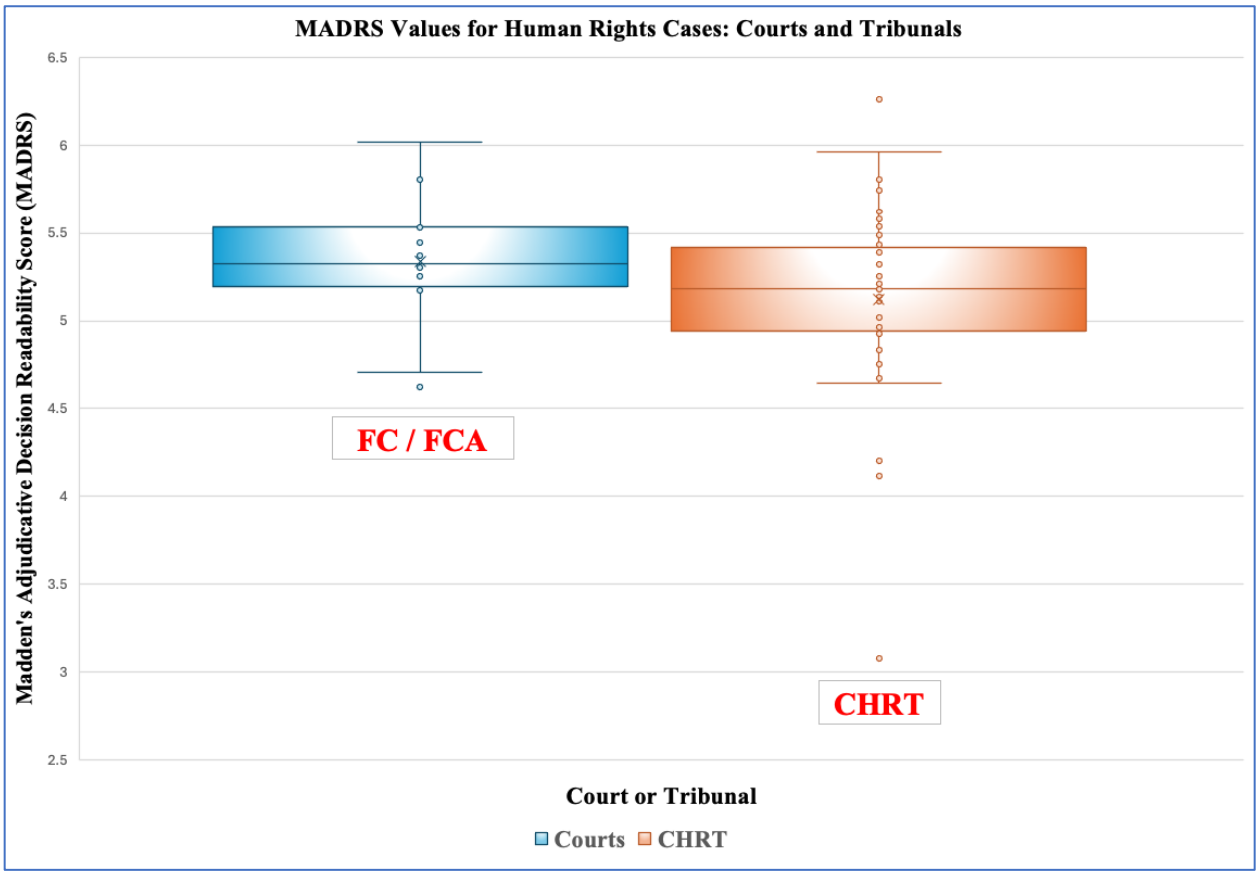
I qualify the SST's results here as only a partial success, though, because more than half of the SST decisions (i.e.: 92 of 177 decisions) have MADRS values at the 4.0 level or higher (requiring some post-secondary education to understand them). So, while many SST decision-makers are producing highly readable decisions in many cases, there still appears to be room for improvement at that tribunal in other cases. My specific recommendations for how courts, and tribunals like the SST, can leverage both the MADRS formula and the results of different studies within this dissertation to identify areas for improvement, and intervention strategies designed to bring about this improvement, are contained in Chapter 7.

4.4.2.2 Human Rights Law Decisions from Courts & Tribunals: Parallel Unreadability

A workbook containing the full results on separate spreadsheets for decisions from the Federal Court / Federal Court of Appeal, and the Canadian Human Rights Tribunal, including component scores for each contributing variable, is publicly available.[103] A graphical representation of MADRS results for human rights law decisions from courts and tribunals is shown below in Chart 4.8.

[103] Mike Madden, "MADRS_MaddenDissertation_Chapter4_HumRtsResults" (2025), online: <https://doi.org/10.7910/DVN/WEOM66>, Harvard Dataverse, V1.

Chart 4.8 – MADRS Values for Human Rights Cases: Courts and Tribunals



After exclusionary criteria were applied, there were a total of 15 court decisions and 44 tribunal decisions within the above results. The number of cases that contribute is relatively low for both courts and tribunals, but it should be recalled that these numbers represent the full populations of the relevant decisions from 2022, so they provide a complete picture of the results from that year without any need for statistical inferences.

Although the CHRT results are generally somewhat more readable than Federal Court and Federal Court of Appeal decisions dealing with human rights, the results are remarkably similar: the courts' average was 5.34, and the CHRT's average was 5.12. Additionally, we can see that both the box plots and the whiskers for these two groups align very closely (albeit at slightly lower points on the scale for the CHRT's results).

The readability levels for CHRT decisions are surprising: the tribunal's own website notes at the top of its main landing page that "[m]ost human rights complainants, and some respondents, prepare their cases without a lawyer."[104] In fact, in 2023, 35 of 66 complainants and 8 of 66 respondents were self-represented.[105] The tribunal's most recent Annual Report also specifically says, "[w]e work hard to be less formal than a court."[106] In light of this statement, and the high proportion of self-represented parties at the tribunal, one might reasonably have expected CHRT decisions to be substantially more readable than Federal Court and Federal Court of Appeal decisions on the same subject. Although the CHRT may have made many of its processes less formal than the processes of a court, there appears to still be room for improvements in the levels of formality, or reading complexity, of the tribunal's decisions.

The CHRT is a somewhat unique tribunal, in the sense that the qualifications for appointment as a member are more restrictive than the qualifications for appointment as a member of many other administrative tribunals. Subsection 48.1(2) of the *Canadian Human Rights Act* requires that "[p]ersons appointed as members of the Tribunal must have experience, expertise and interest in, and sensitivity to, human rights."[107] There is no similar restriction on the appointment of members of the Social Security Tribunal of Canada,[108] or of Ontario's Landlord and Tenant

---

[104] Canadian Human Rights Tribunal, "Welcome to the Canadian Human Rights Tribunal", online: <https://www.chrt-tcdp.gc.ca/en>.
[105] Canadian Human Rights Tribunal, *Annual Report 2023* (Minister of Public Works and Government Services, 2024), online: <https://www.chrt-tcdp.gc.ca/sites/default/files/2024-04/CHRT-AnnualReport-2023-EN.pdf>, at 10.
[106] Canadian Human Rights Tribunal, *Annual Report 2023, supra* note 105 at 5.
[107] *Canadian Human Rights Act*, RSC 1985, c H-6.
[108] *Department of Employment and Social Development Act*, SC 2005, c 34, s 45.

Board,[109] for instance, and a much more permissive rule regarding the appointment of members of the Immigration and Refugee Board.[110]

If the legal discourse surrounding human rights cases – as seen in decisions of courts and tribunals that deal with these cases – is all written at a relatively complex level, then perhaps the statutory criteria for appointment as a member of the tribunal is a barrier to the writing of more readable decisions by the tribunal. After all, if the members who are appointed to preside at the CHRT have all professionally matured in the same environment (that is, the place where they acquired their "experience" and "expertise" in human rights law) by reading the same types of case law, then it may be expecting too much of these individuals to break from the conventions of their field by writing drastically more readable decisions than their predecessors.

So, if the Government of Canada seriously wanted the CHRT to produce more readable decisions (perhaps as a means of improving access to justice) then this government might consider recommending different types of people (such as those with experience and expertise in fields like education or language) to the Governor in Council for appointment. Or – even more drastically – might need to change the statutory criteria for appointment. Perhaps having outsiders from beyond the human rights world would lead to readability improvements. Or, perhaps adding a new criterion requiring a demonstrated ability to write understandable decisions would be helpful. In any case, this CHRT readability example raises the question of whether one can influence readability results just by focusing on the language of decisions, instead of looking more broadly at the characteristics of the people who are appointed to courts and tribunals, and who therefore author the decisions.

---

[109] *Residential Tenancies Act*, 2006, SO 2006, c 17, s 169.
[110] *Immigration and Refugee Protection Act*, SC 2001, c 27, s 153(4), requires that at least 10% of the members of the Immigration Appeal Division, and the Refugee Appeal Division, "must be members of at least five years standing at the bar of a province or notaries of at least five years standing at the Chambre des notaires du Québec."

Stepping back from the above focus on the CHRT, the overall MADRS results here are unfortunate in respect of both the courts' and tribunal's decisions. Given the quasi-constitutional status of human rights law,[111] describing universal rights that inhere in 'people' generally (as opposed to, for instance, corporate law rights that often attach only to a corporation or its shareholders), one might reasonably have expected both courts and the CHRT to have better readability scores. This legal subject area is one wherein many factors (such as public interest, universality, and self-representation) all suggest that the audiences for decisions would benefit from improved readability.[112] If someone from within the large proportion of the Canadian population that has less than a completed undergraduate degree ever had a need or desire to read a Canadian human rights decision – whether as a party to a human rights proceeding, or as an interested member of the public – this individual would probably be unable to easily understand the majority of both current court and tribunal decisions on the subject. The field of human rights law, at least at the federal level, is one wherein decisions seem to be particularly unreadable, based on MADRS predictions about how these decisions would be scored by human experts.

4.5 Concluding Thoughts: Canadian Caselaw Readability Levels

Within this chapter, we have seen the results of 3 different studies that looked at the readability levels of Canadian court and tribunal decisions from different perspectives. As a

---

[111] See *Winnipeg School Division No 1 v Craton*, [1985] 2 SCR 150, at 156 (for reference to this status of Canadian human rights legislation, generally). See also *CN v Canada (Canadian Human Rights Commission)*, [1987] 1 SCR 1114, at 1335-36 (noting specifically that the *Canadian Human Rights Act* falls into the category of quasi-constitutional legislation).

[112] Many of these same considerations apply to every other quasi-constitutional Canadian statute as well, since these statutes "are all rights-based regimes that cover the traditional civil liberties, the protection against discrimination, the right to privacy, the right to access government information, language rights, and the common law protection of a person's good reputation" (John Helis, *Quasi-Constitutional Laws of Canada* (Toronto: Irwin Law, 2018) at 4). In other words, quasi-constitutional laws all seem to describe universal (statutory) rights that inhere in everyone. The same could be said of all constitutional provisions relating to individual rights or liberties, but not necessarily to constitutional provisions that address topics like the composition of the Senate or Parliament's power to create courts, among other topics.

general matter, the results are not encouraging. Across all 3 studies, a total of 1353 decisions or opinions were scored. Not a single decision scored at or below 2.0 on the MADRS scale. And, only 27 decisions (just under 2% of decisions) scored below 3.0, which is to say, they could be easily understood by a person with less than a completed high school education.

The courts and tribunals that were studied in this chapter may want to write their decisions for the Canadian people. And there are many good reasons why these adjudicative bodies would want to reach diverse audiences with their decisions, just as there are many good reasons why the Canadian people might want to read and understand the decisions. But the results discussed in this chapter clearly suggest that an overwhelming proportion of adjudicative decisions are written in ways that are too difficult for millions of Canadians to understand. There is a disconnect between aspiration and reality that is revealed in this chapter. The aspiration to write readable decisions is not matched with a practice of writing readable decisions.

I am prepared here to assume that many authors who realize their decisions are likely not reaching their audiences will strive to do something that could change this situation. This is perhaps a contentious assumption, and one for which I have no objective empirical proof. But it is a reasoned assumption, grounded in my personal (but impressionistic) experiences[113] interacting with judges over the last 15 years, as both a lawyer and scholar of law. These experiences have almost universally demonstrated to me that judges care about (1) performing their functions to a very high standard; and, (2) meeting the needs of those who are affected by their decisions. Since judges' own words have shown that many of them – across different levels of courts and jurisdictions – see their audiences broadly as including the general public, these same judges probably see the function of communicating in ways that can be understood by the general public

[113] I have referred to these experiences already once in this dissertation, above at section 4.2.3.2.

as a function that they ought (or would like) to perform. And, where judges have presented themselves (perhaps only to me, although I doubt this) as people who care about how well they perform their functions and meet the needs of those who are affected by their decisions, it does not strike me as an illogical or unsupportable assumption to conclude that these judges will do something to change the situation if they realize that they are falling short of their own expectations.

It is this assumption that motivates me to discuss several possible ways in which the research that is contained in this chapter, and the preceding chapter, might be helpful to these decision-makers and the institutions that they represent. My recommendations, that discuss how MADRS readability measurements can be used to potentially drive positive change in terms of how well adjudicative decisions are likely to be understood by their audiences, are contained in Chapter 7. But before eventually moving to these recommendations, I first take a closer look at two specific areas – youth criminal court cases, and cases involving self-represented litigants – wherein there may be a special need or justification for highly readable decisions. These studies are described below, in Chapter 5.

5. Targeted Canadian Decision Readability Studies

5.1 Overview

As I noted in the introduction to this dissertation, there is not necessarily a single and universal readability level that would be an appropriate target for all adjudicative decisions, since every decision is likely to have at least subtly different audiences, with different reading capabilities. However, there are some contexts wherein society might reasonably expect court decisions to be particularly readable. These expectations arise in cases where, regardless of who else may be interested in a decision as an audience member, a party to the case (that is, a person who judges routinely describe as a primary audience member)[1] has unique needs that affect their ability to understand a court decision.

Specifically, we might expect or want MADRS levels to be lower for decisions involving charges under the *Youth Criminal Justice Act* (YCJA), and in cases where a party to the proceedings is not represented by a lawyer. In the former class of cases, the accused person will always be someone who was under the age of 18 at the time that the alleged offence was committed,[2] and will regularly also still be under the age of 18 when the decision is rendered. Chronologically, it therefore stands to reason that individuals who are dealt with under the YCJA will often struggle to understand a decision that scores higher than 3.0 on the MADRS scale.

---

[1] Jamie WS Saunders, "Reflections on the Art and Science of Decision-Making and Decision-Writing" (Nov 2017), *Professional Development Conference: Maritime Provinces' Law Societies* (Halifax, NS), online: <https://www.courts.ns.ca/sites/default/files/editor-uploads/From%20the%20Bench/Reflections_on_the_Art_and_Science_of_Decision_Making_11_09_17.pdf>, at 71 ["Reflections"]; Canada, Department of Justice. *The Honourable Steven Wilson's Questionnaire (Questionnaire for the British Columbia Supreme Court Judicial Appointment Process)* (1 Mar 2017), online: <https://www.canada.ca/en/department-justice/news/2019/02/the-honourable-steven-wilsons-questionnaire.html> [*Steven Wilson's Questionnaire*].

[2] *Youth Criminal Justice Act*, SC 2002, c 1, s 2(1), sv: "young person" provides that a young person is someone who appears "to be twelve years old or older, but less than eighteen years old."

People in Canada (outside of Quebec) generally finish secondary school at the age of 17 or 18,[3] so most young persons within the meaning of the YCJA will not have the post-secondary educational and literacy skills that are needed to understand decisions scoring higher than 3. And, in the latter class of cases, where a party does not have the extra resource of a legal representative to help them understand the decision, we might also hope for courts to write more readable decisions – ones that can easily be understood without the need for a lawyer's assistance.

This chapter reports on two studies that are designed to explore readability levels in these situations. The first study discusses readability levels for cases involving youth criminal court decisions. The second study discusses readability levels for cases involving at least one self-represented litigant (SRL). Collectively, these studies provide us with data and insights to better understand the ways in which courts communicate their decisions when one party to the proceeding likely has relatively higher readability needs.

5.2 The Readability of Youth Criminal Court Decisions

The clearest circumstance of any studied so far within this dissertation where there is a need for highly readable decisions – if one subscribes to the idea that the parties are a primary audience for any court decision[4] – is in the case of youth criminal court decisions. The accused

---

[3] Organisation for Economic Cooperation and Development, "Education at a Glance – 2022 Indicators – Canada", online: <https://www.oecd-ilibrary.org/sites/5ea1485d-en/index.html?itemId=/content/component/5ea1485d-en>: "The average age of graduation from general upper secondary programmes varies from 17 to 21 years across OECD countries and is 18 years in Canada."

[4] See Saunders, "Reflections", *supra* note 1 at 71, and, *Steven Wilson's Questionnaire*, *supra* note 1. This point is also made in an American context in Federal Judicial Centre, *Judicial Writing Manual: A Pocket Guide for Judges*, 2nd ed (Federal Judicial Center, 2013) at 5: "Opinions are written primarily for the litigants and their lawyers, and for the lower courts or agencies whose decisions they review." *See also*, Gerald Lebovits, Alifya V Curtin & Lisa Solomon, "Ethical Judicial Opinion Writing" (2008) 21 Georgetown Journal of Legal Ethics 237 at 244 (wherein the authors note that decisions are first written for the parties: "The law forbids vigilante, or 'self help,' justice. If individuals believe they will receive unexplained outcomes in the judicial forum, reliance on self-help might become the norm" (internal citations omitted). In a UK context, *see also* Andrew Stephen Burrows, "Judgment-Writing: A Personal Perspective", *Annual Conference of Judges of the Superior Courts in Ireland* (20 May 2021), online: <https://www.supremecourt.uk/docs/judgment-writing-a-personal-perspective-lord-burrows.pdf>, at 4 (for comments of a sitting judge of the UK Supreme Court: "one's target audience must include the parties themselves, the legal

person in these cases will regularly be less than 18 years old, and therefore not someone who is likely to have acquired any post-secondary education. If courts are really taking the circumstances of their audiences into account when writing decisions, then MADRS scores for this class of cases should be somewhere near the 3.0 level. The study that follows tests this hypothesis.

5.2.1 Method – Measuring the Readability of Youth Criminal Court Decisions

My research method for this study followed the overall approach used in the previous study (described above in Chapter 4) that looked at readability variances across court levels. I built a corpus of judicial decisions that consisted of only cases involving charges under the YCJA. My initial searches into the population size for this class of cases indicated that it was small, so I used the entire population rather than a random sample (subject to exclusionary criteria, discussed below).

In terms of case selection, I applied the following selection when building the corpus of decision files. I searched for cases using the CanLII database. I began with a full text search of the database for documents containing the phrase "youth criminal justice act." I then filtered to include all provincial, territorial, and federal jurisdictions other than Quebec (for the same reasons that I provided when excluding Quebec cases from previous studies). I set the search period to only include results from 2022. I then filtered to only "cases" (excluding secondary sources and legislation). This approach generated a list of 152 results.

All other exclusion criteria that applied to cases (described above) for my first study of readability scores across court levels were also applied here, for all of the same reasons. However, I also screened each decision within the results list and excluded any case where it was apparent

---

advisers to those parties, other judges, other practising lawyers, academic lawyers and students, and last but by no means least the public at large").

that the case was not one that primarily involved the prosecution of charges against a young person. These exclusions arose where, for instance, the decision included the phrase "youth criminal justice act" for only incidental reasons, rather than because the decision was from a court that had been empowered under the YCJA. The exclusions also arose in criminal prosecutions of adults, but where the YCJA was implicated indirectly – for instance, when an adult accused person sought the production of criminal records relating to a young person who was a third party, and whose records were protected under the YCJA.

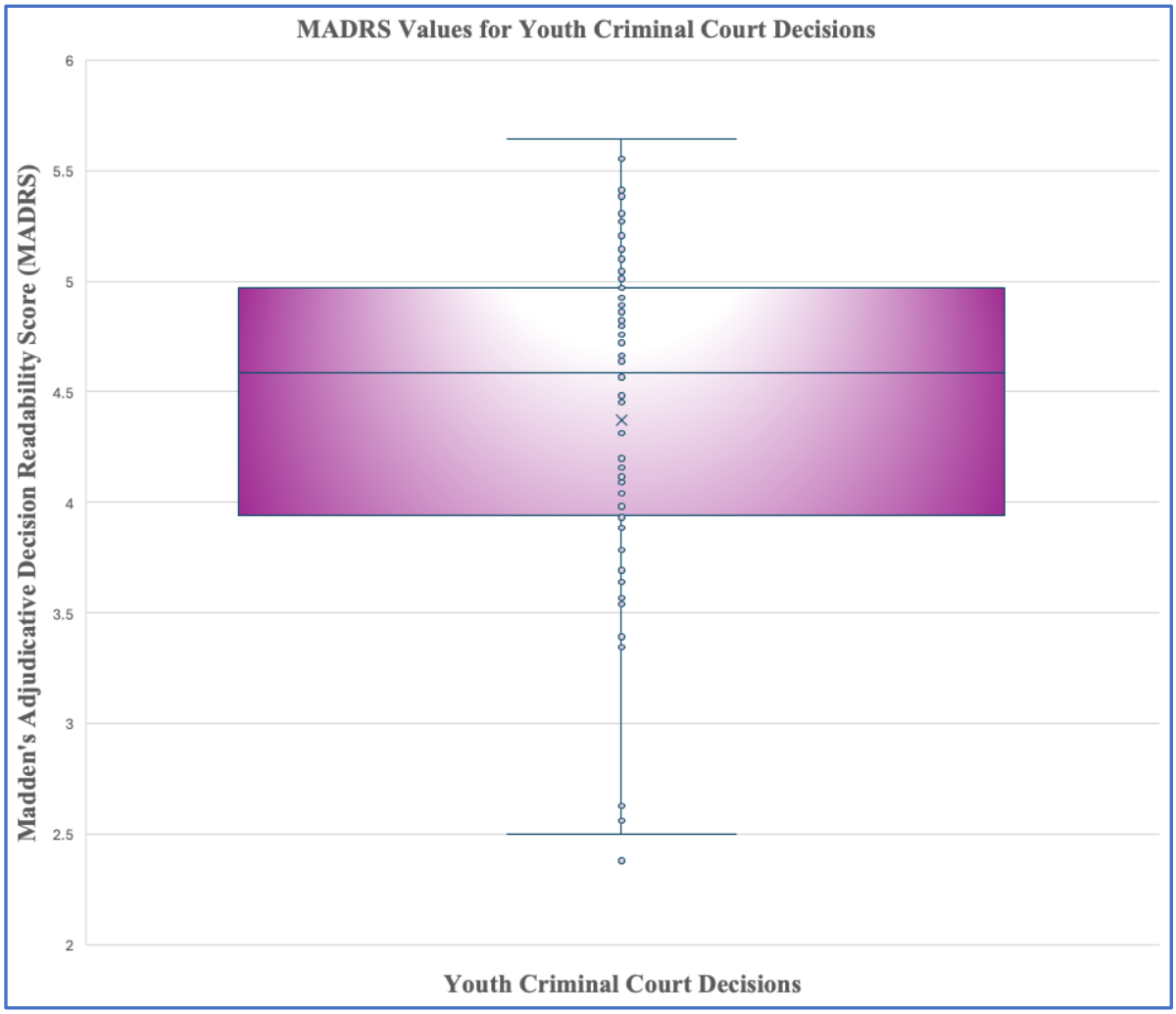
Text was extracted from cases, inserted into files, and linguistically analyzed all in the same way as for the previous studies. Following the above steps allowed me to compute MADRS values for decisions from cases involving only young persons who were prosecuted under the YCJA.

5.2.2 Discussion & Analysis – Readability Scores for Youth Criminal Court Decisions

Once all exclusion criteria were applied, I ultimately ended up with a corpus that consisted of 74 decision files, containing over 546,000 words of text. A spreadsheet containing the full MADRS results for these decisions, including component scores for each contributing variable, is publicly available.[5] A graphical representation of MADRS results for these decisions is shown below in Chart 5.1.

[5] Mike Madden, "MADRS_MaddenDissertation_Chapter5_YouthCourtResults" (2025), online: <https://doi.org/10.7910/DVN/H0DF2O>, Harvard Dataverse, V1.

Chart 5.1 – MADRS Values for Youth Criminal Court Decisions

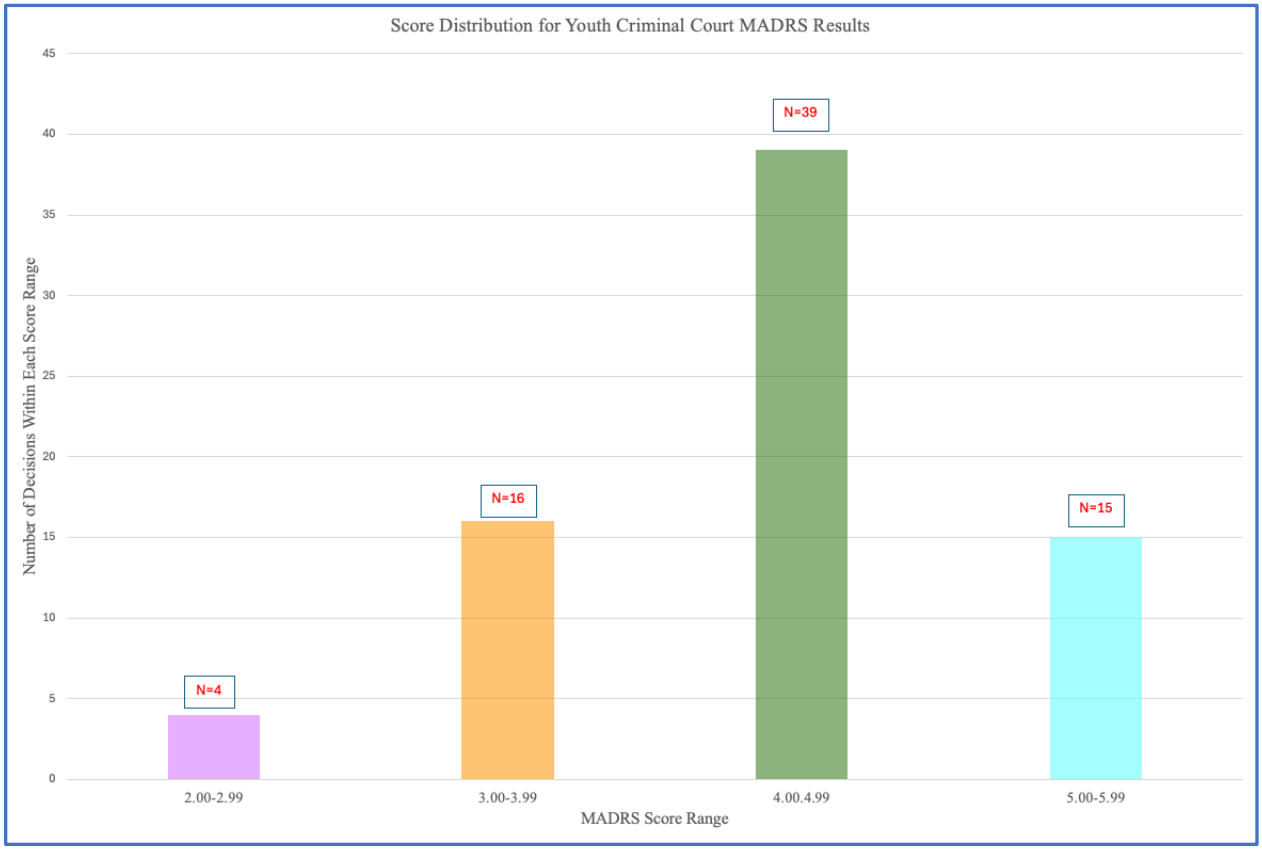


These results, unfortunately, look surprisingly similar to the results for all inferior court decisions (shown above, in Chapter 4). The MADRS average for youth criminal court decisions was 4.43, while the average for all inferior court decisions in the previous study that looked at readability variances across court levels was 4.37. In other words, inferior courts write more readable decisions in general than youth courts write in decisions that involve prosecutions of young persons under the YCJA. That being said, the average score for these YCJA decisions is noticeably more readable than the average score for all criminal law decisions (4.69) that was

reported above in Chapter 4, and this is probably a more relevant comparator for YCJA decisions, since both sets of decisions deal with criminal prosecutions. Overall, the YCJA results suggest that individuals who are subjects of the youth criminal court decisions would need, on average, somewhere between "some post-secondary education" and "a completed undergraduate degree" to understand the decisions that relate to their own cases.

A breakdown of the score distribution for these youth criminal court decisions is shown below, in Chart 5.2.

Chart 5.2 – Score Distribution for Youth Criminal Court MADRS Results



If we assume that the young persons who were the subjects of these decisions would typically not have completed more than secondary school education, then only 5.4% of the decisions (4 of 74) would be understandable to the people who are most affected by the decisions:

the accused persons or offenders in those cases. Over 20% of the decisions (15 of 74) were written at a level that requires a completed undergraduate degree to understand the decisions. In light of what we know or can assume about the education levels of young persons who are prosecuted under the YCJA,[6] there is a massive disconnect between the kind of language that these people can understand, and the kind of language that is used in the decisions that dispose of charges against these people.

The two most readable decisions were written by judges of the Alberta Provincial Court (Judge D'Souza[7] and Judge Lloyd[8]). The next two most readable decisions are from the Ontario Court of Justice,[9] and one of them was written by Justice West,[10] who was mentioned above in Chapter 4 for having written 2 of the 5 most readable decisions from any level of court in the study of readability variances across court levels. These were the only decisions in the study that required less than a completed high school education to understand.

The least readable decision, scoring 5.64 (somewhere between requiring a completed undergraduate degree and a law degree to understand the decision) was from the Ontario Superior Court of Justice.[11] In some ways, one might have expected this decision to have a far lower MADRS result. The decision arises in the context of a first-degree murder charge that also involved an alleged terrorism component. The decision considered the constitutionality of certain provisions

---

[6] In YCJA cases, the accused person will always be someone who was under the age of 18 at the time that the alleged offence was committed (because the *Youth Criminal Justice Act*, SC 2002, c 1, s 2(1), sv: "young person", provides that a young person is someone who appears "to be twelve years old or older, but less than eighteen years old."). Such individuals will also regularly also be ones who are still under the age of 18 when the decision is rendered. It would be exceedingly rare for such a person, who is under the age of 18, to have achieved much more than a completed secondary school diploma. This is because, in most parts of Canada (and in all parts of English-speaking Canada) individuals generally graduate from secondary school at the age of 17, or older.
[7] *R v RM*, 2022 ABPC 230.
[8] *R v TL*, 2022 ABPC 14.
[9] *R v SP*, 2022 ONCJ 454; and, *R v LP*, 2022 ONCJ 55.
[10] *R v LP*, 2022 ONCJ 55.
[11] *R v OS*, 2022 ONSC 3935.

of the *Criminal Code* that apply in the case, and whether these provisions violate the accused person's rights under sections 7 and 11(d) of the *Charter of Rights and Freedoms*. All of these factors arguably create a powerful rationale for the decision to be written readably: the accused is a young person; murder is one of the most serious charges under Canadian criminal law; terrorism offences are also serious, and tend to attract a significant degree of public attention; and, the case has a constitutional dimension to it that may be of broad public interest. In spite of all of these factors – any one of which might have created a reasonable demand from audiences for a highly readable decision – the Court wrote the decision in language that two-thirds of Canada's population cannot easily understand.

From all of the above results, it seems as if Canadian judges make some effort to write more readably than they otherwise might when dealing with YCJA prosecutions (as seen in the lower YCJA scores when compared to all criminal law scores). However, the level at which these YCJA decisions are written is likely not low enough to enable the subjects of the decisions to personally understand the content of the decisions.

In reality, however, all of the accused persons or offenders in the current YCJA data set were represented by lawyers. So, one might argue that the readability levels of the decisions are not problematic as long as the lawyers can understand the decisions, and can explain them to their youthful clients. This argument might be compelling if lawyers had both the time and the ability to explain decisions to their clients so that clients could understand the relevant content – as is apparently required by Canada's *Model Code of Professional Conduct* for lawyers.[12]

---

[12] Federation of Law Societies of Canada, *Model Code of Professional Conduct, as Amended April 2024*, online: <https://flsc.ca/wp-content/uploads/2024/05/2024-Model-Code-of-Professional-Conduct.pdf>, at 20 (Rule 3.1-1, associated commentary para 3: "A lawyer has a duty to communicate effectively with the client. What is effective will vary depending on the nature of the retainer, the needs and sophistication of the client and the need for the client to make fully informed decisions and provide instructions") and 27 (Rule 3.2-9: "When a client's ability to make

We do not know the extent to which lawyers spend time explaining the meaning of a decision to a client, but anecdotal information suggests that legal aid lawyers, who tend to represent young persons in proceedings under the YCJA,[13] are already underpaid or not paid for many of the hours that they work.[14] Results from the current study indicate that the average reported YCJA decision is approximately 7300 words in length. And, recent research suggests that adults read (silently) non-fiction texts at an average of between 175-300 words per minute.[15] So, a lawyer would need to spend approximately 30 minutes reading an average YCJA decision, one time. It is possible that a lawyer would need to read a decision more than once before being in a position to explain it to a client. And, it would presumably take at least as long to explain the decision to a client as it takes to read the decision silently, although the task may take even longer.[16] Under these time and financial constraints that YCJA lawyers would face, it is probably imprudent to assume that a lawyer will explain the meaning of a court decision to their client in a way that gives the client the kind of understanding that the client would have if they could read and understand the decision personally.

---

decisions is impaired because of minority or mental disability, or for some other reason, the lawyer must, as far as reasonably possible, maintain a normal lawyer and client relationship"; and associated commentary, para 1: "The key is whether the client has the ability to understand the information relative to the decision that has to be made and is able to appreciate the reasonably foreseeable consequences of the decision or lack of decision"). Taken together, these rules and their associated commentary imply that lawyers have a special obligation to explain decisions to minors in ways that they can understand.

[13] Young persons generally have the ability to obtain government-funded representation through legal aid schemes in far wider circumstances than adults: Ting Zhang & Josh Hoddenbagh, *The Costs of the Youth Criminal Justice System 2010* (Ottawa: Department of Justice, 2013), at 16.

[14] Breese Davies, "Going Broke in Legal Aid: The High Cost of Cheap Justice," *The Walrus* (Sept 2016), online: <https://thewalrus.ca/going-broke-in-legal-aid/>.

[15] Marc Brysbaert, "How Many Words do we Read per Minute? A Review and Meta-analysis of Reading Rate" (2019) 109 Journal of Memory and Language 104047, at 21.

[16] For instance, Brysbaert found that adults read aloud at 183 words per minute. Using this reading rate, it would take 40 minutes to read the average YCJA decision aloud. If a lawyer stopped at various intervals to provide a gloss on the language in the decision, or to answer a client's questions about a part of the decision, then it could conceivably take more than an hour to explain what a decision means to a client. See Brysbaert, "How Many Words do we Read per Minute? A Review and Meta-analysis of Reading Rate", *supra* note 15 at 11.

The point of this discussion is not to suggest that judges are the only, best, or most efficient individuals who should bear the responsibility for ensuring that their decisions are understandable to the parties in any given case. If lawyers are paid less than judges (which they almost certainly are, at least in the case of legal aid lawyers), but they are just as capable of explaining the meaning of court decisions to the parties, then a persuasive economic argument could be made that judges should not spend time writing more readably (since lawyers will essentially 'translate' any unreadable parts of a judge's decision to their clients in any event at a lower dollar-cost to society).

But this argument only works if lawyers can and will perform these explanatory tasks – presumably because they are compensated for the work. And there are numerous indications that lawyers are not being paid to perform this kind of work, or other similarly essential functions that would benefit low-income and high-needs clients. For instance, the Chief Justice of Ontario recently commented upon cuts to legal aid funding, and the impact of these cuts, in the following way:

> It is not for judges to determine how public funding are to be allocated. The needs of the justice system must be balanced against challenges and demands of public health, education and welfare, and a host of other necessities of modern society. But what we judges can say is that reducing legal representation for the most vulnerable members of society does not save money. It increases trial times, places greater demands on public services, and ultimately delays and increases the cost of legal proceedings for everyone. We can also say that public confidence in the administration of justice is enhanced when the most vulnerable in our society are given a voice, so they can truly be heard.[17]

As these remarks illustrate, legal aid funding is scarce, and this scarcity seems to be shifting at least some of the work that one might expect lawyers to perform onto the shoulders of other actors within the justice system. Consequently, it is not clear that the work of communicating the meaning of court decisions to clients is currently being performed by lawyers in ways that adequately

---

[17] Chief Justice George Strathy, "Opening of Courts Remarks: Opening of the Courts of Ontario for 2019" (September 10, 2019), online: <https://www.ontariocourts.ca/coa/en/archives/ocs/2019.htm>.

compensate for the inability of many parties to personally read and understand decisions. Under these circumstances, the argument for more readable decisions from judges becomes more urgent and persuasive.

Ultimately, as with the previous studies discussed in Chapter 4, the current study of YCJA decisions does not conclusively reveal how many people are actually incapable of reading decisions that they want to read. The current study also cannot tell us whether large gaps between the likely reading competency levels of YCJA accused persons/offenders, and the readability levels of court decisions that affect these people, are adequately bridged by explanations of decisions from lawyers who represent these clients. But the study does show how YCJA decisions are generally written at levels that the parties in those cases will not easily understand without some sort of external help – and this is a potential problem in its own right. Judges who purport to be writing "for the parties" do not tend to qualify their statements by saying things like, "…who will, of course, need to rely on their lawyers in order to understand the decisions." Rather, these judges appear to believe that they are writing directly for the parties, without any references to intermediaries or 'explainers' of the decisions. So the current study points to another disconnect between the judicial statements[18] about the (in this case, primary) audiences for their decisions, on the one hand, and the reality of how unreadable their decisions will likely be to these audiences, on the other hand. This problem is one that merits further consideration and attention from judges, regardless of the extent to which lawyers might step in to help parties understand that which they find unreadable, if only to allow judges to reconsider the coherence of their statements that they are writing for the parties.

---

[18] Saunders, "Reflections", *supra* note 1 at 71; and, *Steven Wilson's Questionnaire*, *supra* note 1.

And, as the next study will illustrate, there are many other circumstances wherein parties will simply not have access to lawyers to help them understand decisions.

5.3 The Readability of Decisions Involving Self-Represented Litigants

One reason for examining readability levels by legal subject area in Chapter 4, and for looking specifically at the readability of family law decisions, was because we know that SRLs are more prevalent in family law cases than in other types of cases. But measuring MADRS results for family law cases is only an indirect way of gauging the readability levels of decisions involving a SRL, because lawyers are involved in many (but perhaps not a majority) of cases. In order to more directly measure readability levels for cases wherein at least one party is unrepresented, the study that follows is designed to isolate and report on the readability results of a corpus of decisions where each decision involves at least one SRL.

5.3.1 Method – Measuring the Readability of Decisions Involving Self-Represented Litigants

The research method that I used in this study was similar to the method for all previous studies in most respects. I assembled a corpus of decisions that each implicated at least one SRL. I began by searching the CanLII database's full text field for the phrase "self-represented." I then filtered the search period to only results from 2022, and to only "cases" (excluding secondary sources and legislation). I selected all jurisdictions other than Quebec, and filtered to "all courts" (excluding tribunals).

My list of search results at this point consisted of 1588 cases, indicating that the full population of cases from 2022 was large enough to merit the use of a random-sampling approach. I again used a 95% confidence interval, and a 7% margin of error. Using these two parameters, and a total population size of 1588 cases, I calculated the required sample size to be N=175, using a

free online statistical calculator.[19] With these parameters, we can infer that results from this sample of self-represented cases would be representative of results (within a maximum range of + / - 7%) for the full population of cases from which the samples are drawn, 19 times out of 20.

All other exclusion criteria that applied to cases (described above in Chapter 4) for my first study of readability scores across court levels were also applied here, for all of the same reasons. Randomization was also achieved in the same manner as in the previous studies. And, text was extracted from cases, inserted into files, and linguistically analyzed all in the same way as for the previous studies. However, I immediately (and expectedly) realized that my full text search for the term "self-represented" was yielding a large number of false positive results – that is, results that contained the phrase "self-represented" within the decision even though all parties were actually represented by lawyers.[20] Consequently, I manually screened each result within my list of randomly selected cases to ensure that only cases that truly involved a SRL were included within my corpus. Following these steps caused me to exclude 46 of the 175 cases that had originally been randomly identified for inclusion – a much higher proportion of exclusions than in any previous study – but I attribute this higher exclusion rate to the relatively broad full text search phrase that I used to generate my list of results. Ultimately, each excluded case was replaced with the next-in-sequence case that met all the necessary criteria, so the uniquely high exclusion rate in this study is unlikely to have had any adverse impact on the results.[21] In the end, the corpus

---

[19] Online: <https://www.qualtrics.com/blog/calculating-sample-size/>.

[20] Many of these false positives arose because the authoring judge referred, within their reasons, to other cases wherein someone was self-represented (e.g.: to distinguish the judge's present case from a previous case that would otherwise have had precedential value). Other false positives periodically arose because a party in the authoring judge's case had, at some earlier point in the legal process, been self-represented – even though they were represented by counsel when the judge heard and decided the case that was the subject of the decision.

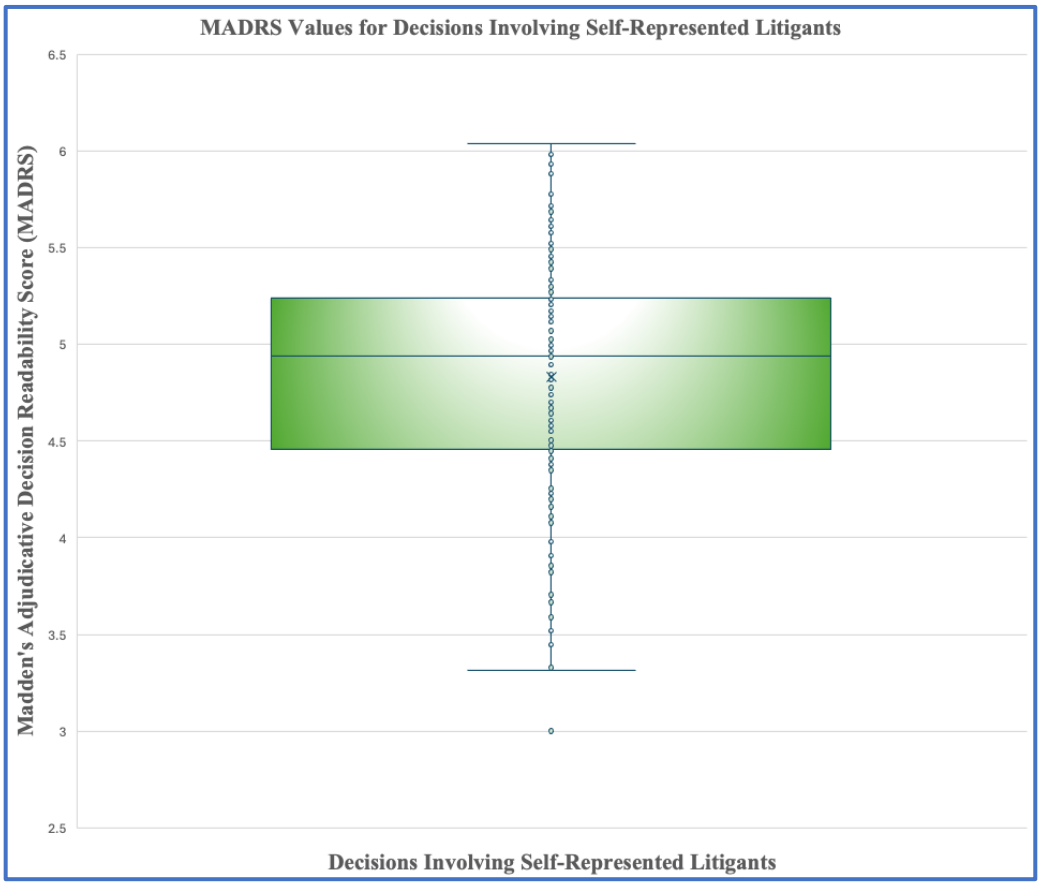
[21] To the contrary, the exclusions indicate that I likely over-estimated the full population size of true "self-represented" decisions, which means that my sample size of N=175 would likely produce slightly better than a 7% margin of error.

consisted of decisions that had been verified to each include at least one SRL, which allowed me to compute focused MADRS results for this special class of decisions.

5.3.2 Discussion & Analysis – Readability Scores for SRL Decisions

The 175 decisions in this corpus contained over 761,000 words of text. A spreadsheet containing the full MADRS results for these decisions, including component scores for each contributing variable and my notes for any instance where a randomly selected case was excluded and replaced with the next-in-sequence case, is publicly available.[22] A graphical representation of the MADRS results is shown below in Chart 5.3.

Chart 5.3 – MADRS Values for Decisions Involving Self-Represented Litigants

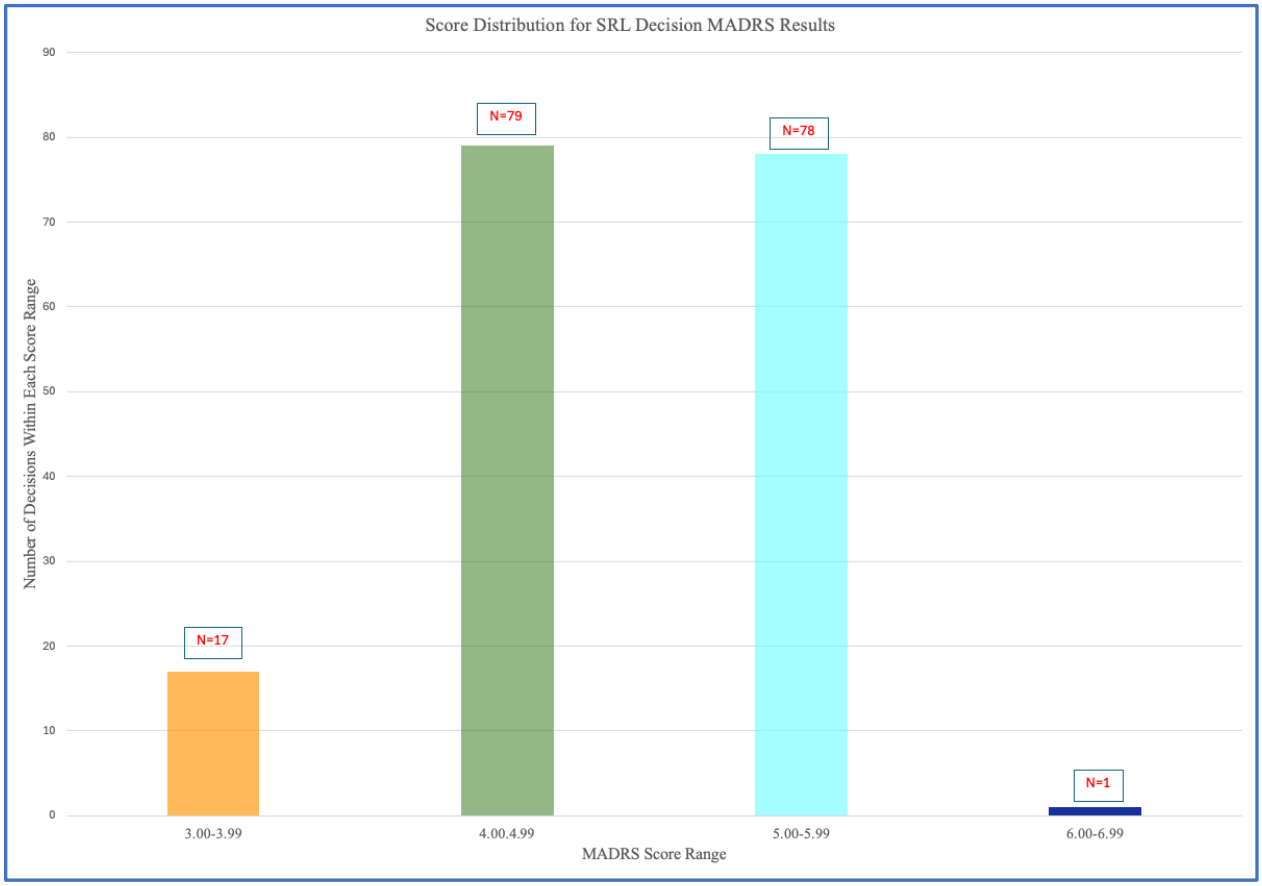


[22] Mike Madden, "MADRS_MaddenDissertation_Chapter5_SelfRepResults" (2025), online: <https://doi.org/10.7910/DVN/8YMQUE>, Harvard Dataverse, V1.

These results are, in many ways, the most surprising and disappointing of any that are reported within this dissertation. The MADRS average for decisions involving SRLs is 4.83, which means that something close to a completed undergraduate degree is needed to understand the average decision. This average result is noticeably higher than the average of 4.69 for criminal law decisions that was reported above, in Chapter 4 – which is striking when one considers that all of the decisions within the criminal law corpus involved accused persons or offenders who were represented by counsel.

A breakdown of the score distribution for these SRL decisions is shown below, in Chart 5.4.

Chart 5.4 – Score Distribution for SRL Decision MADRS Results



As this chart shows, only 9.7% of the decisions scored below 4 on the MADRS scale; 45% scored between 4 and 5 on the scale; and 45% of decisions scored above 5 on the scale (signaling that something between a completed undergraduate degree and a completed law degree would be needed to understand these last decisions). No decisions scored below 3.

I characterize these results as surprising because a judge's responsibility to assist a SRL is now well-established. In 2006, the Canadian Judicial Council published guidance to judges that described their responsibilities "to ensure that self-represented persons are provided with fair access and equal treatment by the court,"[23] including by making "all aspects of the court process to be, as much as possible, open, transparent, clearly defined, simple, convenient and accommodating."[24] Judges have acknowledged these responsibilities in case law:

> It is well settled that where an accused is self-represented at trial, the presiding judge has a duty to ensure that the accused has a fair trial. To fulfill that duty, the trial judge must provide guidance to the accused to the extent that the circumstances of the case and those of the particular accused may require it. Within reason and without becoming counsel for the accused, trial judge must provide assistance to aid the accused in the proper conduct of his defence and to guide him, as the trial unfolds, in such a way that the defence is brought out with its full force and effect.[25]

And, judges have similarly acknowledged these responsibilities in their extra-judicial writing and commentary.[26]

All of the above discourse, however, seems to be preoccupied with helping SRLs who cannot afford lawyers, and who must therefore represent themselves in order to obtain access to

---

[23] Canadian Judicial Council, "Statement of Principles on Self-represented Litigants and Accused Persons" (September 2006), online: <https://cjc-ccm.ca/sites/default/files/documents/2020/Final-Statement-of-Principles-SRL.pdf>, at 1 ["Principles"].
[24] Canadian Judicial Council, "Principles", *supra* note 23 at 2.
[25] *R v Breton*, 2018 ONCA 753, at para 13.
[26] Wayne K Gorman, "Access to Justice in Canada: The Unrepresented Accused and the Role of the Trial Judge" (2020) 56:1 Court Review 48.

justice.[27] These litigants seem to benefit from judicial sympathy and compassion.[28] One might therefore reasonably expect judges to address these SRLs more directly in their decisions, by using language that the SRLs would likely understand without any need for external resources (like lawyers, who they have not engaged). If judges accept that they have obligations to help these types of SRLs to understand and access the processes that ultimately lead to a decision from the court, then presumably judges would also accept that they have a responsibility to help these SRLs to understand the decision itself – by writing the decision in a way that is capable of being understood by the relevant SRLs.

However, former Chief Justice of Nova Scotia MacDonald noted that the above class of self-represented litigants (that is, people who cannot afford a lawyer but who still need to go to court) is only one of three common classes of SRLs. A second class consists of people who "probably could afford lawyers but think, 'There's nothing to this,' so they do it themselves."[29] Chief Justice Macdonald explains that "the more we try to provide access and help for people who really do need help, the more we accommodate people who really should and could have lawyers, but have chosen not to."[30] MacDonald seems to suggest that this second class of SRLs is perhaps not as worthy of judicial sympathy and assistance as the first class of SRLs.

Chief Justice MacDonald also describes a third class of SRLs, who present a different kind of problem for the judiciary:

---

[27] Sean MacDonald, "Balance and Independence: The Judicial Process in the Charter Era" (2007) 16 Dalhousie Journal of Legal Studies 161, at 169 ["Balance and Independence"]. I note that this article, although styled as one that is authored by Sean MacDonald, is actually the transcript of an interview that took place in January 2007 between Sean MacDonald (a third-year law student at the time) and the Honourable Michael MacDonald (who was the Chief Justice of Nova Scotia at that time).

[28] MacDonald, "Balance and Independence", *supra* note 27 at 169: "There is a body of litigants out there who cannot afford lawyers. We could spend the whole day talking about trying to solve that problem and we have to do what we can to help those people. […] Those are the people we want to target and help as much as we can."

[29] MacDonald, "Balance and Independence", *supra* note 27 at 169.

[30] MacDonald, "Balance and Independence", *supra* note 27 at 169.

> [l]astly, and, tragically, we have a smaller group of litigants who may be small in number, but take up an enormous amount of court time. Their problems are intertwined with mental health issues and it becomes very much a challenge for the judiciary when they bring cases; these cases can go on and on, and they don't necessarily appreciate the process. For instance, they will issue subpoenas for the Pope, the Governor General, the Premier, the Prime Minister. It's very sad.[31]

While MacDonald suggests that this final group of SRLs might benefit from judicial sympathy, it is unclear whether he feels that this third class of SRLs should benefit from assistance in the same way as the first class.

I have described these classes of SRLs here because it seems to me, impressionistically, that the MADRS results for cases involving SRLs might be affected by the heterogenous nature of the SRL population. In other words, most SRL discourse seems to focus on the plight of ostensibly worthy or deserving SRLs. From a readability perspective, we might therefore expect decisions in cases involving SRLs to be highly readable. After all, judges seem to show sympathy, and offer various forms of assistance, to these SRLs. But judges interact with all 3 classes of SRLs. And in some places, or for some judges, the interactions might be far more frequent with the second and third classes of SRLs, who judges may perceive to be less worthy of assistance from the court. To be clear, I am offering a hypothesis that might explain the results shown above in Chart 5.3, rather than declaring a truth that can be demonstrably substantiated by evidence. But perhaps the ensuing example relating to a judge from Alberta will help to support my hypothesis.

Consider, for instance, the situation of the Honourable John D. Rooke, who was Associate Chief Justice (ACJ) of the Court of King's Bench of Alberta during 2022, until he retired on December 22 of that year.[32] ACJ Rooke's name appeared frequently within decisions that had been

---

[31] MacDonald, "Balance and Independence", *supra* note 27 at 169.
[32] Canada, Prime Minister's Office, "Prime Minister announces appointment of new Associate Chief Justice of the Court of King's Bench of Alberta" (October 10, 2023), online: <https://www.pm.gc.ca/en/news/news-releases/2023/10/10/prime-minister-announces-appointment-new-associate-chief-justice>.

randomly selected for inclusion: he authored 11 (6.3%) of the 175 decisions within the corpus. That one author would have contributed so many decisions, when there are currently 1195 federally appointed judges[33] and hundreds of other provincially appointed judges in Canada, is statistically unlikely. To even have two decisions from the same judge within a random sample would be unusual, if judges were presiding over cases involving SRLs at equal rates. It seems, therefore, that ACJ Rooke was deciding cases involving SRLs with disproportionate frequency.

But a closer look at ACJ Rooke's decisions within the sample reveals that he was not dealing with the kind of people who fit into Chief Justice MacDonald's first class of SRLs. For instance, in the first sentence of *Royal Bank of Canada v Anderson*,[34] ACJ Rooke describes the SRL in that case as follows: "Sandra Ann Anderson [Ms. Anderson] is a highly active Organized Pseudolaw Commercial Argument [OPCA] litigant whose illegal pseudolaw activities are documented in a number of Court Decisions."[35] In describing some of the evidence that this defendant submitted, ACJ Rooke lists, among other things,

> A nonsensical pseudolaw document, titled "Writ-of-Quo-Warranto", that is reproduced in part in Appendix "A" below, that purportedly would be filed in the *RBC v Anderson* Action, but, startlingly, adds the "Treasury Inspector General of the United States of Alberta" as a "co-Plaintiff" with Ms. Anderson, and Queen Elizabeth II, personally, as a "Defendant", along with Counsel for RBC. In the interests of judicial economy, I will simply state that this document, at its most generous, appears to attempt to establish some kind of counterclaim or counter-lawsuit, in the *RBC v Anderson* Action. The "Writ" is an absurd conglomerate of quotations, citations, and nonsense.[36]

---

[33] Office of the Commissioner for Federal Judicial Affairs Canada, "Number of Federally Appointed Judges as of May 1, 2024" (May 1, 2024), online: <https://www.fja.gc.ca/appointments-nominations/judges-juges-eng.aspx>.

[34] 2022 ABQB 577, online: <https://www.canlii.org/en/ab/abqb/doc/2022/2022abqb577/2022abqb577.html>. The full style of cause for this case is "*Royal Bank of Canada v Sandra Ann Anderson A.K.A. Sandra-ann, Authorized representative A.K.A. Sandra-Ann: Anderson, A.K.A. SANDRA ANDERSON, WOMAN SANDRA OF THE ANDERSON FAMILY, A.K.A. i woman sandra-ann, acting as agent for and on behalf of SANDRA ANDERSON, A.K.A. sandra-ann:woman, A.K.A. Sandra-Ann, Executor, A.K.A. sandra-ann, A.K.A. Sandra-Ann: Anderson All Rights Reserve Cestui Que Use, Principal Beneficial Owner*", based on the way in which the defendant identifies herself within this litigation.

[35] 2022 ABQB 577, online: <https://www.canlii.org/en/ab/abqb/doc/2022/2022abqb577/2022abqb577.html>, at para 1.

[36] 2022 ABQB 577, online: <https://www.canlii.org/en/ab/abqb/doc/2022/2022abqb577/2022abqb577.html>, at para 17 (internal citations omitted).

ACJ Rooke's description and treatment of the SRL defendant's documents in this case suggests that she falls into the third class of SRL, and that she does not deserve the court's assistance in any way.

The above passages also highlight how a judge who is dealing with this type of litigant may be partially limited in their ability to use simple language within a decision – because the judge must refer to the nonsensical language of the SRL to at least some extent when writing the decision. However, it also illustrates how a sense of frustration toward the SRL, or potentially some other factor within the judge's control, may also contribute to the production of a less readable decision – in this case, one that uses uncommon words like "purportedly" and "conglomerate",[37] and phrases like "judicial economy."[38] We cannot be certain that there is a causative relationship here that connects the SRL's conduct to any frustration from the presiding judge, and ultimately to the readability level of the decision. After all, many other decisions that form part of different studies in this dissertation also scored as high or higher on the MADRS scale when compared to ACJ Rooke's decision, and I have no plausible basis for suggesting that any of those other judges' emotions were factors affecting the readability scores of the other decisions. Again, I am simply hypothesizing here that these factors may all be connected in cases involving certain types of SRLs, in ways that lead to less readable decisions.

The above decision from ACJ Rooke was his third-most readable decision of the 11 within the current study's sample. It scored 5.07 on the MADRS scale. Based on the way in which the

[37] In the COCA corpus, the word "purportedly" is ranked #17149 based on frequency of appearance in the corpus. The word "conglomerate" is ranked #12555 (online: <https://www.english-corpora.org/coca/>).
[38] In the COCA corpus, the phrase "judicial economy" only appears a total of 14 times across the more than 1 billion words of text in the corpus.

defendant was portrayed within the decision by ACJ Rooke, one suspects that Ms. Anderson would not have easily understood the decision.

ACJ Rooke's decisions scored an average of 5.22 on the MADRS scale. The least readable decision was a short one that scored 5.78,[39] and that also involved SRLs who had litigation restrictions[40] imposed upon them by the Alberta courts. However, ACJ Rooke was not particularly critical of these litigants, and seemed to address them with practical direction about how they could continue to advance their litigation while complying with the litigation restrictions. The relatively few sentences that ACJ Rooke used in this case described procedural law using relatively uncommon terms, and were generally long with multiple clauses. The high MADRS value for the decision flows from these factors.

ACJ Rooke's second-least readable decision was, in my view, much more representative of his decisions within the sample as a whole. This decision (*Hayden v Hayden*)[41] scored 5.72 on the MADRS scale. The first paragraph sets the scene: the plaintiff is described as "an aggressive, relentless, and seemingly unstoppable abusive litigant. [Her] successive hopeless lawsuits, and illegal applications, have now continued in the Alberta Court of King's Bench and other Alberta tribunals for well over a decade. [She] is also a highly active but consistently unsuccessful appeal court litigant."[42] ACJ Rooke goes on to deal with the law about vexatious litigants, court access gatekeeping, and judge-shopping before offering his concluding thoughts:

> I have exhausted the Court's mechanisms to manage Ms. Hayden […]. [P]redictably, Ms. Hayden will take additional future steps to achieve her intended objectives: to do whatever she wants, however she wants it. This ridiculous farcical festival will now continue to cycle

[39] *Sun v Tesla*, 2022 ABQB 445.
[40] Litigation restrictions seem to consist mostly of court-ordered restrictions that impose obligations on litigants who have shown a pattern of frivolous, vexatious, or abusive litigation to seek leave from the court before filing any new originating documents: *see Re Onischuk*, 2017 ABQB 663, at para 23.
[41] 2022 ABKB 648, online: <https://www.canlii.org/en/ab/abkb/doc/2022/2022abkb648/2022abkb648.html>.
[42] 2022 ABKB 648, online: <https://www.canlii.org/en/ab/abkb/doc/2022/2022abkb648/2022abkb648.html>, at para 1.

> over and over, around and around, with no end in sight. And Ms. Hayden is far from unique. Others also operate the same way, ignoring court procedure, resisting, or simply defying, any attempt at management, such as Chinedu Gideon Ubah, Keenan A Feeney, Janice Dmyterko, Sandra Ann Anderson, Salim Rana, and many more. They do whatever they want, squander court resources, harass and threaten staff, and incrementally erode the operation of courts and tribunals, often across many jurisdictions. I do not see that as "access to justice". But such is the state of things in Alberta courts, because that is "the law".[43]

Several of the parties mentioned in this passage (Anderson, Ubah, Feeney) are also SRLs whose decisions appear in the current study's sample of SRL decisions. ACJ Rooke's frustration with these individuals is almost palpable from his sarcastic and hyperbolic language. He seemingly has no desire to assist these people.

The irony, though, is that these people are perhaps the ones who most need a readable decision from the judge in order to understand how and why their litigation is fruitless. However, whether from frustration or ill will, ACJ Rooke writes his decisions in these SRL cases using language that the litigants will likely find difficult to understand. And ACJ Rooke is not unique in adopting this writing style. In *Durkin v Facebook, Inc*,[44] a different judge from British Columbia noted that "considerable qualification is required when describing [the plaintiff] as self-represented in the sense of inexperienced. He has been a party in the following cases [list of 7 cases in which the SRL plaintiff was involved]."[45] The judge goes on to explain how, in one of those prior cases, the SRL plaintiff had been found to have deliberately sworn a false affidavit,[46] and in another, he had been characterized by the judge as "evasive, combative, deliberately vague and frequently

---

[43] 2022 ABKB 648, online: <https://www.canlii.org/en/ab/abkb/doc/2022/2022abkb648/2022abkb648.html>, at paras 13-14 (internal citations omitted).
[44] 2022 BCSC 1305, online: <https://www.canlii.org/en/bc/bcsc/doc/2022/2022bcsc1305/2022bcsc1305.html>.
[45] 2022 BCSC 1305, online: <https://www.canlii.org/en/bc/bcsc/doc/2022/2022bcsc1305/2022bcsc1305.html>, at para 19.
[46] 2022 BCSC 1305, online: <https://www.canlii.org/en/bc/bcsc/doc/2022/2022bcsc1305/2022bcsc1305.html>, at para 21.

dishonest."[47] The MADRS result for this decision was 5.52. In a case where the SRL plaintiff is not a particularly sympathetic character, one wonders about how much effort a judge is inclined to make in order to write a decision that this plaintiff will easily understand.

Ultimately, judges themselves will need to decide whether, and to what extent, they ought to help SRLs who behave in the ways that some of the above SRLs behave. The frustration that these judges might feel when confronted with SRLs who try to use the court's processes for anti-social purposes is somewhat understandable. But Chief Justice MacDonald's description of the situation – that is, one where SRLs often make litigation decisions in ways that are likely affected by their mental health issues – as being sad and tragic is perhaps worth recalling here. To the extent that many of these SRLs might be less equipped to control and manage their litigation impulses than others in society, due to personality disorders or other mental health conditions, it is perhaps unjust to punish these individuals with a form of judicial 'cold shoulder' that other SRLs do not receive. After all, even criminal offenders who engage in anti-social behaviour can still expect respectful (and more readable) decisions from the courts – although their circumstances are different in the sense that they find themselves before the courts involuntarily, unlike the SRLs who instigate proceedings. Under these circumstances, it may not be fair or appropriate for SRLs like those described above to be denied the same treatment from the courts within the courts' decisions.

In spite of the above discussion about how a particularly difficult class of SRLs are treated and written about in court decisions, these individuals comprise a minority of the overall number of SRLs from decisions within the current study. The 3 least readable decisions within the study,

---

[47] 2022 BCSC 1305, online: <https://www.canlii.org/en/bc/bcsc/doc/2022/2022bcsc1305/2022bcsc1305.html>, at para 22, citing *SHH Management Limited v Philip*, 2020 BCSC 1411, at para 22.

for instance, all involve SRLs whose cases are unsuccessful but who are not characterized as abusive litigants. In *Curtis v Canada (Employment, Workforce Development and Disability Inclusion)*[48] (MADRS = 6.04), the plaintiff's claim against the Bank of Nova Scotia and several Crown actors was struck out as disclosing no cause of action. In *Toronto (City) v Nawrocki and Zest For Living Etobicoke Inc*[49] (MADRS = 5.98), the appellant municipal government successfully appealed the sentence imposed on the SRL respondent in relation to charges for operating a business in violation of COVID-19 restrictions. And, in *Tewari v Sekhorn*[50] (MADRS = 5.93), the SRL plaintiff's statement of claim against the law firm that had successfully represented his employer in the plaintiff's action against the employer for unjust dismissal was again struck out for disclosing no reasonable cause of action.

It is possible that the SRLs in these cases fall within the second category described by Chief Justice MacDonald – people who have enough money to afford a lawyer, but who choose not to engage one because they feel capable of representing themselves.[51] There is not enough information about the financial circumstances of the SRLs in any of these decisions to allow us to make this determination. But despite the high MADRS results in each of these cases, the results do not seem to flow from feelings of frustration with, or a lack of sympathy for, the SRLs in the cases. So regardless of whether these people fall into the first or second category of SRL, it is somewhat surprising that the decisions score so high on the MADRS scale: there are no indicators that the judges in these decisions are annoyed with the SRLs, and yet the decisions are still less readable than the average decision from every other contexts seen within this dissertation. In each

[48] 2022 FC 826.
[49] 2022 ONCJ 307.
[50] 2022 ONSC 417.
[51] MacDonald, "Balance and Independence", *supra* note 27 at 169.

case, where a completed law degree would effectively be required to understand the decisions, the courts have likely failed to communicate in language that the SRLs will understand (unless by some coincidence, each of the SRLs in these cases also possessed a law degree). The high MADRS values within this study as a whole are therefore not well explained just by reference to the phenomenon of especially difficult SRLs who appear before the courts.

The four most readable decisions in the current study all relate to family law.[52] None of the litigants are characterized as abusive. However, the *CC v SC*[53] decision (MADRS = 3.45) is interesting, because the judge in this decision expresses considerable frustration with the SRL mother's conduct, while still showing a measure of empathy and concern for her. After recounting how the mother keeps sabotaging her case for greater access to her child by calling the police to the father's home without any basis, Justice Conlan makes the following statements:

> I will see the mother in a few months. She should continue with her teaching in the meantime. She should get some counselling and therapy. She should focus on her own mental health. She should stop visiting websites where the common denominator is a disdain for and an overwhelming feeling of having been screwed by the justice system. […] I think that this Court is in danger of becoming the person who cries wolf, but the mother must realize that there comes a time when even the most fervent hopes, even those for the sake of a child, are gone forever. There comes a time when the Court has no reasonable alternative but to make a final order that removes one parent completely from the life of her child. No judge wants to do that to [the child].[54]

I have quoted from this part of the decision at length to illustrate how it is possible for a judge to be frustrated with the conduct of a SRL, while still writing a readable decision that the SRL can likely understand. Contrary to what was seen above in the decisions of ACJ Rooke and other judges who deal with abusive litigants, it is not necessarily the case that a judge must write a more

---

[52] *Robert v Louisseize*, 2022 ONSC 1756; *Leavitt v Leavitt*, 2022 ABPC 200; *Walton v Davis*, 2022 ONSC 1784; and, *CC v SC*, 2022 ONSC 5225.
[53] *CC v SC*, 2022 ONSC 5225.
[54] *CC v SC*, 2022 ONSC 5225, at paras 14-15.

complex or less readable decision just because one of the parties is difficult to deal with during the proceedings.

I acknowledge the difficult task that faces judges who preside over cases involving SRLs. The level of emotion in these cases is going to be higher than in cases where a dispassionate and objective lawyer represents each party. In contrast, the levels of legal and procedural knowledge possessed by the SRL parties in these cases is far lower than in cases where the parties are represented by lawyers. Judging these cases is hard. But the results of the present study are nonetheless disappointing. SRL parties are arguably people who are most in need of a readable decision. While all parties might ideally hope to be able to personally understand the decisions in their cases, those who are represented by lawyers at least have someone skilled to whom they can turn if they fail to understand the decision on their own. SRLs are alone. If judges do not write in ways that SRL parties can understand, then there is very little hope for these parties to ever feel as if they have received a just decision. The "overwhelming feeling of having been screwed by the justice system" that Justice Conlan acknowledges in the SRL mother who appears before him[55] would probably be aggravated if the judge had chosen to write his decision in language that she cannot understand, instead of in the simple and straightforward way that Justice Conlan actually wrote the decision. And the same kind of feeling has likely resulted in the many other cases within the current study where judges have written decisions that cannot be understood by the SRL parties in those cases. Whether they engage in poor conduct before the courts, or are simply untrained in the law, SRLs deserve decisions that they can understand. Our courts should strive to provide them with these kinds of decisions.

---

[55] *CC v SC*, 2022 ONSC 5225, at para 14.

5.4 Key Takeaway: Readability Deficits in Unfortunate Places

The results from Chapter 4 revealed many things about the readability levels of Canadian court and tribunal decisions. Most of what was reported in that chapter was not very encouraging. But there were elements of the results in that chapter that illustrated what is possible – both from individual judges, and more systemically – in terms of producing more readable decisions. A number of criminal court judges demonstrated how they can write individual decisions that would be easily understood by people with only some high school education. At least one judge (Justice West) demonstrated that he can consistently write decisions at less than the 4.0 level on the MADRS scale. And, one tribunal (the SST) has shown that its decision-makers can also consistently write at less than 4.0 levels, when courts writing on the same subject score almost 1.2 points higher on the MADRS scale for their decisions.

Having seen these promising results in areas of law where we had no particular expectation for decisions to be unusually readable, my sense was that the readability results within this chapter ought to have been more consistently low-scoring. The only cases that featured within this chapter were, after all, cases involving either YCJA prosecutions or SRLs. In both groups of cases, there are clear and compelling justifications for why a decision ought to be readable. Furthermore, these justifications are not linked to peripheral audiences who may not ever be interested in reading the decisions; rather, they relate to primary audiences, who almost certainly have reason to read and understand the decisions in their cases, because each individual is a party to a case – someone who is directly affected by its outcome.

In the case of young persons who are prosecuted under the YCJA, decisions generally need to be written in simple language if they are to be understood, because the young persons will rarely be old enough to have attained anything beyond a secondary school level of education. Even

though these individuals are typically represented by lawyers, they will struggle to understand the decisions personally unless they are written at something close to the 3.0 level on the MADRS scale. The current average of 4.37 is likely too high to be understood by most subjects of the decisions.

And, in the case of SRLs, the average decision readability score is even higher: 4.83. These scores might be justified if almost everyone in Canada had something close to a completed undergraduate degree. However, only one-third of the adult population has an undergraduate degree.[56] These numbers suggest that most people in Canada, if they chose to represent themselves in legal proceedings, would struggle to understand the most important document that emerges from litigation: the decision that explains the results of the process.

Again, these quantitative results do not allow us to quickly make individualized determinations about whether decision X or Y was written to effectively reach audience member A or B for the decision. It is theoretically possible that each decision was carefully calibrated to use just the right level of linguistic complexity for the specific audiences who would read the decisions. But a macro-level view of the results highlights the improbability of this hypothesis being true. Most decisions are written at levels that most members of the audience would likely not understand – particularly in the types of cases that have been studied in this chapter that involve young persons and SRLs whose readability needs are probably greater than those of adults who are represented by lawyers. The results in this chapter, perhaps even more than anywhere else in this dissertation, suggest that judges can and should do more to write readably when they are dealing with people who are not well-equipped to understand the judge's decision.

---

[56] Statistics Canada. "Table 98-10-0384-01: Highest level of education by census year: Canada, provinces and territories, census metropolitan areas and census agglomerations" (December 9, 2022), online: <https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=9810038401>.

6. Adjudicative Decision Readability: Getting Personal

6.1 Overview

This chapter attempts to place a more personal spin on applications of the MADRS formula to adjudicative decisions, in two ways. The first way is researcher-centric. This chapter describes and provides readers with access to MADRS results from a large corpus of decisions that have been individually coded for a number of different variables. The goal here 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 chapter gets personal is more author-centric: it identifies the authoring judge or tribunal member for each decision within the list of results. The justification for creating a more personalized, and personalizable, set of MADRS results is included below, before a description of the methods for this effort, and a brief discussion of some of the results.

6.2 Motivations: Why Get Personal?

As I noted above in both Chapters 1 and 5, I do not believe that there is 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, as I have also noted elsewhere in this dissertation,[1] the general public is very frequently going to be part of the intended audience for adjudicative decisions, and decisions should probably 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 in order 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 and analysis from the two previous chapters now show, we have strong reasons to believe that many, if not most, adjudicative decisions

[1] See above, at section 4.3.2, in the context of a discussion about the readability levels of family law decisions.

are being written in ways that 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 the dissertation 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 is one that would communicate the key content of my dissertation to what I feel would be the most appropriate group of people – the judges and tribunal members who write decisions – for them to consider how readability levels might or might not align with the expected reading competency levels of their intended audience members. As this statement implies (and as I discuss at greater length in the next chapter, dealing with my recommendations), 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 for those audience members.[2] 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.

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 this dissertation. It is well-established in psychological literature that individuals tend to over-estimate their own performance relative to their peers.[3] This phenomenon seems to apply (as predicted) to legal authors who self-assess their writing:[4] 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 this dissertation to tell themselves that the results probably reflect their peers' decisions, but not their own decisions.

It is partially for this reason that my dissertation now gets personal. In this chapter, I describe and provide access to readability results from a study of 1621 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

---

[2] 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. So 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.

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

[4] Elizabeth Ruiz Frost, "Overcoming 'Illusory Superiority': Mentoring Legal Writers" (2013) 73:9 Oregon State Bar Bulletin 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 (at 13): "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."

total of 840 different authors have written decisions whose results are included in this Table of Results. My hope is that, by providing the legal and judicial/tribunal member communities with access to these personalized results, authors will find personal relevance to the contents of this dissertation – and motivation to explore the results.

The current study also classifies each decision based on several other variables: court level; jurisdiction; legal subject; author's gender; opinion type; number of judges agreeing with an opinion; percentage of judges agreeing with an opinion; and, word count. The inclusion of all 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 chapter 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. In other words, this chapter describes a study that was not designed to answer my own research questions. The study was instead designed to offer data and results up to others who might, for all different types of reasons, want information about readability levels of Canadian court and tribunal decisions – subject to the limitations that I discuss below.

6.3 Method: Getting MADRS Results for a Large Sample of Decisions

My research method for this study was similar in many respects to the methods that I used in studies described in Chapters 4 and 5 of this dissertation. In simple terms, 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.

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 produce approximately only 0.2% as many decisions as Ontario produces – that is, they produce only 2 decisions for every 1000 decisions from Ontario).[5]

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 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

[5] 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.

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.

In order to ensure that the corpus contained a wide range of different decision types based on the different variables that 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. These files are the same as those used in my study of readability differences across court levels that was reported in Chapter 4.

I included a random sample of 25 decisions from each provincial appeal court, and from the Federal Court of Appeal. In any cases where a provincial/territorial appeal court had released less than 25 decisions in 2022, I took all decisions from that court meeting the inclusion criteria.

I included a random sample of 25 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 25 decisions in 2022, I took all decisions from that court meeting the inclusion criteria.

I included a random sample of 25 decisions from each province's or territory's inferior court(s). In any cases where provincial/territorial inferior court(s) had released less than 25 decisions in 2022, I took all decisions from that level of court meeting the inclusion criteria.

I included a random sample of 25 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.

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;
    - 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.

All other exclusion criteria that applied to cases (described above, in Chapter 4) for my first study of readability scores across court levels were also applied here, for all the same reasons. Text was extracted from cases, inserted into files, and linguistically analyzed all in the same way as for the previous studies.

However, for each file, I manually reviewed the decision/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 7 cases by searching publicly available information like court or tribunal websites, or announcements of appointment, to ascertain how the different judges/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).

Following these steps allowed me to compute MADRS values for all decisions/opinions within the broad corpus, where each file was also coded for each of the variables listed above.

6.4 Results: Some Examples and a Resource for Further Study

MADRS results were calculated for 1621 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 corpus of opinion/decision files are shown graphically, below.

Chart 6.1 – Breakdown of Corpus Files by Court Level

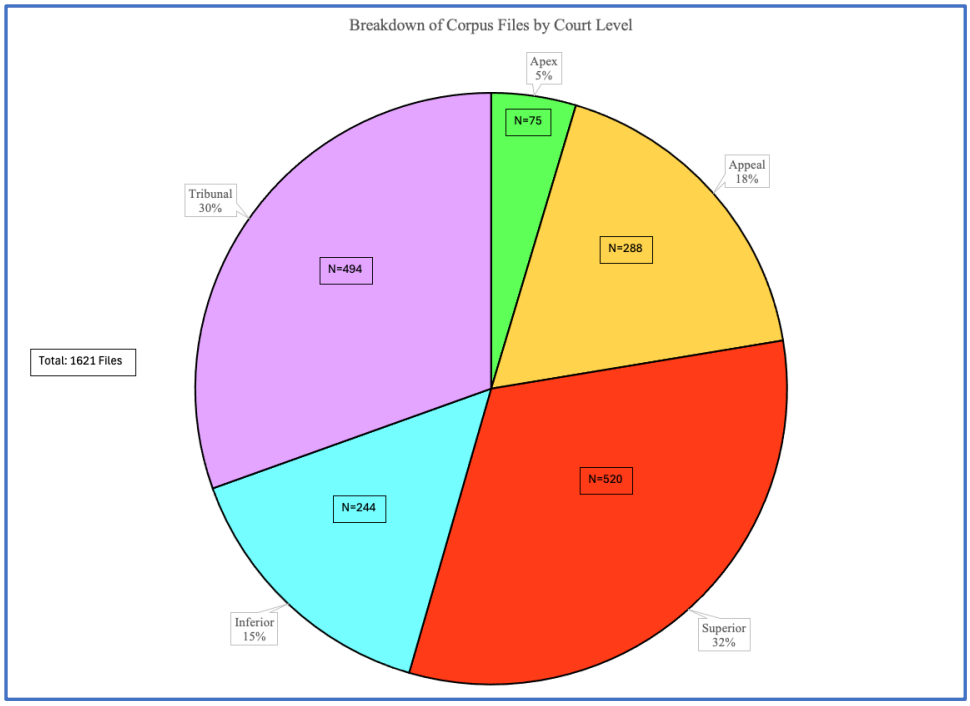
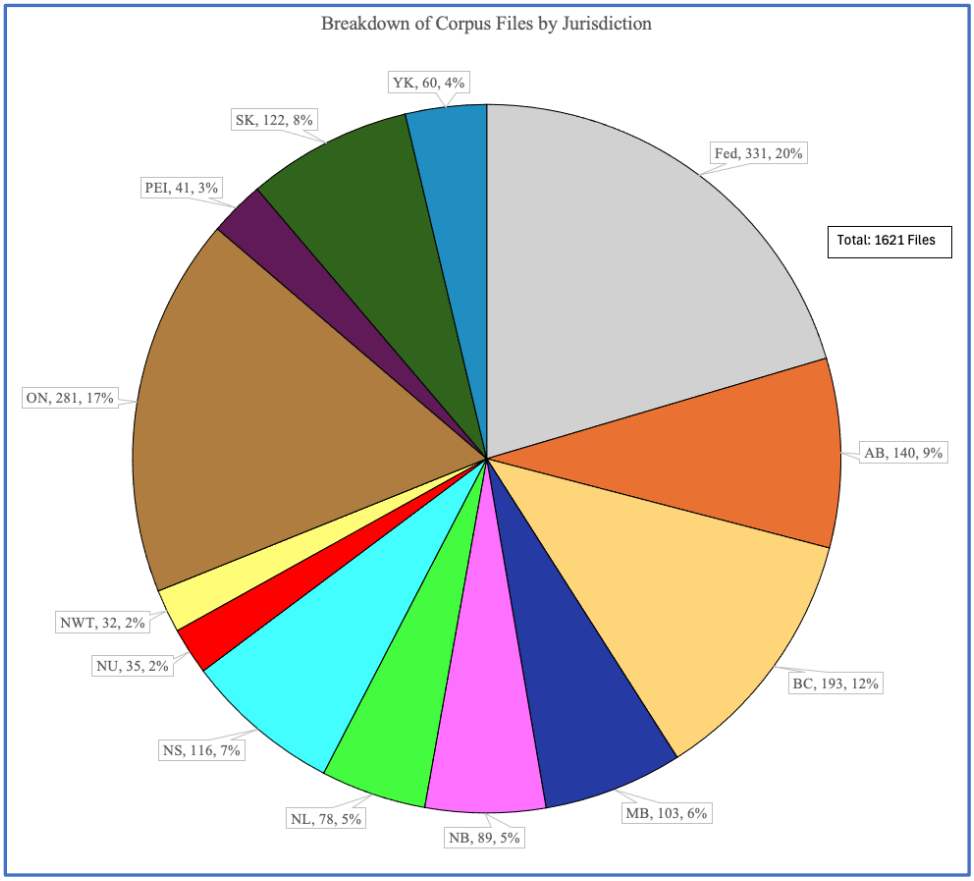


Chart 6.2 – Breakdown of Corpus Files by Jurisdiction



I discuss some of the readability results that emerge from decisions within this corpus for illustrative purposes, below. However, for all of this discussion, it is important to recall that the results from this non-representative study are not generalizable to the broader population of Canadian cases as a whole. So, conclusions drawn about the MADRS results from this sample of 1621 cases are only certain to be valid in respect of these cases, but not necessarily to other Canadian cases.

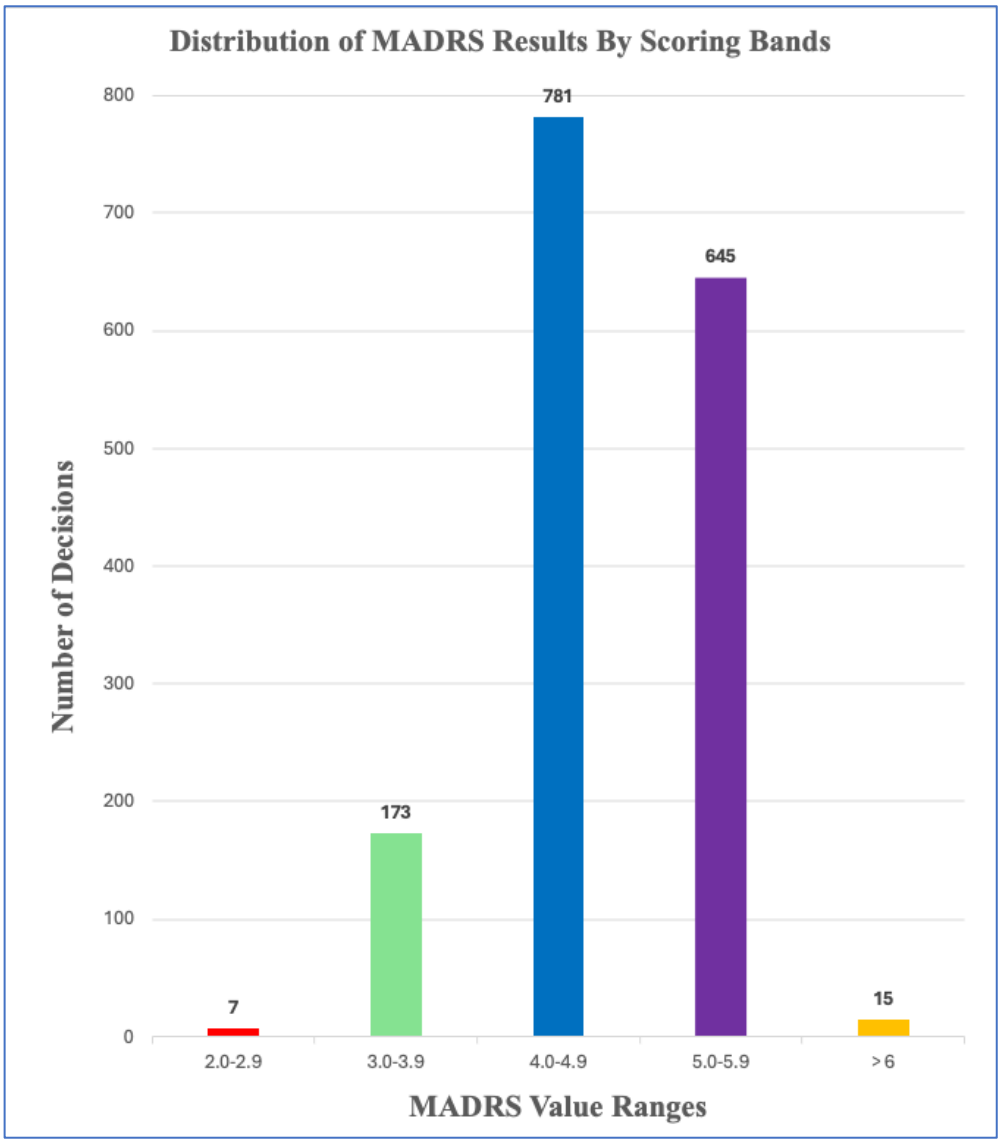
Nonetheless, the analyses that follow may illuminate research questions or paths that would be fruitful for further inquiry in the future. This discussion may essentially suggest that, if a particular phenomenon was observed in this study of a broad (but not representative) sample of decisions, then 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.

6.4.1 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 in order for the reader 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 at Chart 6.3.

Chart 6.3 – Distribution of MADRS Results by Scoring Bands

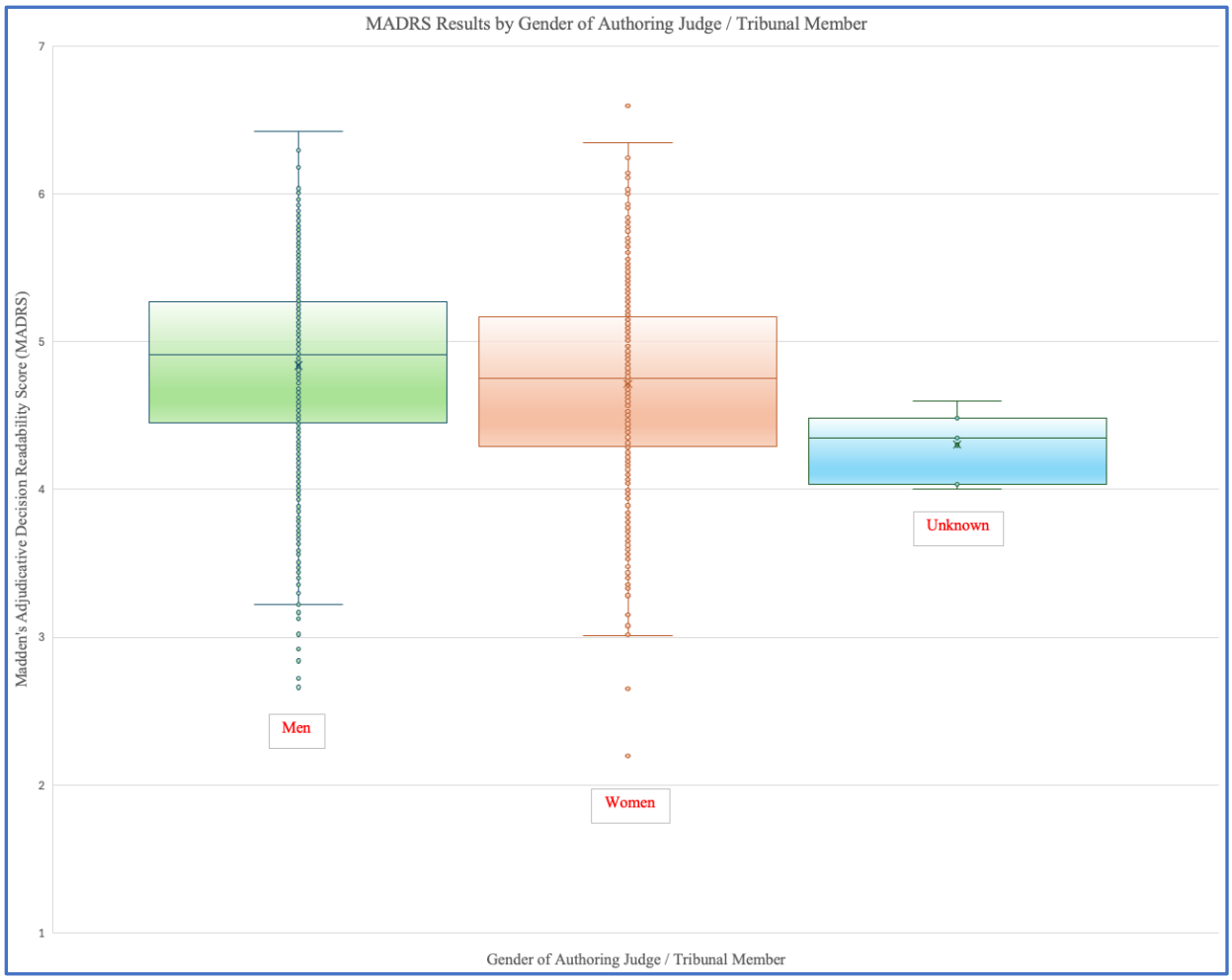


The interquartile MADRS range for these decisions (that is, the scoring range for the middle 50% of the decisions) was between 4.41 and 5.25. With respect to the 7 most readable decisions, 4 were from inferior courts, and 3 were from superior courts. Six of these most readable decisions involved criminal cases, and 1 was a family law decision. With respect to the 15 least readable decisions, there were none from inferior courts, but multiple decisions from every other level (tribunals – 2; superior courts – 4; , appeal courts – 6; and, Supreme Court of Canada – 3). They covered many subjects, but 8 of the least readable decisions were about criminal law.

6.4.2 Gender and Individual/Collaborative Authorship

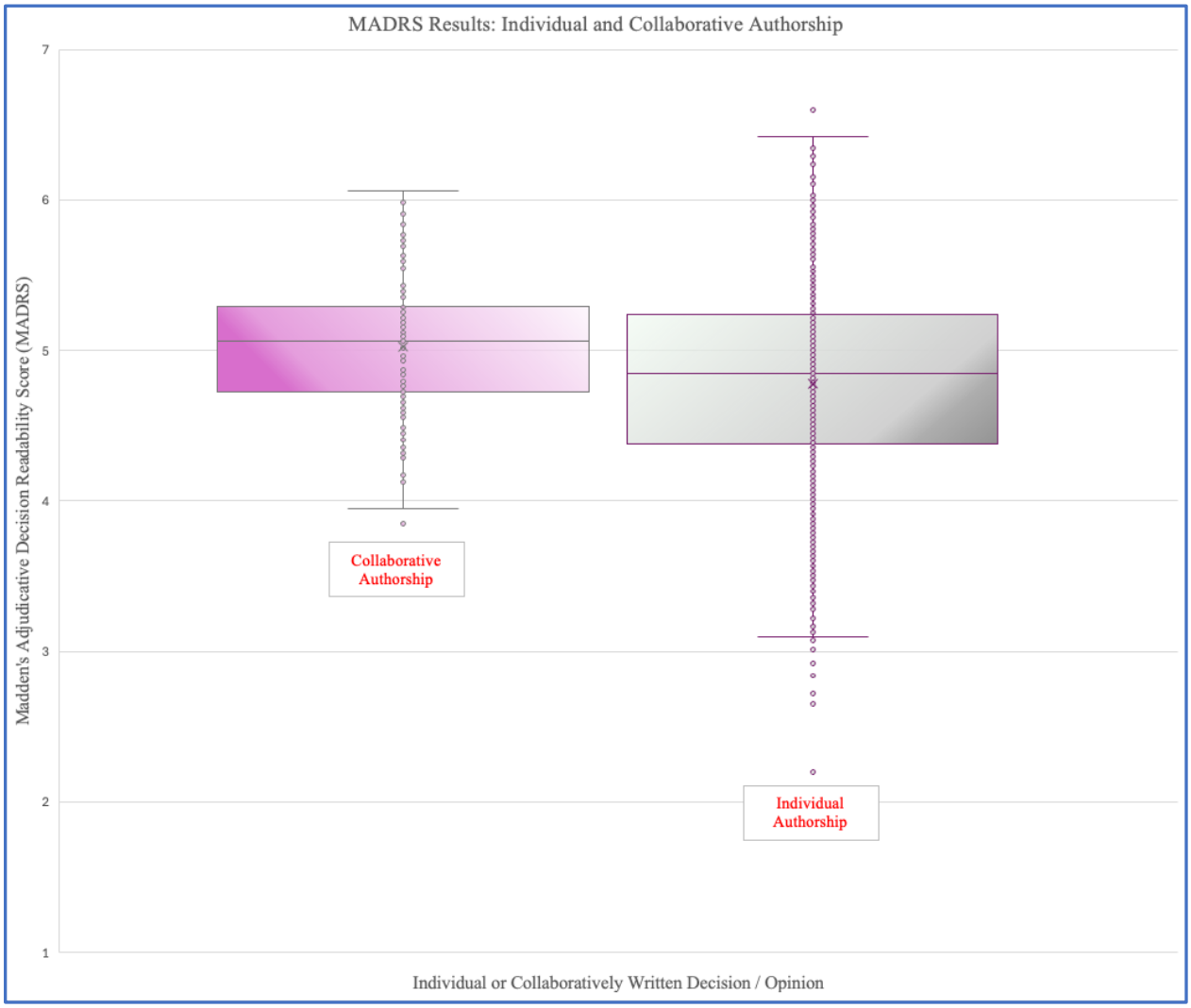
Decisions written by men (N=805) were slightly less readable than decisions written by women (N=698): the average for men was 4.84, and the average for women was 4.71. Only 7 decisions were written by authors whose gender I was unable to identify, and these decisions had an average MADRS value of 4.30. None of the 833 authors whose gender information was available to me identified as anything other than a man or a woman. A visual representation of MADRS results by gender of the authoring judge is shown below, in the boxplots contained in Chart 6.4.

Chart 6.4 – MADRS Results by Gender of Authoring Judge / Tribunal Member



There were also 117 collaboratively written decisions/opinions (that is, decisions written by 2 or more identified authors, "the Court", or some other collective group of decision-makers) within the corpus, and the average MADRS value for these files was 5.03. The 1504 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 6.5.

Chart 6.5 – 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. It would be worthwhile to assess whether this result persists within a purpose-specific study that randomly samples from individually and collectively written decisions to help us better understand the relationships between readability scores and collaborative decision-making and/or decision-writing processes that courts and tribunals might use.

This comparison of results for collectively and individually authored decisions highlights an interesting and unanswered question about causation. As I mentioned previously in Chapter 4, courts of appeal and the Supreme Court of Canada (SCC) tend to decide cases as panels – thereby creating the possibility of collective authorship. Judges of 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. I also noted in Chapter 4 how average MADRS results are lower for inferior and superior courts than they are for appeal courts and the SCC. This data therefore raises the following causation question: are average MADRS values higher for collectively authored decisions within this study because these decisions are only possible from appeal and apex courts that generate higher MADRS results for other reasons? Or, are appeal and apex court average MADRS values generally higher than inferior and superior court values because the former courts tend to have more decisions that are collectively authored, and because collectively authored opinions cause higher MADRS results? Or, perhaps the court level variable and the collective/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 that manifest within the non-random pool of cases that form the basis of the study, which might signal where a closer look (through a study that is designed to expose causative factors) would be warranted.

6.4.3 Scoring Across Jurisdictions

The breakdown of average MADRS result for opinions/decisions based on jurisdiction, and the number from each jurisdiction, is shown below in Table 6.1.

Table 6.1 – 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
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 in this study, 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 that are dealt with by courts and tribunals in the provinces and territories, because of the way in which legislative competence is divided

between levels of government in Canada.[6] But 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 the jurisdiction-related results from this study persist in a future study of randomly selected cases from each jurisdiction, the results might suggest where opportunities for dialogue across courts and tribunals from the different jurisdictions could 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 to identify strategies to improve readability scores across the board, or to pull less readable jurisdictions more toward the readability levels achieved by British Columbia's courts and tribunals.

6.4.4 Scoring Across Legal Subject Areas

The breakdown of average MADRS results for opinions/decisions based on legal subject area, and the number from each subject, is shown below in Table 6.2.

Table 6.2 – 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
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

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

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 generally be 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.[7] It is disappointing to see that decisions in this area of law from within the current study's database are markedly less readable than both the average decision (across all subjects), and than decisions in many other areas of law. Where only one-third of Canada's adult population has a completed undergraduate

[7] Lorne Sossin, "Designing Administrative Justice" (2017) 24:1 Windsor Yearbook of Access to Justice 87 at 94.

degree,[8] the average administrative law decision within these results is likely to be beyond the comprehension of a large majority of Canadians.

Similarly, it is unfortunate that procedural law decisions in this study 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.[9] 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 next do (procedurally) to move their cases forward.[10] Where the average procedural law decision from within the current study is written at a level that requires somewhere between a completed undergraduate degree and a law degree in order for a reader to understand the content of the decision, it seems likely that many litigants will find these decisions to be unreadable.

6.4.5 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. However, before discussing some of the potentially more interesting individual results, a reiteration of my previous caution about how these results must be understood is needed here. The decisions that contributed to the results

---

[8] Statistics Canada. "Table 98-10-0384-01: Highest level of education by census year: Canada, provinces and territories, census metropolitan areas and census agglomerations" (December 9, 2022), online: <https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=9810038401>.

[9] See, e.g, *Chisan v Klooster*, 2022 ABQB 502.

[10] Social Security Tribunal of Canada, "Evolution of Plain Language Decision Writing" (February 28, 2023), at s 5.4, online: <https://www.sst-tss.gc.ca/en/our-work-our-people/evaluation-plain-language-decision-writing>. 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.

described in this chapter 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 1504 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,[11] there were no decision-makers within the current dataset for whom there were more than 13 decisions. A small minority of authors had more than 3 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 being said, several decision-makers with at least 3 results from the current study demonstrated a consistent ability (in those cases) to write highly readable decisions. The most readable average for an author with at least 3 decisions was Jessica Derynck of the British Columbia Human Rights Tribunal (BCHRT). Her average MADRS value was 3.48, based on 3 decision scores. Nothing in Derynck's biography obviously suggests that she would produce more readable decisions than her peers:

> 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.[12]

---

[11] Ronald J. Kruzeniski, who is Saskatchewan's Information and Privacy Commissioner, wrote 23 decisions that appear within the current set of results.
[12] Online: <https://www.bchrt.bc.ca/tribunal/organization/members/#Derynck>.

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 4 decisions within the current study, and a MADRS average of 3.67 for these decisions. Ohler's biography indicates that she "developed and delivered innovative human rights-related educational and mentorship programs […] and has taught international and common law as Lecturer and Adjunct Professor at the University of British Columbia."[13] Perhaps this experience in teaching (explaining hard things to people who are unfamiliar with these things) helped to prepare Ohler to write readable decisions. In any case, it is encouraging to see how readably different members of this tribunal write within this study on a legal topic that has universal relevance to the people of Canada (who all have and presumably exercise human rights every day) – especially when one considers the readability levels reported for decisions from the Canadian Human Rights Tribunal (above, in Chapter 4). In Ohler's case, it is also nice to see some 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

[13] Online: <https://www.bchrt.bc.ca/tribunal/organization/members/#Ohler>.

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 3 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 3 results that averaged to 3.71 – well below the average of 4.37 that was reported for inferior court decisions on the whole in Chapter 4. Judge van der Hoek's biography reveals that she spent time as both a Crown prosecutor and a legal aid lawyer before being appointed to the bench,[14] so she presumably acquired some experience in communicating with people at the lower ends of the socio-economic and educational spectrums in her past work – although I observe (impressionistically) that this path to the inferior court bench is not unusual, and is likely shared by many others who do not necessarily write as readably as Judge van der Hoek.

As I noted above,[15] 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 identify more holistically who writes exemplary decisions. All of these three profiled decision-makers, above, would seem to be good candidates for further study. Regardless of whether they always, or even consistently, write decisions using simple and familiar language, they have shown that they are capable of this kind of writing. The results noted above (particularly if they persist in broader quantitative studies that look at the full populations of their decisions, and/or if they are matched with similar qualitative assessments of

[14] Online: <https://www.courts.ns.ca/resources/news/provincial-court-welcomes-two-new-associate-chief-judges>.
[15] At section 4.2.3.2, in the context of a discussion about Judge Peter West's highly readable inferior court decisions.

writing strength) could imbue these authors with a certain measure of quantitative credibility as strong writers. Perhaps further study involving their decisions, or their personal writing processes and philosophies, could identify practical tips to be shared with other authors about how to write readable decisions. Or perhaps it could lead to further work that asks whether other interested judges or tribunal members can adapt their own writing styles to produce more readable decisions simply by trying to emulate decisions of their peers that have been found to be highly readable. The point here is not to state conclusively how decision-makers or society will benefit from our ability to identify the decisions and authors whose scores (at least some of the time) 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 Justice Charlesworth from the Nunavut Court of Justice. Justice Charlesworth had 4 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.[16] In light of her location in Nunavut, where 57% of the population does not have a completed secondary school (or equivalent) level of education,[17] it would be interesting to see whether Justice Charlesworth'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

[16] Online: <https://www.canada.ca/en/department-justice/news/2018/06/government-of-canada-announces-judicial-appointments-in-nunavut.html>.
[17] Statistics Canada. "Table 98-10-0385-01 – High school completion by census year: Canada, provinces and territories, census metropolitan areas and census agglomerations" (Oct 4, 2023), online: <https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=9810038501>.

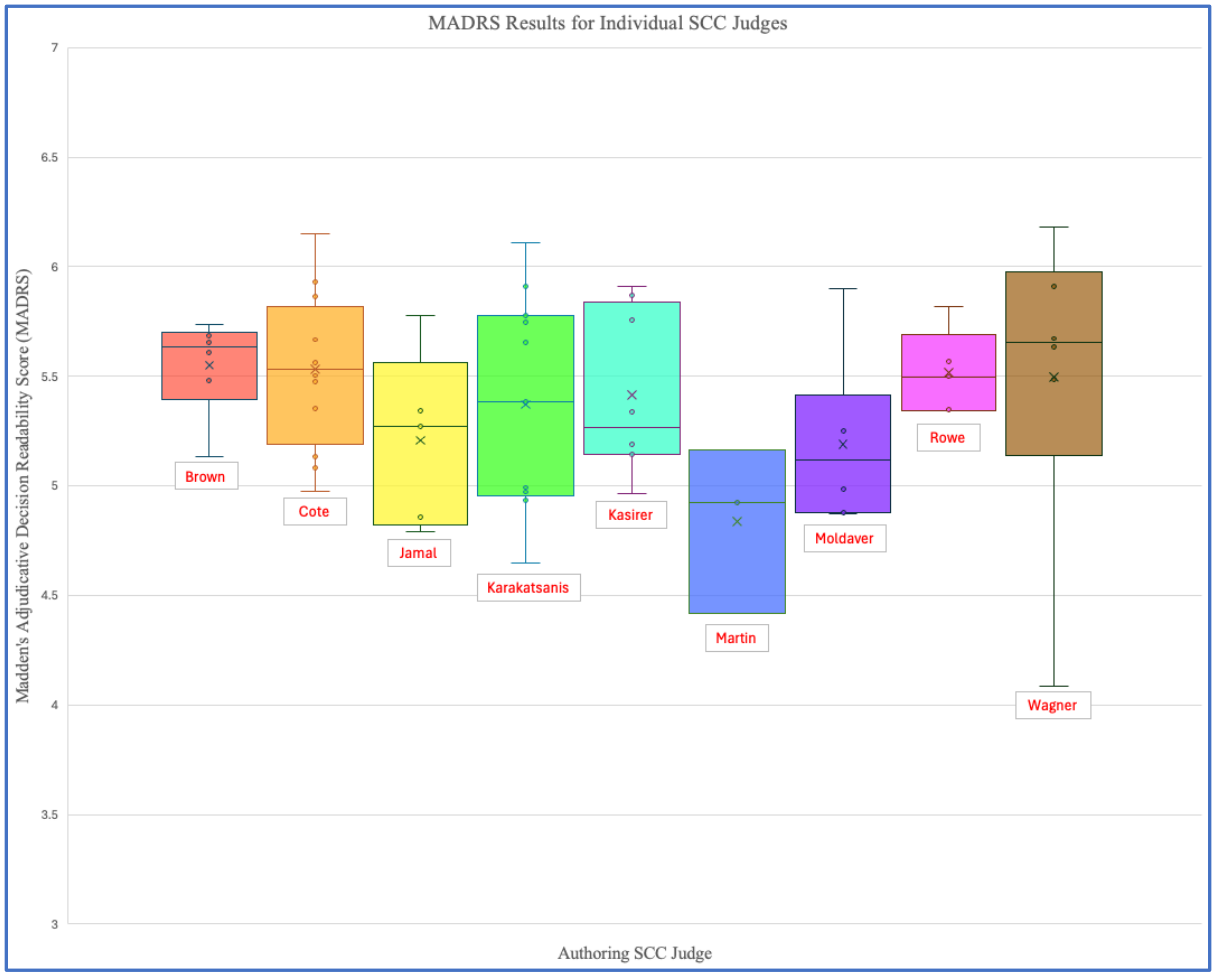
data for this jurisdiction suggests that readability scores ought to be much closer to, or below, 3.0 on the MADRS scale in order for the decisions to be understood by most members of the population there.

Justice Pfuetzner of the Court of Appeal of Manitoba had the second-least readable average decision score for 3 or more decisions within the current study. Her 3 decisions averaged a score of 5.56 on the MADRS scale. Justice Campbell of the Supreme Court of Yukon, and Justice Brown (formerly of the Supreme Court of Canada) were tied with the third-least readable average decision score, coming in at 5.55. Justice Campbell had 4 decisions within the current study, while Justice Brown had 6 decisions. Again, in most of these cases, it would be interesting to see whether these judges maintained their high average MADRS results across all, or a representative large sample, of their decisions. In Justice Brown'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.

6.4.6 Individual Author Scores – Judges of the Supreme Court of Canada

The study within this chapter included all 62 individually written opinions from the SCC. Chart 6.6 shows each judge's scores in alphabetical order from left to right, in the form of boxplots.

Chart 6.6 – MADRS Results for Individual SCC Judges



In the above boxplots, the "x" marks represent each judge's average readability score. These averages are also shown in Table 6.3, below, along with the total number of opinions by each judge in parentheses. The results in this table are sorted from least to most readable average opinion score.

Table 6.3 – Average MADRS Results for Individual SCC Judges

Judge Name (number of opinions in parentheses)	**MADRS Average**
Justice Russell Brown (N=6)	5.55
Justice Suzanne Côté (N=12)	5.53
Justice Malcolm Rowe (N=5)	5.51
Chief Justice Richard Wagner (N=6)	5.49

Justice Nicholas Kasirer (N=8)	5.42
Justice Andromache Karakatsanis (N=11)	5.37
Justice Mahmud Jamal (N=5)	5.21
Justice Michael Moldaver (N=5)	5.19
Justice Sheilah Martin (N=3)	4.84

If one compares the above average readability scores for certain judges with their answers provided on questionnaires submitted as part of the application process for appointment to the SCC, then the relative rankings are not necessarily surprising. Specifically, Justices Rowe, Martin, Kasirer, and Jamal were all appointed after the Canadian government began publishing questionnaire responses from successful applicants for judicial appointments. Each of these SCC judges therefore has shown us their answers to the question, "Who is the audience for Supreme Court of Canada decisions?"

Justice Rowe's response is in many ways the least indicative of a desire to speak directly to the Canadian public through his decisions. He acknowledges that

> Criminal law will remain one of the most closely watched areas of the Court's jurisprudence, given the large number of criminal cases that are always underway and the relatively frequent restatements of law that occur. Beyond the Bar & Bench, there are, of course, the accused, those affected by crime and various advocacy groups. As well, **journalists have found that the public has an appetite for reporting on criminal matters**.[18] (emphasis added)

As this passage suggests, Justice Rowe seems to see members of the general public as – at most – an indirect audience for SCC decisions, who can access these decisions through media reporting about the cases. He also says that, in constitutional equality rights cases, "given the importance of such decisions for how Canadian society sees itself, the audience can readily be the broad

[18] Canada, Department of Justice. *The Honourable Malcolm Rowe's Questionnaire, (Questionnaire for the Supreme Court of Canada Judicial Appointment Process)* (4 July 2017) at Question 4. Online: <https://www.fja-cmf.gc.ca/scc-csc/2016-MalcolmRowe/nominee-candidat-eng.html> [*Rowe's Questionnaire*].

public."[19] He does not mention self-represented litigants in his response, and beyond the two relatively narrow examples cited above, he construes his audience as being quite small. It is perhaps understandable (although not necessarily desirable) that Justice Rowe has the highest average MADRS result from within the group of SCC judges whose answers to the audience question are publicly available. Justice Rowe does not see the public as an audience for many SCC decisions, so he may not make as much effort as others do to linguistically reach people with lower levels of reading competency within his decisions.

Justices Kasirer and Jamal both include the broad public and people with less education as parts of their audiences for decisions. Justice Kasirer suggests that SCC decisions teach the population about the law, and "this function speaks to a specific audience: legal professionals and judges in all jurisdictions, law professors and researchers in related fields, students from high school to the doctorate level, etc."[20] He also implies that members of the public need to be able to understand SCC decisions as a matter of legitimacy: "a proper explanation of a rule not only allows the outcome of the case to be properly grasped, but also gives the public and the media an understanding of the legitimacy of the exercise of public power that every judgment of the Court represents."[21] Justice Jamal similarly affirms that the "public has a right to understand how their highest court is grappling with the most pressing legal issues of the day."[22] He also believes that the decisions should convey meaning to all readers, regardless of whether they are lawyers.[23]

---

[19] *Rowe's Questionnaire, supra* note 18 at Question 4.
[20] Canada, Department of Justice. *The Honourable Nicholas Kasirer's Questionnaire (Questionnaire for the Supreme Court of Canada Judicial Appointment Process)* (18 April 2019), at Question 4. Online: <https://www.fja.gc.ca/scc-csc/2019/nominee-candidat-eng.html> [*Kasirer's Questionnaire*].
[21] *Kasirer's Questionnaire*, *supra* note 20 at Question 4.
[22] Canada, Department of Justice. *The Honourable Mahmud Jamal's Questionnaire (Questionnaire for the Supreme Court of Canada Judicial Appointment Process)* (17 Jun 2021), at Question 4. Online: <https://fja-cmf.gc.ca/scc-csc/2021/nominee-candidat-eng.html> [Jamal's Questionnaire].
[23] *Jamal's Questionnaire*, *supra* note 22 at Question 4.

Both Justices Kasirer and Jamal mention self-represented litigants within their answers about who the audience is for SCC decisions.[24] Where these two judges have turned their minds to the potentially broad need for ordinary members of the population to understand their decisions, there is at least some basis for us to link this consideration with their average MADRS results that are lower than Justice Rowe's average. But this link requires some explanation, and reliance on some inferences that may or may not be valid.

To begin with, one would need to accept my impressionistic assessment that judges generally care about performing their functions to a very high standard and about meeting the needs of those who are affected by their decisions.[25] If judges have considered the reading needs of the broad public, and if they care about meeting those needs, then they may also make efforts to meet those needs by trying to write more readable decisions. But is this enough? Is it possible for someone who wants or tries to write more readable decisions to actually achieve this result?

There is no research showing that desire, alone, is enough to improve the linguistic readability of one's written communications. It is difficult to conceive of how such a study would be structured. But there is research showing that qualitative writing interventions, in the form of training for decision-makers,[26] and readability-checked templates accompanied by training,[27] do

---

[24] *Kasirer's Questionnaire*, *supra* note 20. *Jamal's Questionnaire*, *supra* note 22 at Question 4.

[25] See above, at sections 4.2.3.2 and 4.5, wherein I discuss the basis for my characterization of judges in this manner.

[26] Social Security Tribunal of Canada, "An Evaluation of How Easy it is to Read Decisions of the Social Security Tribunal" (last modified 10 May 2022), online: <https://sst-tss.gc.ca/en/our-work-our-people/evaluation-easy-it-read-decisions-social-security-tribunal> ["Evaluation"]. Within this study that considered – among other things, the impact of plain language training for tribunal members on their subsequent decisions – the authors found that "[t]he median score of all decisions before the most recent language training was 9 out of 17 points. That improved to 11 after training. Before training only 34% of decisions scored between 11 and 17 points whereas 51% scored in this upper range after training."

[27] Kristie B Hadden, et al, "Improving Readability of Informed Consents for Research at an Academic Medical Institution" (2017) 1 Journal of Clinical and Translational Science 361. In this study, the authors reviewed 217 previously approved, investigator-written Informed Consent forms for research participants, and measured the readability of these forms using 3 older quantitative readability formulas. The mean readability score was 10th grade reading level for these forms. The authors then developed and tested a template form that was written at the 5th grade level, that future investigators could complete for their own purposes. Finally, the authors measured the readability of

help to improve quality or readability measurements that are taken after the training is completed.[28] This research suggests that it is possible to improve the readability of one's writing if one learns more about what constitutes understandable writing.

And it does not seem that it would be difficult to acquire this knowledge. The National Judicial Institute (NJI) and the Canadian Institute for the Administration of Justice (CIAJ) jointly run multiday programs twice per year that all new federally appointed judges attend within one year of appointment, and these seminars include sessions on judgement writing and effective communication in the courtroom.[29] The NJI/CIAJ and the Council of Canadian Administrative Tribunals run separate programs (of 4 days and 2 days, for judges and tribunal members, respectively) that are focused exclusively on decision-writing.[30] The Social Security Tribunal has published a Style Guide for its members, which includes advice like, "write how you speak", "use short sentences", "use common words", and, "explain legal wording."[31] These elements of advice are, incidentally, closely related to different MADRS variables that measure elements of a text such as how frequently its language uses the most common trigrams in the COCA "Spoken" sub-corpus, the density of its legal terminology, and the syntactic complexity of its sentences.[32]

---

82 forms that had been written by investigators after the intervention, and found that the mean readability score was 8th grade reading level. They also found that the "intervention has resulted in a 658% increase in the percentage of written informed consents at eighth grade readability or below compared with baseline" (at 363-364).

[28] See Social Security Tribunal of Canada, "Evaluation", *supra* note 26, and Hadden, et al, "Improving Readability of Informed Consents for Research at an Academic Medical Institution", *supra* note 27.

[29] National Judicial Institute, "Course Descriptions", online: <https://www.nji-inm.ca/index.cfm/publications/public-resources/nji-course-descriptions/>, at 6.

[30] Canadian Institute for the Administration of Justice, "Mastering the Skill of Judgment Writing (2024 Edition): Joint program with the NJI", online: <https://ciaj-icaj.ca/en/upcoming-programs/mastering-judgment-writing-2024/>; and Council of Canadian Administrative Tribunals, "Decision Writing Seminar 2024", online: <https://www.ccat-ctac.org/national-administrative-law-week/decision-writing-seminar-june-2024/>.

[31] Social Security Tribunal of Canada, "Style Guide: Social Security Tribunal of Canada Decisions" (15 May 2024), online: <https://www.sst-tss.gc.ca/en/our-work-our-people/style-guide-social-security-tribunal-canada-decisions>.

[32] See above, at section 3.5.1., for a discussion relating to each of the MADRS variables.

It is not illogical, therefore, to think that Justices Kasirer and Jamal may write more readable decisions than Justice Rowe, at least as measured by the MADRS formula, because these two judges have more clearly identified members of the general public as part of the audiences for SCC decisions. Justices Kasirer and Jamal appear to see members of the broad public within their respective audiences. Perhaps this factor has caused the judges to have invested some additional time and energy in learning how to reach members of the public through the language that the judges use in their decisions – in ways that are reflected in these judges' readability scores.

Justice Martin's response to the question of who she sees as the audience for SCC decisions, however, dedicates 5 complete paragraphs to her explanation of how the public is an audience,[33] after spending the first 5 paragraphs explaining (more briefly) how the litigants, lawyers and judges, Canadian governments, academics, and international stakeholders are all audiences as well. Justice Martin's emphasis on communicating to the public is very prominent within her response.

---

[33] Canada, Department of Justice. *The Honourable Sheilah Martin's Questionnaire (Questionnaire for the Supreme Court of Canada Judicial Appointment Process)* (21 December 2021), at Question 4. Online: < https://www.fja.gc.ca/scc-csc/2017-SheilahMartin/nominee-candidat-eng.html>. Excerpts from Justice Martin's long answer are included below:

> In Canada, a constitutional democracy built on the rule of law, the members of the public receive the wisdom from the Supreme Court decisions and incorporate it into the national conversation. Through its decisions, the Supreme Court speaks to the people on matters that lie at the heart of what it means to live in a constitutional democracy, at one end of the spectrum, to matters that lie at the heart of their day-to-day lives, at the other end.
>
> […]
>
> The public, including any victim and/or interested communities, may also have a pressing interest in criminal matters proceeding as of right. Public interest will be high whenever the federal government poses a reference question directly to the Court because this power is used rarely and is reserved for cases that involve fundamental Canadian rights, values, and institutions.
>
> […]
>
> The cases before the Supreme Court often call for a choice between conflicting rights, values, principles and directions. When the correct conclusion on a matter of public importance is contested socially and legally, the 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. Experience and research suggests that people are capable of respecting decisions, even ones they disagree with strongly, when the decision-making process is respectable, rational and legitimate.
>
> […]
>
> [D]ecisions that provide a full explanation to the parties will also inform the wider public. In the end, the decisions of the Supreme Court must demand public respect. It is the way judges ensure public respect for, belief in and a commitment to the rule of law.

This consideration of her audience may inform Justice Martin's approach to her opinion-writing. It may also help to explain why she has the most readable average MADRS decision result of the SCC judges within this study, and why she is the only judge whose average decision requires less than a completed undergraduate degree to understand.

In reality, however, the results shown above in Table 6.3 demonstrate that none of the SCC judges wrote, in 2022, in ways that most of the Canadian population can easily understand. Statistics Canada data from the 2021 census indicates that only 32.9% of the Canadian population between the ages of 25-64 have a "Bachelor's degree or higher."[34] However, 8 of the 9 SCC judges within the present study have average MADRS decision scores of greater than 5.0, indicating that a completed undergraduate degree would be needed to understand their average decisions. What is striking about this information is the obvious discrepancy between what some of the SCC judges say they are trying to do within their decisions (that is, reach their audiences, that often include the Canadian public) on the one hand, and their failure to mobilize language within their decisions in ways that will accomplish their stated goals, on the other hand. The SCC judges' efforts to communicate with their audiences are no doubt sincere, but the results in this study suggest that more work is needed to succeed in these efforts. Perhaps it is actually more difficult than I have suggested (above within this section) for motivated judges to find out how they can write more readable decisions, even though there is an abundance of judicial education on the subject of decision-writing. I return to this idea when discussing the recommendations that flow from this dissertation, in the next chapter.

6.5 Self-Directed Exploration of the Results – Data for the Masses?

[34] Statistics Canada. "Table 98-10-0384-01: Highest level of education by census year: Canada, provinces and territories, census metropolitan areas and census agglomerations" (December 9, 2022), online: <https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=9810038401>.

I have highlighted above some of the results from this chapter'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 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.[35]

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, within the dataset. 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/territorial jurisdictions, or between family law decisions from inferior versus superior courts, from within the dataset. All of these search queries are possible by extracting the relevant subset of coded results from the publicly available results file.

But 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 1621 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 results file. This information might help

[35] Mike Madden, "MADRS_MaddenDissertation_Chapter6Results" (30 May 2024), Harvard Dataverse - V1, online: <https://doi.org/10.7910/DVN/NIQ2KJ>.

authors to 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 to 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 to 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 all be successful strategies that drive change in the subject's writing.[36] Writers can become better at their craft by observing and reflecting upon the models produced by others.

---

[36] *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); and Martine AH Braaksma, et al, "What Observational Learning Entails: A Multiple Case Study" (2006) 6:1 L1 – Educational Studies in Language and Literature 31, especially at 53 (in the context of secondary school students' writing): "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

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 that is 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.

As I note in the following chapter, 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, then it may not be necessary to focus much on producing highly readable decisions.

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

---

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."

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;

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.

Herein lies a scent of what the next chapter contains: an outline of the different ways in which the MADRS formula could be used as part of a larger process to help individual authors become better at writing more readable decisions, among other recommendations. These recommendations will ideally lead to improved readability results in the medium- to long-term, such that more people who would currently struggle to understand most adjudicative decisions that could be of relevance or interest to them might eventually find most of these decisions to be easily understandable. Time will tell whether such changes materialize, 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 potentially not being met.

7. Recommendations: Using MADRS Effectively

7.1 Overview

Chapter 3 introduced *Madden's Adjudicative Decision Readability Score*, a novel readability formula that was shown to outperform several other known formulas when used on adjudicative decision texts. Chapters 4-6 demonstrated how the formula can be used to generate insights about the readability levels of different classes of decisions, and gave us contemporary data about how Canadian courts and tribunals are currently performing on readability fronts. But the work contained within Chapters 4-6 is really just a small example of how the MADRS formula could potentially be used to propel our knowledge forward, or to make access to justice gains in Canada.

In this chapter, I will therefore describe the larger potential that I believe the MADRS formula holds for legal, government, academic, and political communities, by making recommendations about how this tool can be used in different contexts. Where Chapters 4-6 revealed some of the different ways in which MADRS can be used in empirical legal research, this chapter will identify other areas for future readability research and applications using the MADRS formula. This chapter contains my recommendations about how others within Canada's court, tribunal, and other systems or environments can do the same type of work for their own purposes. But before moving ahead with any recommendations, it is perhaps worthwhile here to reflect on what MADRS can and cannot do, and to consider from a broader perspective what might be lost or gained through different uses or misuses of this readability formula.

7.2 Of Diseases and Cures: Should We Emphasize Readability?

Among other things, this dissertation has shown how it is possible to measure the readability of Canadian adjudicative decisions using a law-specific formula. This fact, alone, does

not necessarily point us to the conclusion that we ought to measure and emphasize the importance of readability within these decisions. After all, just because something can be measured does not mean that it should be measured. Before deciding whether it is worthwhile to spend time and energy on the task of measuring (and then presumably doing something useful with) decision readability scores, it is therefore important to first ask ourselves, is there a problem here that needs to be solved? If we determine that a problem exists, then we can think of the problem metaphorically as a type of legal disease. The next logical question from the perspective of a readability scholar would then be, can readability measurements play a role in ameliorating this problem? Or, in other words, is readability part of the potential cure? Finally, we would likely want to assess whether the benefits of measuring and focusing on readability outweigh the costs or adverse consequences of using this approach to study (and ideally fix) the problem: is the cure better or worse than the disease?

Much of the previous content of this dissertation has hinted at, or outright pointed to, a problem that probably exists today: many adjudicative decisions are written in ways that – according to the MADRS formula – would likely make the decisions impossible for some audience members to easily and independently understand.[1] And parts of the dissertation have suggested (up to this point, in relatively general terms) how readability measurements could play a role in improving this problem.[2] Much more specific discussions about how the MADRS formula could be used to assist in correcting for unintended or undesirable readability levels will follow in the

[1] For examples, see above, at section 4.4.2.2 (in the context of discussions about how human rights law decisions are score relatively poorly on the MADRS scale at both the federal tribunal and court level), and at section 5.3.2 (in the context of a discussion of how decisions involving self-represented litigants also score relatively poorly on the MADRS scale).

[2] For example, see above at section 6.5 (for a discussion of how authors might use MADRS scores to identify highly readable decisions that they could study for the purposes of learning – through emulation or observation – how to adapt their own writing to be more readable).

remaining sections of this chapter. Within this section, however, I focus on the as-yet unaddressed question about what might be lost or gained by measuring and emphasizing the importance of quantitative readability scores for the decisions of judges and tribunal members.

In some ways, it is perhaps unusual to situate a discussion about whether it is a good idea to study the thing that has been studied in such depth within a dissertation in a chapter that falls toward the end of the dissertation. But the concept of quantitative readability is generally not widely understood by, or familiar to, those who study law. And it is quite technical, and interdisciplinary, in ways that make readability somewhat difficult to explain to legal audiences without delving into the reading, educational, and linguistic theories that ground the concept. For these and other reasons, I have structured this dissertation in a way that allowed me to first describe the idea of quantitative readability, then demonstrate how it can be calculated in a domain-specific way to measure the readability of adjudicative decisions, and then show what these measurements look like in Canada for different classes of decisions, before finally returning to the question of whether it is useful to dedicate our energy to studies like the ones within this dissertation. The time for answering this last question has arrived.

7.2.1 Possible Reactions to Increased Emphasis on Readability

Let us first imagine, for a moment, a situation wherein Chief Justices and Chairpersons of various courts and administrative tribunals have decided to consider the MADRS readability levels of decisions produced by their respective judges and tribunal members. Let us also imagine that Cabinet Ministers, law and political science professors, and some members of the media have also taken an interest in these readability levels. And, let us imagine that this interest in readability levels has led to pressure on decision-makers to produce more readable decisions, with a view to

narrowing any gaps that exist between the language that is used within specific decisions and the language that most members of the public are able to easily understand.

It is not obvious in this situation how individual judges and tribunal members would react to the focus on readability. We cannot know for certain what might transpire. But we can hypothesize about possible reactions.

A first hypothesis would be that this focus on readability would have absolutely no impact on how judges and tribunal members write decisions. As noted above in Chapter 1,[3] judges operate in a low-accountability environment,[4] where they are not subject to the same kinds of workplace performance reviews or other systems of performance incentives and correctives that are common in most other workplaces. Tribunal members also benefit from a form of adjudicative independence that protects them from undue influence in respect of the decisions that they make.[5] Under these circumstances, any pressure for decision-makers to write more readable decisions may be perceived as irrelevant to these decision-makers. If a decision-maker in this situation believes that the ways in which they write their decisions are protected by judicial or adjudicative independence, then they do not necessarily have any reason to be affected by the pressure.

A second hypothesis would be that this pressure to produce more readable decisions results in widespread effort by judges and tribunal members to achieve this goal. If one accepts my impressionistic assessment that judges generally work hard to do what is expected of them, and

---

[3] At section 1.3.1.

[4] *See* David M Brown, "The Five Habits of Highly Effective and Accessible Civil Courts: Transparency; Perspicacity; Creativity; Simplicity; and Accountability. Reflections at the End of a Judicial Career", *International Access to Justice Forum* (Toronto: 25 Oct 2024) at 10 ["5 Habits"]: "One could make the argument that our constitution has succeeded in making the courts the least accountable branch of our country's government."

[5] See *IWA v Consolidated-Bathurst Packaging Ltd*, [1990] 1 SCR 282 (in the context of a challenge to a consultation practice of the Ontario Labour Relations Board that was alleged to violate adjudicators' independence): "Independence is an essential ingredient of the capacity to act fairly and judicially and any procedure or practice which unduly reduces this capacity must surely be contrary to the rules of natural justice."

care about the needs of people affected by their decisions,[6] then it is possible that judges would take seriously any call for them to write more readable decisions, by dedicating great energy toward ensuring that the language in their decisions will be more widely understood.

There are elements of the first and second hypotheses that are both plausible, but these hypotheses are both relatively absolute or extreme in their descriptions of what might happen. And they are inconsistent with one another. So a third hypothesis that incorporates the plausible elements of each of the first two, and that occupies something of a middle ground between them, is perhaps the most likely to reflect reality. Under this hypothesis, some decision-makers would disregard any pressure to write more readably. Some would dedicate great energy to the task of writing more readably. And some would dedicate a little additional energy to the work of writing readably, because (as these decision-makers might say to themselves), "writing readably is just one of my many obligations in terms of producing good decisions." What the balance between these three classes of decision-makers might be is difficult to estimate.

If the balance leaned overwhelmingly toward decision-makers who simply ignore any pressure to write more readably, then the benefit of studying the readability of adjudicative decisions is questionable. Producing information that will be ignored by the people who have the greatest ability to act upon the information is not, in my opinion, a particularly worthwhile scholarly or professional pursuit.

If the balance showed that a reasonable number of decision-makers were inclined to reconsider how they write their decisions – either to a large or small extent – because of heightened pressure to produce readable decisions, then it could be worthwhile to continue researching this

[6] I have referred to this assessment above at sections 4.2.3.2, 4.5, and 6.4.6.

topic and maintaining the emphasis on readability, if the costs outweigh the benefits. And so we must identify the costs and benefits, and weigh them to see how they compare against one another.

7.2.2 Readability and its Relationship with Decision Quality

One concern with a focus on readability is that this aspect of writing is not the only, or even the most important, consideration for an author of an adjudicative decision. As former Chief Justice of Canada McLachlin has noted, trial judges must clearly set out the findings of fact, and get the law right;[7] appeal judges must set "out the law as clearly as possible, thereby resolving the case justly and giving guidance to lawyers, judges and the public";[8] and, SCC judges must ideally write to achieve consensus, be brief, and settle legal issues of public importance.[9] In the context of administrative tribunals, the Ontario government has published quality standards for the decisions of provincial tribunal members, which include considerations about whether the law was correctly stated and applied, whether the analysis is persuasive, and whether the result of the decision and any associated order is clear – in addition to considerations about style and comprehensibility of the decision.[10] As these examples suggest, there are multiple objectives that authors should try to achieve within their decisions, and multiple criteria that we might use to assess the quality of the decisions.

But there are several other cascading factors that could also be related to the overall quality of a decision. For instance, if the decision is intended to give guidance to lawyers, judges, and the

---

[7] Beverly McLachlin, "Remarks of the Rt Hon Beverley McLachlin, PC, Chief Justice of Canada" (May 16, 2010), *NJI and CIAJ Advanced Judgment Writing Seminar* (Perth, Ontario), online: <https://ciaj-icaj.ca/wp-content/uploads/podcasts/2020/11/podcast1b_16may10-nji-ciaj-judgment-writing-course-copy.pdf>, at 6 ["Remarks"].
[8] McLachlin, "Remarks", *supra* note 7 at 6-7.
[9] McLachlin, "Remarks", *supra* note 7 at 7-8.
[10] Tribunals Ontario, "Guidelines for Reasons Review and Decision Quality Standards" (20 May 2022), online: <https://tribunalsontario.ca/en/practice-directions/>.

public, then perhaps the decision ought to be memorable in some way,[11] so that it – and the guidance that it contains – can be easily recalled by these individuals when they need to apply the relevant guidance. I say that there are cascading factors in play here because a memorable decision might be one that has a certain literary or aesthetic dimension to it, or an element of humour, or bluntness, or some other characteristic that distinguishes the decision from thousands of others. In this sense, the primary objective of providing guidance to others within a decision might generate a need for the decision to be memorable, which – in turn – might require the decision to be aesthetically pleasing, elegant, funny, or something else in terms of its style. In this sense, style and language considerations within a decision flow from other higher-order imperatives for the decision (e.g.: to be memorable, so that it can effectively instruct).

But some people would characterize even these stylistic elements of adjudicative decisions as primary markers of quality, rather than as subsidiary ones. The Right Honourable Sir Frank Kitto (a former Justice of the High Court of Australia), for instance, suggests that decisions should be written with precise and "elegant" English; he laments how an "ability and a concern to express oneself in clear and graceful English are becoming alarmingly rarer even in the universities [due to] the modern neglect of the study of Latin," but he argues that "a demonstration of that ability and that concern still commands and will always command respect."[12] According to Kitto, this elegant style is important because "every piece of distinguished writing in a judgment tends to enhance the prestige of the Judge's court and to enhance the prestige of the law."[13] The suggestion

---

[11] Richard A Posner, "Judges' Writing Styles (And Do They Matter?)" (1995) 62 University of Chicago Law Review 1420 at 1425 suggests that a memorable opinion has more enduring value than a bland opinion: "The sparkling, vivid, memorable opinion is not so chained to the immediate context of its creation. It can be pulled out and made exemplary of law's abiding concerns."

[12] Frank Kitto, "Why Write Judgments?" in Ruth Sheard, ed, *A Matter of Judgment: Judicial Decision-Making and Judgment Writing* (Sydney: Judicial Commission of New South Wales, 2003) 69 at 74.

[13] Kitto, "Why Write Judgments?", *supra* note 12 at 74.

here is that one must write elegantly in order to preserve the reputation of the law and the judiciary. Many people would likely agree with Kitto about how formal or distinguished language can preserve respect for an adjudicative decision.[14] And many others would likely disagree with elements of Kitto's arguments.[15]

Thus, one problem that we encounter in considering what constitutes a high-quality adjudicative decision is the lack of consensus about the criteria for measuring quality. Posner, for instance, says that "[m]any judges and lawyers are disdainful of 'fine' writing. They think it unprofessional, 'literary,' affected, overrefined. They are mistaken."[16] However, Posner also seems to endorse – rather strongly, if not explicitly – the use of what he calls the impure style over the pure style for writing opinions. The impure style (characterized by an avoidance of things like jargon, solemnity, excessive detail, and long quotations, and by the use of a more personal tone),[17] according to Posner, is the minority[18] style of a more thoughtful judge who looks "below the verbal surface of the doctrines that he is interpreting and applying"[19] and who believes "that values (not just 'feelings'), history, and policy are legitimate considerations"[20] for inclusion within a decision. And nothing within Posner's more notorious works of judicial or extra-judicial writing suggests

---

[14] Gerald Lebovits, Alifya V Curtin & Lisa Solomon, "Ethical Judicial Opinion Writing" (2008) 21 Georgetown Journal of Legal Ethics 237 at 308: "Crafting a judicial opinion that is respectful, well-reasoned, factually honest, and carefully written encourages public respect for the judiciary and acceptance of its opinions."

[15] See Brenda Danet, "Language in the Legal Process" (1980) 14 Law & Society Review 445 at 466-467 (recounting debates among scholars about whether to reduce the formality and Latinate content of legal English). See also George Orwell, "Politics and the English Language," (1968), online: <http://www.orwell.ru/library/essays/politics/english/e_polit/>: "Bad writers, and especially scientific, political, and sociological writers, are nearly always haunted by the notion that Latin or Greek words are grander than Saxon ones."

[16] Posner, "Judges' Writing Styles (And Do They Matter?)", *supra* note 11 at 1424.

[17] Posner, "Judges' Writing Styles (And Do They Matter?)", *supra* note 11 at 1430.

[18] Posner, "Judges' Writing Styles (And Do They Matter?)", *supra* note 11 at 1430: Posner suggests that there is only a "handful of impure judicial stylists."

[19] Posner, "Judges' Writing Styles (And Do They Matter?)", *supra* note 11 at 1447.

[20] Posner, "Judges' Writing Styles (And Do They Matter?)", *supra* note 11 at 1448.

that he considers himself anything **other** than a thoughtful judge who believes that values and policy have important roles to play within a judge's thought processes.

Posner's statements therefore contain an apparent contradiction. The impure style that he appears to prefer is described as conversational – the style of one who writes "as if it were for the ear rather than for the eye"[21] – which suggests that the style is intended to come across as something that does not strive to be literary or refined. The impure style does not fall within the parameters of what most people would consider to be fine writing. But it is still Posner's preferred style? Even though those who disdain 'fine' writing in decisions are incorrect? It is difficult to reconcile even this one judge's complex views about how we ought to assess the quality of judicial writing.[22]

These kinds of contradictions can be found elsewhere in debates about what makes for a well-written adjudicative decision. For instance, some commentators "believe that the use of humor and imagery can be effective, and that it adds life to the otherwise rigid format of judicial opinions."[23] These individuals suggest that, "because [these qualities] show the lighter side of the law, imagery and humor further illustrate the inherently human nature of law and of those who are charged with administering it."[24] Others, however, are strongly opposed to the use of these writing devices within judicial opinions, because "[l]itigants consciously place the court in a position of power to resolve controversies; they expect to be treated fairly and with dignity. Humor can defy both expectations."[25] These same commentators categorically suggest that "[h]umor only reduces

---

[21] Posner, "Judges' Writing Styles (And Do They Matter?)", *supra* note 11 at 1430.
[22] To be fair to Posner (whose thoughts are complex even though they are often articulated in simple language), I infer from the larger context of his writing that he would view the skilled use of an impure style by a judge to be an example of fine, but less conventional writing. And he would likely view the unskilled and conventional use of the pure style by a judge to be exemplary of less-than-fine writing, even if many of his peers would disagree with him.
[23] Adalberto Jordan, "Imagery, Humor, and the Judicial Opinion" (1987) 41:3 U Miami L Rev 693 at 699-700.
[24] Jordan, "Imagery, Humor, and the Judicial Opinion", *supra* note 23 at 701.
[25] Lebovits, Curtin & Solomon, "Ethical Judicial Opinion Writing", *supra* note 14 at 252.

the authority of the opinion and the judge."[26] So, depending on who we choose to believe, humour might add to or detract from the written quality of a judicial opinion.

I do not describe these various differences in opinions between judges, or others who profess to know what makes for 'good' writing in judicial decisions, because I am prepared to take a definitive stand on the question. Rather, I simply note the lack of consensus that exists in relation to the criteria through which we evaluate the quality of these texts. And I call attention to the wide range of criteria that can be found, depending on who one looks to for authority on this matter. If we accept that there are many different characteristics of an author's writing that might have an effect on the overall quality of this writing, then it is worthwhile considering how an increased focus on readability might impact upon any of these other dimensions of writing that can potentially affect quality.

If what legal commentators perceive to be elegant writing consists of unique and memorable turns of phrase, reference to Latin maxims (e.g.: "*de minimis non curat lex*") or other technical terms for legal doctrines (e.g.: "issue estoppel"), longer sentences with deeper syntactic structures, and abstract language (perhaps referencing lofty ideals and values, like "dignity" or "freedom"), then any push for more elegant writing of this sort will be in tension with efforts to improve readability. The MADRS formula suggests that more common (less unique) turns of phrase improve readability. Reducing legalese improves readability. Using sentences with simple syntax (typically found in shorter sentences) helps readability. And using concrete, imageable language enhances readability.

As one can likely see from the above characterization of readable language, it is rather constraining to write this type of text. The goal of producing readable decisions requires authors

---

[26] Lebovits, Curtin & Solomon, "Ethical Judicial Opinion Writing", *supra* note 14 at 274.

to sacrifice linguistic freedoms that they might otherwise want to exercise (perhaps in order to render their decisions more profound, memorable, unique, or clever). And, if one takes readability goals to an extreme, then adjudicative decisions could begin to look like an increasingly homogenous form of texts, where each author's individual voice and idiosyncrasies are suppressed in favour of a narrower and narrower linguistic register that relies on only the most basic and well-known words or phrases, and the simplest sentence structures.

In this situation, something important could be lost from the language of these decisions. Recall, above, how one commentator suggested that using humour in a decision can "illustrate the inherently human nature of law and of those who are charged with administering it."[27] Recall, also, how another commentator (in the context of discussions about whether machine judges could ever replace human judges)[28] made the following observation: "human beings share an *umwelt* that allows for mutual recognition, comprehension, and communication."[29] These two authors both seem to suggest that it is important for the humanity of those who are doing the judging to be recognizable to those who are being judged, and to the rest of society. In other words, people who read adjudicative decisions want to know that a case was justly decided by a real person who heard the parties' concerns, and who thought carefully about the proper outcome for the case. If the language of adjudicative decisions becomes homogenously simple through our efforts to make them more readable, then there is a danger that each decision-maker's humanity will cease to be recognizable within these decisions.

---

[27] Jordan, "Imagery, Humor, and the Judicial Opinion", *supra* note 23 at 701.
[28] Above, at section 1.7.
[29] Michael A Livermore, "Rule by Rules" in Ryan Whalen, ed, *Computational Legal Studies: The Promise and Challenge of Data-Driven Research* (Cheltenham, UK: Elgar, 2020) 238 at 257.

Along similar lines, many of the dimensions of a text that make it memorable, or aesthetically elegant, could also be lost if readability pressures drove authors to write more reductively. This consequence could make the texts both less enjoyable to read (for those who are capable of reading and understanding more advanced uses of language), and less enjoyable to write. Such costs are not necessarily troubling ones, since no one has argued that a main function of adjudicative decisions is to use language in ways that produce pleasure in either the reader or the writer; rather, these decisions seem to serve more functional goals related to the needs of the litigants and the legal system. But it is possible that even these functional goals could be undermined if more readable language led to decisions that were less memorable, and therefore less capable of instructing lawyers, judges, and society on the law.

We must also be alive to the possibility that more readable language, consisting (for instance) of frequently-used words and phrases expressed in short and simple sentences without much legalese, will sacrifice legal precision for the sake of readability.[30] As Assy notes, "[f]or the lawyer, clarity is not merely linguistic – in fact, linguistic clarity (simplicity) is subordinate and instrumental to legal clarity (precision) and must not undermine it."[31] And Assy goes on to argue that when law is correctly committed to precision "to coordinate the wide range of social and economical activities in which people in modern society may engage, simplicity is bound to be compromised."[32] In other words, linguistic readability will necessarily suffer when more emphasis is placed on precision, and vice-versa.

---

[30] Rabeea Assy, "Can the Law Speak Directly to Its Subjects? The Limitation of Plain Language" (2011) 38:3 Journal of Law & Society 376 at 390 ["Can the Law Speak Directly to Its Subjects?"].
[31] Assy, "Can the Law Speak Directly to Its Subjects?", *supra* note 30 at 392.
[32] Assy, "Can the Law Speak Directly to Its Subjects?", *supra* note 30 at 393.

But Assy's assertion is contradicted by other recent research which seems to suggest that legal texts are performatively drafted in unnaturally complex ways for reasons that are unrelated to communicative efficiency.[33] In this study, the authors tasked participants to draft legal rules that prohibited a crime, and a story that involved the commission of the same crime, in order to assess whether the participants used more complex language in the former text than in the latter text.[34] This experiment tested the authors' "magic spell hypothesis", wherein it is "posited that lawyers and lawmakers write in a convoluted manner in order to lend legal documents a ritualistic, spell-like element."[35]

The authors found that "people tasked with drafting laws wrote in a more convoluted manner than when tasked with drafting various control texts of plausibly equivalent conceptual complexity."[36] The authors went on to suggest that "these results add to an emerging body of literature demonstrating that the language of legal documents can be simplified without a loss or distortion of legal content," and:

> Jurists have long acknowledged the tension between the doctrinal mandate that laws be understandable to the common person and the observation that laws are not understandable to the common person. Whereas recent proposals to resolve this tension have taken for granted the necessity of law's complexity and have called for scaling back the mandate that laws be accessible to the common person, our results suggest such compromises may not be necessary. Instead, our results indicate that lawmakers can faithfully comply with this mandate while simultaneously preserving the desired level of conceptual complexity.[37]

This research suggests that the use of complex legal language – at least in tasks involving the drafting of laws – is a choice, not a necessity. Perhaps the same suggestion would apply to the

---

[33] Eric Martínez, Francis Mollica & Edward Gibson, "Even Laypeople Use Legalese" (2024) 121:35 Proceedings of the National Academy of Sciences e2405564121.
[34] Martínez, Mollica & Gibson, "Even Laypeople Use Legalese", *supra* note 33 at 3-4.
[35] Martínez, Mollica & Gibson, "Even Laypeople Use Legalese", *supra* note 33 at 3.
[36] Martínez, Mollica & Gibson, "Even Laypeople Use Legalese", *supra* note 33 at 6.
[37] Martínez, Mollica & Gibson, "Even Laypeople Use Legalese", *supra* note 33 at 6.

drafting of adjudicative decisions. If so, then writing more readable decisions would not necessarily require a sacrifice in terms of precision within these decisions.

Furthermore, it is also worth considering whether an increased emphasis on producing readable decisions might actually lead to the production of decisions that are less comprehensible to their audiences, essentially proving the accuracy of Goodhart's law.[38] Davison & Kantor concluded from their study (qualitatively assessing changes that were made to texts explicitly to improve the texts' readability scores) that "the most successful changes in the text often run directly counter to what readability formulas would suggest, and that the most unsuccessful changes are those motivated by the strictures of the readability formula."[39] Charrow & Charrow reached a similar conclusion about jury instructions that had been rewritten in ways that were designed to, and actually did, increase real mock jurors' abilities to understand the instructions: two of the rewritten instructions that were actually understood better by the human participants in the study scored worse on Flesch's readability scale than the original versions, and another one scored the same on the readability scale despite being significantly better-understood by the mock jurors.[40] To be fair, for half of the rewritten jury instructions within Charrow & Charrow's study, better human comprehension levels were reflected (as readability scholars would expect) in improved Flesch readability scores for those instructions.[41] But this effect was not uniform across

---

[38] Goodhart's law suggests that the more one strives to achieve a certain target that represents a proxy for an objective that one actually wants to achieve, the weaker the relationship becomes between the proxy and the actual objective. *See* Tamara Goriely, "The English Legal Aid White Paper and the LAG Conference" (1996) 3:3 International Journal of the Legal Profession 353 at 358-59. In the context of quantitative readability formulas, the Goodhart's law implies that, when a writer tries to improve readability levels by targeting the variables that contribute to the formula, the writer may actually be making the text more difficult (rather than less difficult) to understand.

[39] Alice Davison & Robert N Kantor, "On the Failure of Readability Formulas to Define Readable Texts: A Case Study from Adaptations" (1982) 17:2 Reading Research Quarterly 187 at 191 ["On the Failure of Readability Formulas"].

[40] Robert P Charrow & Veda R Charrow, "Making Legal Language Understandable: A Psycholinguistic Study of Jury Instructions" (1979) 79:7 Columbia Law Review 1306 at 1340-1341 ["Making Legal Language Understandable"].

[41] Charrow & Charrow, "Making Legal Language Understandable", *supra* note 40 at 1340.

all of the rewritten instructions. The point here is that efforts to improve a text just by changing its language to score better on any particular readability formula may actually decrease the comprehensibility of the text, because these changes would be made "without a view of the entire text as connected discourse."[42] Readability 'improvements' that are made in unskilled or mechanical ways could be counter-productive.

Finally, we must consider the realistic possibility that efforts to produce more readable decisions will result in less timely decisions. One dimension of the quality of a decision is arguably the speed at which it is provided to the parties, since they cannot find closure and a way forward from their dispute or situation until the adjudicative process is completed, and a final decision is issued. If pressure to produce more readable decisions were to lead to longer delays before decisions are issued (because decision-makers need to spend additional time editing and calibrating the language of their decisions to better match the reading needs of the audiences), then access to justice would suffer. And it is almost certain that such delays would occur in at least some cases, since there are time costs associated with most types of systemic changes within a justice system – at least until the change becomes fully adopted, and new procedures or processes become routine. It is difficult, and arguably not important for the time being, to quantify what additional delays might flow from efforts to make adjudicative decisions more readable. The point is that there is likely to be some additional delay in some cases, and this cost must be accounted for when assessing whether to pursue improved decision readability levels.

In short, pressures to improve readability within adjudicative decisions will likely have opposite effects on at least some other indicators of quality for the decisions, or perhaps even on ultimate comprehension of the decisions. Whether this will be a problem depends on the following

---

[42] Davison & Kantor, "On the Failure of Readability Formulas", *supra* note 39 at 207.

considerations: (1) whether one accepts that criteria X is actually a valid dimension of a decision's overall quality; (2) whether criteria X is actually likely to suffer when readability increases; and (3) the relative weight that one assigns to criteria X and readability in terms of their contributions to a decision's overall quality.

As the above discussion about elegance, humour, and other dimensions of decision-writing has illustrated, however, consensus has not been achieved in terms of identifying all of the various factors that make the writing within a judicial decision 'good'. Furthermore, it is often debatable whether any particular factor is inversely related to readability. For instance, even though Assy suggests that linguistic simplicity and legal precision are inevitably in tension with one another,[43] both Adler and Duckworth assert the opposite point – that plain and simple language leads to greater precision and certainty within a text.[44] And most authors are likely to assign at least subtly different weights to the variables that they feel contribute to the quality of their writing. So it is challenging to propose any universal answers to questions about whether an increased emphasis on readability will create other problems within the writing of an adjudicative decision. The best answer that can be given is this: probably. Better readability will probably create other problems within a text for at least some authors or some texts, at least some of the time. But the magnitude and frequency with which these problems might manifest is unclear.

From a personal perspective, I tend to believe that it is more important for an adjudicative decision to be understandable than elegant or profound (although I accept that these latter attributes can also be markers of a decision's quality, and that they can, at times, be in tension with

---

[43] Assy, "Can the Law Speak Directly to Its Subjects?", *supra* note 30 at 392.
[44] Mark Adler, "The Plain Language Movement" in Lawrence M Solan & Peter M Tiersma, eds, *The Oxford Handbook of Language and Law* (London: Oxford UP, 2018) 67 at 71; Mark Duckworth, "Clarity and the Rule of Law: The Role of Plain Judicial Language" in Ruth Sheard, ed, *A Matter of Judgment: Judicial Decision-Making and Judgment Writing* (Sydney: Judicial Commission of New South Wales, 2003) 91 at 94-95.

readability). And, like both Craig and Varsava,[45] I tend to believe that it is improper to privilege a decision-maker's personal voice (whether through humour, the language of crime fiction novels, or some other element of personal linguistic style) over the more institutional voice of the court or tribunal that the decision-maker represents. These are value choices. They cannot be objectively defended as being correct or proven incorrect. They simply reflect the weights that I would assign to potentially competing variables that may have an impact on the overall written quality of adjudicative decisions. And because of my value choices in these respects, I will generally place more importance on the use of language that makes a decision more readable than the use of language that increases a decision's written quality in terms of some dimension other than readability. But as all of the above discussion indicates, there are many reasons why we should carefully consider what is likely to be lost and gained, in terms of the written quality of an adjudicative decision, if we decide to pressure decision-writers to improve their readability scores.

7.2.3 Readability Pressure and the Losses or Gains Felt Beyond the Texts of Decisions

Regardless of whether any new pressure on judges and tribunal members leads to overall gains in the written quality of the decisions from these authors, it is important to consider how this pressure might have consequences that extend beyond just the texts of decisions. The decisions are produced by real people, with real emotions, preferences, and prejudices. If the expectations about how they are supposed to produce their decisions change, then these decision-makers might not all react in uniformly positive or productive ways.

Let us first consider an exaggerated and unlikely hypothesis, wherein all decision-makers respond similarly to new pressures to write more readable decisions by placing additional, but not

---

[45] Nina Varsava, "Professional Irresponsibility and Judicial Opinions" (2021) 59:1 Houston Law Review 103. *See also*, Elaine Craig, "Judicial Audiences: A Case Study of Justice David Watt's Literary Judgments" (2018) 64:2 McGill Law Journal 309.

undue, weight on readability considerations when drafting their decisions. If these decision-makers carefully considered how improving readability might also cause some other marker of the written quality of their decisions to suffer, and weighed these competing considerations carefully to achieve an appropriate balance within the language of their decisions, then this situation probably does not pose any major concerns for the decision-makers, or for society.

At the other extreme end of potential ways in which a heightened emphasis on readability might have effects beyond the texts of adjudicative decisions, it is possible that all decision-makers would react defensively, aggressively, dismissively, or in some other non-productive (from the perspective of those who created the new readability pressure) way to the expectational changes for their decisions. Perhaps some of them would deliberately write in less readable language to demonstrate their independence from perceived bureaucratic interferences in their work. Perhaps some of them would quit the profession of adjudicating – believing themselves to be either unable or unwilling to write in the new way. Potential applicants for adjudicative positions might be similarly deterred from applying, if they believe their strengths to lie in knowing and correctly applying the law, not in packaging their decisions into highly understandable language. In this way, the profession of adjudicators may begin to lose members who are excellent decision-makers in all respects other than in communicating their decisions readably (or, ones who are excellent in all respects, but who have no interest in being told how to write their decisions). All of these potential consequences would be harmful to the legal system and to society in at least some ways.

But again, neither of these extreme scenarios would be likely to manifest. It seems far more likely that only a small minority of judges and tribunal members (if any) would deliberately try to frustrate a broader movement to write more readable decisions. And some may leave the profession, but it is unlikely that many would leave for this reason alone; readability pressures may

instead just be the final irritant that causes an already-irritated decision-maker to quit or retire. Similarly, some otherwise-strong candidates might decide never to apply to adjudicative positions because they do not have the ability or desire to write readable decisions, or because they do not want to be directed or pressured into performing their functions in any particular way by those outside of the profession of judges or tribunal members.

All of these possibilities would probably represent a loss of some sort – a cost to be borne by society as a result of the new readability pressures. These costs would be more than negligible, in the sense that they could include worse readability outcomes in some cases, and economic costs (e.g.: in the form of selection and training costs for new judges who are needed to replace any who leave) in other cases. These and other similar, realistic consequences of pressuring judges and tribunal members to write more readably should be considered when deciding whether it is worthwhile to begin studying and emphasizing decision readability levels.

But these adverse consequences of promoting better readability levels for adjudicative decisions do not automatically mean that the project should be abandoned. Again, the key question is whether the benefits outweigh the costs. Losing a good judge, for instance, is an unfortunate cost. But if one element of what it means to be a good judge is now (in a world where society has put pressure on decision-makers to write more readable decisions) considered to be an ability and willingness to write decisions that the intended audience would probably understand with ease, then perhaps the loss of judges who do not have this ability and/or willingness is not so consequential (other than for these judges) after all. If the judge can be replaced by another suitably qualified individual[46] who also has a willingness and ability to write more readable decisions,

[46] This does not seem like it would be a problem. Between October 2022-October 2023, for instance, 73 of the 315 candidates who were assessed for suitability as federally appointed judges were "highly recommended" for appointment. Only 68 new appointments were made during that same period – suggesting that, even if every appointed

without too much economic or time cost, then the overall value proposition is not necessarily a negative one. I would characterize it as a net gain for the system.

The focus thus far has been on potential costs. On the benefits side of the equation, pressure to produce more readable decisions is likely to have at least some effect, for some authors, in some decisions, that actually yields decisions capable of being easily understood by more people. As I have noted elsewhere in this dissertation,[47] we have no easily obtainable way of knowing how many people try and fail to read and understand adjudicative decisions, or how many people never make the effort to try reading adjudicative decisions because they believe that these decisions would be beyond their comprehension, so it is difficult to quantify the magnitude of a benefit that would arise from improved readability levels. We do know that the number of audience members for any adjudicative decision is almost certain to be far greater than the number of authors for a decision (given that the highest possible number for this latter group in Canada is currently 9 – the number of judges of the SCC who could jointly write a unanimous decision for that court).

On a simplistic level, then, it seems as if the benefits from improved readability levels for adjudicative decisions would likely flow to a greater number of people than the number of people who would be adversely affected by associated costs. After all, the above discussion of costs highlights how most of the potential costs would likely be borne by the authors of decisions, who must likely take additional time to write more readably, or whose preferences and individuality may suffer from the use of a more readable writing style.

---

judge had received the highest recommendation of "highly recommended," there would still be a surplus of highly recommended candidates. And, an additional 68 candidates were "recommended" for appointment. See Office of the Commissioner for Federal Judicial Affairs Canada, "Demographic Statistics on Diversity in the Judiciary" (6 Mar 2024), online: <https://www.fja.gc.ca/appointments-nominations/StatisticsCandidate-StatistiquesCandidat-2023-eng.html>.

[47] Above, at section 4.2.3.1.

I do not mean to suggest here that a completely utilitarian calculation involving measurements of the number of people who benefit from a change, from which the number of people who lose from a change is subtracted, is the appropriate way to determine whether the costs of focusing more heavily on readability outweigh the benefits. Really, neither the costs nor the benefits can be easily quantified – which makes it hard to determine how they compare against one another. But even with this reality in mind, it seems as if more people would benefit than lose, and there is some comfort for readability advocates in this information.

In the end, however, it must be acknowledged that different conclusions will likely be reached on this question depending on the value judgements that one makes. I have set out my own values above, in favour of comprehensibility over elegance and other stylistic preferences (where they are in conflict with comprehensibility), and in favour of the institutional voice over the individual voice. These value judgements cause me to weigh concerns about authorial preferences, aesthetic pleasures of reading and writing, and satisfaction levels of adjudicative decision-makers somewhat more lightly than I weigh concerns about achieving functional success in communicating the contents of a decision to wide audiences of differing reading competence levels.

But this weighing is not always relevant, since (as the above discussion has shown) there will often be no conflict between these different considerations. In such cases, there is no reason not to continue studying, and advocating for improved, readability levels. In other cases, where conflicts exist, one must choose what considerations are more important. I have outlined above how there are rational reasons to prioritize many different considerations, but I have also indicated why I find that high readability levels should be a paramount objective for adjudicative decisions.

There are compelling reasons to continue studying, and emphasizing the importance of, highly readable decisions.

7.3 Recommendations for Researchers: Future Studies

I now begin considering more specifically how the MADRS formula might be used for productive purposes. The types of research questions that can be asked and answered by using the MADRS formula are extremely broad. The recommendations that follow suggest a variety of different research projects that would each add to our collective knowledge of how courts and tribunals operate, and may also help us to better understand where and how to use readability-based strategies within courts, tribunals, and society more generally, with a view to generating either more readable decisions (in many cases), or other appropriate substitutes for more readable decisions.

7.3.1 Study Readability Across Provincial / Territorial / Federal Jurisdictions

Although the results within Chapter 6 are coded to indicate the jurisdiction from which each decision emerged, and although this data suggests that there are appreciable differences in readability levels across jurisdictions, the cases studied within that chapter were not randomly selected and are not necessarily representative of the broader populations of cases from each jurisdiction. Nonetheless, the results suggest that some jurisdictions are consistently producing more readable decisions than other jurisdictions. It would be worthwhile to explore whether this finding holds up within a larger and more statistically robust future study.

Knowing which jurisdictions tend to produce the most readable decisions might be relevant information for the various Attorneys General or Ministers of Justice who tend to appoint judges

within the different jurisdictions,[48] and who are responsible for the administration of justice within their jurisdictions – particularly if the results show that decisions from their jurisdictions are not as readable as decisions from other places. This is particularly so in light of the fact that education and reading competency levels vary across jurisdictions. In Nunavut, 57.2% of the population does not have a completed secondary school, or equivalent, level of education; in Ontario, this figure is 16.7%.[49] Any readability data that is not disaggregated by province or territory is therefore of limited use in helping us to understand the extent to which the people of Nunavut, or Ontario, are likely to easily understand a decision, given how different the education levels are between these two jurisdictions. The information from a study like this might also expose whether there are local or cultural factors that influence how judges and tribunal members write their decisions, and that could potentially be transplanted from higher-performing jurisdictions to lower-performing jurisdictions, if readability performance improvements were a priority for the lower-performing jurisdictions.

7.3.2 Assess Whether Readability Relates to Being Upheld on Appeal / Review

Much of the dissertation up to this point has focused on readability measurements as something that could be used to (indirectly) benefit readers of adjudicative decisions. However, decision-writers might also have an interest in knowing whether the readability levels of their decisions correlate with being either upheld or overturned on appeal or judicial review. It is generally accepted that decision-makers strive to avoid being overturned by higher courts or

---

[48] Provincial or territorial governments have the authority under s 92(14) of *The Constitution Act, 1867*, 30 & 31 Vict, c 3, to appoint judges of inferior courts in their jurisdictions, while the federal government has the authority under ss 96 and 101 of that Act to appoint superior court judges, and judges of other federally created courts, like the Federal Court, and the Supreme Court of Canada.

[49] Statistics Canada, "Table 98-10-0385-01 – High school completion by census year: Canada, provinces and territories, census metropolitan areas and census agglomerations" (Oct 4, 2023), online: <https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=9810038501>.

tribunals: "[t]rial judges are also likely impacted by loss aversion through their desire to avoid having cases overturned on appeal. Trial judges do not want to risk their cases being lost on appeal, cast into doubt, and potentially raising a cloud of uncertainty concerning their other disposed cases."[50] And, previous empirical research from a related context found judges from both trial and appeal courts in the United States have strong and clear preferences for reading counsel submissions that are written in plain language rather than more complicated legal language.[51] Furthermore, Spencer & Feldman found in their American empirical study of the relationship between the readability of briefs submitted in motions for summary judgement that more readable written briefs were significantly associated with successful litigation outcomes, especially at the federal court levels, even after controlling for other potentially confounding variables.[52]

Where previous research like the two studies above have suggested that there is a meaningful relationship between the readability of a legal text and the ability of that text to persuade a higher-level judge of its correctness or reasonableness – at least in the United States – first-line adjudicative decision-makers in Canada might benefit from a study that looks more specifically at the readability of decisions in relation to the ways in which those decisions are treated on review or appeal in this country.

7.3.3 Praising and Shaming – MADRS as a Performance Appraisal Tool

---

[50] Theodore Wilson, "The Promise of Behavioral Economics for Understanding Decision-Making in the Court" (2019) 18:4 Criminology & Pub Policy 785 at 793.
[51] Sean Flammer, "An Empirical Analysis of Writing Style, Persuasion, and the Use of Plain Language" (2010) 16 Legal Writing 183 (finding that American judges, including appellate judges, prefer plain language over legalese, and observing (at 212) that "[j]udges prefer the Plain English style so much that they would rather have litigants submit informal pleadings, filled with contractions and first person, than formal Legalese").
[52] Shaun B Spencer & Adam Feldman, "Words Count: The Empirical Relationship between Brief Writing and Summary Judgment Success" (2018) 22 Legal Writing: Journal of the Legal Writing Institute 61.

Judicial independence in Canada strongly protects judges from partisan vulnerability, and from non-legal pressures to decide particular cases in particular ways.[53] Most administrative tribunal members in Canada benefit from similar – but less powerful – forms of adjudicative independence that protect the members from improper influences when it comes to deciding their cases.[54] In other words, tribunal members have functional independence in their decision-making, even if they do not have the same security of tenure, financial security, and administrative independence as judges.

One possible drawback of this independence that protects judges and tribunal members is that it may shield them from appropriate social, political, or professional pressures to perform their jobs better (to the extent that we can convincingly state the criteria that contribute to "better" job performance). Strong independence may create weak motivation to do things – like write decisions – differently than how one currently does those things. After all, if one is professionally secure regardless of how well one performs their job according to someone else's standards, then a person's calculations about what might be lost and gained by adapting their behaviour will be different than in more typical workplace cases where an employee or contractor must generally perform up to others' expectations in order to continue earning their income.

A comprehensive study of individual judges' or tribunal members' readability scores using the MADRS formula may represent one way to incentivize the writing of decisions that are likely to be more easily understood by wide audiences, and disincentivize the writing of decisions that

---

[53] See, generally, *Reference re Remuneration of Judges of the Provincial Court (PEI)*, [1997] 3 SCR 3.
[54] See, generally, *IWA v Consolidated-Bathurst Packaging Ltd*, [1990] 1 SCR 282, *per* Gonthier J (for the majority)(for an overview of the idea of the independence of Canadian tribunal members, and a ruling that a practice of having non-binding consultative discussions among all tribunal members before a 3-person panel issued a ruling in a particular case did not breach the panel members' independence).

cannot easily be understood by these audiences. As I noted earlier in this chapter,[55] there are risks associated with any movement to build this kind of readability pressure upon judges. But it may nonetheless be worth doing so, if these risks are outweighed by the potential rewards. If authors know that their readability scores may be calculated and published, then perhaps their own assessments as to what might be gained (like pride, job satisfaction, reputational gains, access to justice improvements, etc) by writing very readable decisions, or what might be lost (credibility, respect among peers and members of the legal community, higher-level appointment opportunities within the judiciary, reappointment to one's tribunal at the end of an initial term, etc) by writing very unreadable decisions may cause these authors to dedicate more time and energy to the task of writing for their audiences.

But the incentives and disincentives are not necessarily just personal ones. They may be institutional as well. Perhaps a study like the one described above could motivate both individual authors and entire courts/tribunals to perform better at writing decisions that their audiences can understand. For instance, if institutional actors or processes influence things like how much money courts and tribunals receive for education and training about decision-writing, a well-structured readability study that exposes high- and low-performing courts might help to create pressure within these institutions to either add or shift some focus onto the goal of writing more readable decisions.

My own research objective in writing this dissertation was not to praise or shame particular authors or courts/tribunals for their readability performance – it was simply to draw attention to questions about readability levels, and to provide data and tools that allow all of us to consider decision readability levels with greater precision. However, it is clear that other studies using the

[55] See above, at section 7.2.

MADRS formula could be used for such purposes, by highlighting whether, according to this formula, the decisions are likely to be easily understood by particular audiences.

7.3.4 Beyond Decisions – Testing MADRS on Other Legal Documents

As I noted above, in Chapter 3, the MADRS formula was derived from a corpus of text files extracted from real Canadian adjudicative decisions that were published in 2021. Its reliability has been validated on an independent set of adjudicative decisions texts. But, the formula has not yet been positively established as performing well on other types of legal texts, like legal information websites, factums, opinion letters, or law journal articles. The broader legal community (beyond just adjudicative decision-makers) might stand to benefit from a readability formula that accurately measures how well other legal documents are likely to be understood by different audiences for many of the same reasons that have been explored throughout this dissertation. So, a study that tests MADRS' performance on other different types of legal texts could be useful in expanding the scope of MADRS' legitimate application across a wide range of contexts.

Regression formulas, like MADRS, are not necessarily reliable when used for extrapolation purposes – that is, on data that falls outside of the scope of the formula's original data.[56] One relatively simple way to explore MADRS' validity on different types of legal documents would involve calculating scores for each of the 6 linguistic component-variables that contribute to the MADRS formula for each text in a new set of non-decision (but still legal) text files, to see if the scores for these 6 variables still fall within the range of scores that existed in the initial 250 text files from which MADRS was derived. If the scoring ranges for each variable in the new set of

[56] See, generally, Gerald J Hahn, "The Hazards of Extrapolation in Regression Analysis" (1977) 9 Journal of Quarterly Technology 159.

documents does not exceed the ranges within the 250-file "formula development" set of files (used in Chapter 3), then MADRS should perform effectively on the new set of files. As long as the language properties (as represented by the scoring range for each MADRS variable) for the new set of documents falls within the scoring range for the formula development set of documents, then the formula would remain valid for use on the new set of documents.

Another way of approaching the same problem could involve measuring MADRS scores for a group of legal documents that are not adjudicative decisions, and testing this formula against human-generated scores for the same documents that use the same MADRS scale. If the correlations between MADRS and human-assigned scores are statistically significant and strong ($r > 0.5$), then MADRS might be assessed as sufficiently reliable for use on such other documents. If the correlations are even stronger ($r > 0.756$),[57] then MADRS would perform better in predicting the human scores of these other documents than it performed within this dissertation at predicting human expert scores for adjudicative decision text files.

In order to facilitate this type of future study that could extend the MADRS formula's uses to new types of legal writing contexts, I have published an openly accessible spreadsheet of the linguistic variable scores for the variables that contribute to the MADRS formula, for each of the 250 files that were used to create the formula.[58]

At this point, however, it must again be acknowledged that specialized and domain-specific readability formulas are not necessarily the best solution for all purposes, due to the time and energy costs associated with the development and use of these formulas.[59] Although the MADRS

[57] See above, at section 3.4.4.
[58] Mike Madden, "MADRS_MaddenDissertation_Chapter3_LinguisticVariableScores250Files" (May 30, 2024) online: <https://doi.org/10.7910/DVN/NCQYJ1>, Harvard Dataverse, V1.
[59] See above at section 4.2.3.4.

formula might perform well on a range of legal texts beyond just adjudicative decisions, it is not obvious, *a priori*, that the formula will outperform general-purpose readability formulas on these documents, or – if MADRS does outperform such formulas – that the difference in performance is significant enough to justify spending extra time and energy to validate and use the specialized MADRS formula on these texts. After all, an off-the-shelf readability solution already exists in the form of the Flesch Reading Ease and Flesch-Kincaid Grade Level formulas, which can be quickly and easily calculated for any document using Microsoft Word's "Editor" feature. A cost-benefit analysis (or estimate) should therefore be performed to gauge whether the anticipated benefit of using MADRS on different classes of legal documents is worth the effort. In Chapter 4 (at section 4.2.3.4), I concluded that MADRS offered performance gains over other readability formulas – when these formulas were applied to adjudicative decision texts – that justified the time and energy spent in developing the MADRS formula. But a different conclusion may be reached when one considers using MADRS on other types of legal texts.

The caution that I have offered here about efforts to extend MADRS' applications beyond the domain of Canadian adjudicative decision texts exposes a broader dilemma for researchers, and a tension between values of specificity/accuracy, on the one hand, and universality, on the other hand. Spending time and energy to develop a readability formula that works well in one narrow context (in this case, in the context of Canadian adjudicative decisions) is not necessarily as useful as spending the same time and energy to develop a less specific or accurate formula that works effectively on a much larger class of texts (such as "adjudicative decisions from all English-speaking countries" or "Canadian legal texts, writ large"). There are trade-offs inherent in any decision that limits the intended scope of a readability formula: more specific formulas are likely to be better at predicting comprehension levels, but for fewer documents or types of documents;

more general formulas are likely to be less accurate at predicting comprehension levels, but they can be used in a wider range of contexts.

In Chapter 4 (at section 4.2.3.4), I found that the MADRS formula produced meaningfully different, and more accurate, results than 3 other general-purpose readability formulas. But perhaps a legal readability formula that was created from, and intended for use on, legal textbooks, websites, factums, and adjudicative decisions would still have achieved these (or similar) performance gains relative to general-purpose formulas. If so, then I have missed an opportunity to create a tool that would be useful to a wider segment of the Canadian legal community by creating the MADRS formula from, and validated for use on, just adjudicative decision files.

The challenge here is that one cannot really know the extent to which a new formula will be better or worse than an old formula until after the work of creating and testing the new formula is completed. Consequently, in spite of this discussion about some of the costs and benefits that might flow from attempts to extend the MADRS formula's use to other documents, I cannot offer future researchers any definitive answers about whether the efforts would be worth the gains. Undertaking this type of future research involves a gamble. As a minimum, one should be aware of the possibility that such research might result in a wasted effort if, for instance, it reveals that pre-existing readability formulas outperform the MADRS formula when assessing the readability of legal text files other than adjudicative decisions.

7.4 Recommendation for Courts and Tribunals: Track MADRS Results

The MADRS formula could easily be applied by staff members within Canada's courts and tribunals to continue progressing the types of research into decision readability levels that are contained in this dissertation. There are several potential benefits to these courts and tribunals if they choose to assess and disclose the readability of their decisions in the future. One benefit is

transparency. As Justice Brown of the Court of Appeal for Ontario recently noted in a speech about access to justice, transparency is prerequisite for access to justice: "[i]f courts do not publish or make available operational and performance data, then any hope of providing access to justice is mere chimera since without knowing how well a court is providing its adjudicative services and where any problems lie in a court's operations, it is not possible to propose, let alone implement, evidence-based fixes to such problems."[60] Failing to assess readability levels, if one believes that an audience's likelihood of understanding a decision is a relevant performance or quality consideration for a decision, would make it impossible, according to Justice Brown, to fix any readability problems within the court.

But Justice Brown's remarks suggest another benefit of measuring readability levels: he (and potentially other judges) appear to want operational and performance scrutiny from others. His remarks do not suggest that he would respond unproductively to either internal or external studies that assess readability levels of his decisions. In this sense, if courts and tribunals took it upon themselves to measure the readability levels of decisions written by their judges/members, these institutions might just be giving the judges and members what they have been wanting. At least for some decision-makers, this kind of (albeit uni-dimensional) writing performance measurement might actually serve to increase job satisfaction for those who are eager to receive such data.

Chief Justices or tribunal administrators who are responsible for file assignment decisions, and who have accurate readability scores for their decision-makers, might also choose to use this data for their own purposes. For instance, perhaps they would use the data to match decision-makers to files in new ways. They might choose to avoid giving cases involving unrepresented

---

[60] Brown, "5 Habits", *supra* note 4 at 4.

litigants to specific decision-makers who have demonstrated a past inability to write decisions that are likely to be understood by such litigants. To the extent that Chief Justices and tribunal administrators already consider a decision-maker's strengths and weaknesses when assigning files,[61] the MADRS formula may simply offer them one additional piece of information that is potentially relevant to the assignment process.

Ideally, all authors within a particular court or tribunal would want to know – at least periodically – how likely it is that their decisions would actually be understood by their audiences. Presumably, these individuals care to at least some extent whether they are writing decisions that audience members (that is, people who the authors themselves have determined to be potential readers of the decisions) can understand – because if not, then the point of this aspect of their work is called into question. Why bother writing decisions if the people for whom the decisions are intended are unlikely to understand the decisions?

The MADRS formula offers these decision-makers a tool that they could use to self-assess in this respect. As I have stressed throughout this dissertation, the tool is not perfect. It does not measure the actual comprehensibility of decisions – it only predicts what score the decisions would be assigned by human experts, and it does not even perform this task in a way that yields perfect correlations between MADRS and actual expert scores. But it is the best tool among the many that were tested in this dissertation for predicting human-assigned scores for adjudicative decision files. So it can tell decision-makers something more about the language of their decisions than they are likely to be able to figure out without using the tool.

---

[61] Connor Bildfell, "The Role of Chief Justices in British Columbia's Superior Courts" (2018) 76:1 The Advocate 25 at 29-30: "Chief Justice Hinkson noted, for example, that a recently appointed judge would not ordinarily be assigned a complex first-degree murder case in his or her first week; it is only right, in the interests of fairness and justice, that a more senior judge who has accumulated sufficient experience in the area be assigned the case."

Just as a decision-maker might want to know when and why their decision has been overturned on appeal or review, so that they can improve for the next time, decision-makers could also benefit from personalized readability score snapshots that tell them something about whether they are likely to be understood by their intended audiences. And, if judges find that their individual readability scores are not where the judges want these scores to be, then they might choose to undertake continuing professional development or other training to improve the readability of their decisions.

These are just some of the potential benefits for a court or tribunal that chooses to compute MADRS results for the institution's decision-makers. Any tribunal that wanted to start calculating MADRS results would not find it difficult to perform this task. The steps are very straightforward, and could be undertaken in the following way:

1. Each decision/opinion (from the first word to the last word of the decision, but excluding all other front- and end-matter) for each judge/member is saved as a discrete plain text (.txt) file within a directory reserved for that author;
2. Linguistic variable scores are calculated for each of the 6 MADRS variables, using the TAALES, SEANCE, and TAASSC software applications,[62] following the example that is described above at section 4.2.1.2.
3. Scores for each of the 6 variables, for all of an author's decision files, are copied and pasted into a Excel file.

[62] Online: <www.linguisticanalysistools.org>.

4. The variable scores for each file are combined together in accordance with the regression equation that creates the MADRS formula[63] to produce a MADRS value for each decision.
5. An average score is produced for that author.

These steps could all be performed by a data-entry clerk who has moderate competence using Microsoft's Excel software. It would also require a computer that has the relevant linguistic software onboard – which could present a minor bureaucratic hurdle for some courts and tribunals that are bound by government information technology systems rules that are (in my anecdotal experience) extremely restrictive by default.

If an author wrote 45 decisions in a year that were to be scored using the MADRS formula, my conservative estimates about how long it would take to calculate these scores based on the above 5 steps are as follows:

- Step 1 – collecting files – 2 minutes per file: **90 minutes**;
- Step 2 – processing files through linguistic software: **15 minutes**;
- Step 3 – transposing results into an Excel file: **5 minutes**;
- Step 4 – calculating MADRS scores using regression equation: **5 minutes**; and,
- Step 5 – calculating an author's average score: **5 minutes**.

In other words, it would likely take only 2 person-hours of a relatively low-level employee's time to calculate a single author's MADRS results for an entire year. I also note that time involved in completing steps 2-5 are largely independent of the number of files being processed, so one could likely calculate a group of 50 decision-maker's scores (based on 45 decisions per author) in

[63] The MADRS regression equation is reproduced here, for reference: **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).

something close to 80 hours of work. This seems like a small input of institutional time, particularly if one suspects that the results would be useful to and/or appreciated by the judges or tribunal members to whom they pertain.

7.5 Recommendation for Authors: Write for Audience, Not MADRS

Decision-makers who are able to access their MADRS results may wish to review these results. The results probably represent the best large-scale indicator of how likely an audience is to understand adjudicative decisions that is practically available for the time being. But judges and tribunal members should be focused on their audiences, rather than on any particular readability formula, when they write future decisions. It is widely acknowledged that improving (or trying to improve) one's score based on a particular readability formula is not a useful or coherent way to improve the overall readability and quality of one's written document.[64] The paramount concern for authors should be whether their language will be understood by likely readers of a decision, rather than on the formula-based score that the decision might generate – keeping in mind that audiences can be narrower or wider in different cases, just as they can be sophisticated or naïve in law depending on the case.

One way to mitigate against over-reliance on MADRS results as a means of communicating the overall written quality level of a decision, or even the likelihood of a decision being understood by a particular audience, might involve the publication of a disclaimer to accompany any published results. This dissertation is full of qualifications and disclaimers about the limitations of what MADRS-related information can tell an author of a text. Perhaps something like the following

[64] Rebekah George Benjamin, "Reconstructing Readability: Recent Developments and Recommendations in the Analysis of Text Difficulty" (2012) 24 Educational Psychology Review 63 at 64.

would be helpful to prevent authors from fixating counter-productively on a goal of improving MADRS scores by writing specifically and only in ways that will affect MADRS variables:

> *The following scores represent statistically- and linguistically-generated predictions of whether a text is likely to be easily understood by a reader with a particular level of education. The scores are not fool-proof, and the formula that is used to compute the scores can potentially be confounded by many different factors. They should not be considered in isolation. To develop a more complete sense of whether a reader with a particular education level is likely to understand a particular text, a qualitative assessment of the text and its language should also be undertaken, and the two assessments should be considered together.*

That being said, the readability principles that are reflected within the MADRS formula (e.g.: more imageable and concrete language is easier to understand; more common words and phrases are easier to understand; legalese is harder to understand; and, sentences with more complicated syntax, and more embedded clauses, are harder to understand) should be kept in mind when writing. But this consideration does not require authors to constantly cross-check their words against linguistic databases and corpora during their decision-drafting processes to see if they have maximized their scores on all MADRS variables. Crossley, Allen & McNamara's research has noted how writers often use an "intuitive simplification" approach rather than a more mechanical word list approach in order to simplify their writing.[65] These authors also found that scores calculated using the Coh-Metrix (CML2RI) formula (which was tested alongside the MADRS formula above, and was the next best readability formula after MADRS in terms of predicting human experts' scores for adjudicative decision files) changed in relation to texts that were

---

[65] Scott A Crossley, David B Allen & Danielle S McNamara, "Text Readability and Intuitive Simplification: A Comparison of Readability Formulas" (2011) 23:1 Reading in a Foreign Language 84 at 86 ["Text Readability"]:

> Under an intuitive approach, the author's experiences as a language teacher, language learner, materials writer, or any combination of these guide the process of simplification and allow the authors to rely on their own subjective approximations of what learners at a particular level should be able to understand. Even with recourse to word and structure lists for reference, most authors following a structural approach still report relying mainly on their intuition. Research […] also reported that most writers depend on their intuition even when publishers offered advice on how to adapt texts.

simplified intuitively: the texts scored better on the readability formula after being intuitively simplified.[66] In other words, decision-makers likely already have the language tools that they would need, intuitively, to write their decisions in simpler language that is more likely to be understood by wider audiences.

Again, in order to caution decision-makers against using MADRS results and the formula itself for unhelpful purposes, it might also be worthwhile to include the following disclaimer with any publication or disclosure of MADRS results:

> *Research has shown that writers can intuitively simplify their writing based on key principles (such as – use more common rather than less common words; use simpler rather than more complex sentence structures; avoid using foreign-language or technical terms that non-lawyers would not use in their own writing; and try to use concrete and imageable language that describes things in the physical world, more than abstract language). Trying to intuitively write in more simple ways by following these principles will probably help you to write more readable decisions more than thinking about any statistical readability formula will help you.*

Furthermore, authors should acknowledge the limits of their control in terms of the language that they can use to dispose of a case that is in front of them. It is generally helpful to know that terms like "beach" are highly imageable. But this knowledge is of little assistance to a judge or tribunal member who is adjudicating a specific dispute about a human rights or constitutional equality case, wherein the facts all relate to questions of dignity – a concept that is neither imageable nor concrete.[67] The decision-maker's freedom to choose their own language is necessarily limited by the nature and the facts of the case that is before them, and this reality must be kept in mind when thinking about any decision's readability level.

---

[66] Crossley, Allen & McNamara, "Text Readability", *supra* note 65 at 94-95.

[67] The word "dignity" scores 280 on the Medical Research Council's Concreteness scale within its Psycholinguistic Database. Online: <https://websites.psychology.uwa.edu.au/school/MRCDatabase/uwa_mrc.htm>. The challenges of picturing "dignity" are also illustrated in *The Simpsons*, "A Millhouse Divided" (Fox Television broadcast, 1 December 1996), during a Pictionary-style drawing game played between various characters. The relevant (short) segment of this episode is available online: <https://www.youtube.com/watch?v=Uw3AghZHCdk>.

But authors always have some degree of freedom to choose their language. So, my recommendation to decision-makers is to seek out and consider their MADRS results in a spirit of self-improvement. Regardless of how well each author believes that they now write, I encourage them to be open to the possibility that they might be better able to reach their audiences in at least some cases by thinking through their readability scores. MADRS results can help these authors to see in specific cases where they have scored higher (less readably) than they might have expected, and where they have scored at or better than the levels that they would hope to achieve. These real-world examples, particularly if they are considered alongside parallel qualitative assessments of their writing in at least some of the same decisions, could give judges and tribunal members feedback on their language that, until now, has been difficult to provide on a large scale.

7.6 AI AND ALTERNATE COMMUNICATION APPROACHES: BEYOND MADRS

Readers of this dissertation may recall my earlier consideration[68] about whether there is any need to study the readability of adjudicative decisions today, in an age when artificial intelligence (AI) and natural language processing (NLP) technologies are evolving very quickly to assist humans with language-related tasks. I concluded this discussion by suggesting that AI and NLP technologies are not so advanced right now as to render the topic of readability irrelevant or obsolete, because current text-summarization tools do not seem to be adequate to meet the needs of some people who will need or want to understand an adjudicative decision, and because text generation tools do not yet seem to be widely used by decision-makers in order to write decisions.

But AI and NLP technologies, and other substitutes or processes for communicating the meaning of adjudicative decisions, might nonetheless have useful roles to play as part of any

[68] Above, at section 1.7.

broader project to accurately and effectively transmit the key content of these decisions to wider audiences. In this section, I will therefore step back from just looking at the linguistic readability of decisions in isolation (as if improving the readability of these decisions was the only, or even the most important way to communicate the decisions' contents to intended audiences). Instead, I will consider here how other approaches might be used either instead of, or alongside, efforts to measure and enhance readability levels of the decisions themselves.

7.6.1 AI Summarization Tools

As I noted previously in section 1.6 of this dissertation, research into the effectiveness of some AI summarization tools used on adjudicative decisions suggests that legal knowledge may be necessary in order to effectively prompt the AI tools to generate the kind of output that a reader is looking for, and to verify that the output is accurate.[69] And I also noted that the CanLII database now has AI-generated summaries for almost all case law from Alberta, Saskatchewan, Manitoba, and Prince Edward Island.[70] So, there are many different potential AI summarization tools available to the general public, but it is not yet clear how well they perform for the purposes of communicating the important content of a decision to a member of the general public.

It is not my intention to study this question in detail. A separate study that empirically assesses how well non-lawyers with different levels of education understand summarized legal decisions from AI tools, compared to how well these people understand the original decisions,

---

[69] Aniket Deroy et al, "An Analytical Study of Algorithmic and Expert Summaries of Legal Cases" in Erich Schweighofer, ed, *Legal Knowledge and Information Systems* (Amsterdam: IOS Press, 2021) 90 ["Expert Summaries of Legal Cases"]; and, Elliott Ash et al, "Translating Legalese: Enhancing Public Understanding of Court Opinions with Legal Summarizers" in CSLAW '24: Proceedings of the Symposium on Computer Science and Law (New York: Association for Computing Machinery, 2024) 136 ["Translating Legalese"].

[70] Lexum, "AI-Powered Case Analysis Added on CanLII for All Case Law From Alberta, Saskatchewan, Manitoba, and Prince Edward Island" (3 Jun 2024), online: <https://lexum.com/en/blog/ai-powered-case-analysis-added-on-canlii-for-all-case-law-from-alberta-saskatchewan-manitoba-and-prince-edward-island/>.

would probably be needed to offer a definitive answer to the question of how well the summarization tools work.

For the time being, however, I will offer a few observations that are based on my impressionistic assessment of how well AI summarization tools work on adjudicative decisions. My impression is informed from, among other things, my experimentation with different AI platforms in asking these platforms to summarize a family law case from Saskatchewan.[71]

CanLII's AI-generated summary for this case[72] is relatively short, and contains some information about the key details of the case. However, the language of the summary appears to be drawn primarily from the decision itself, and there does not appear to be any simplification of the language from the decision within the summary. Furthermore, a central legal issue in the case involves the "*parens patriae* jurisdiction." This doctrine is referred to and explained (by references to older cases dealing with the doctrine) in the decision.[73] However, the Latin term is used in the summary without any explanation as to what it means. This use of a technical, Latin term within the summary, with no explanation of what the term means, would likely hinder a non-lawyer's ability to understand what the case is about from just reading the summary.

When I prompted ChatGPT to summarize the decision, the response did not correspond at all to the actual case.[74] The response was utterly unreliable, and it is not clear where the information contained in the response may have originated from.

---

[71] *Aalbers v Aalbers*, 2024 SKCA 83.
[72] Online: <https://www.canlii.org/en/sk/skca/doc/2024/2024skca83/2024skca83.html#summary>.
[73] *Aalbers v Aalbers*, 2024 SKCA 83, at paras 70-71.
[74] Prompt: "Summarize the court decision from *Aalbers v Aalbers*, 2024 SKCA 83." Response: "In Aalbers v. Aalbers, 2024 SKCA 83, the Saskatchewan Court of Appeal addressed issues related to family law, particularly concerning the division of property and spousal support following a separation. The court upheld the trial judge's decision regarding the equalization of property. […]" (response trimmed for length). The ChatGPT response missed all the key details of this case except that it was a family law decision. The decision concerns a child with Down syndrome, whose father sought to have her undergo a psychological capacity assessment. It had nothing to do with divorce, separation, or the division of property.

Microsoft Copilot responded to my prompt by indicating that the application "can't provide a detailed summary of specific court decisions." No other program or application that I found would provide a free document summary of this decision (consisting of 8000 words, when many applications that provided free summaries were limited to 600 words). My inability to locate other direct-to-public tools that could be used without cost for summarizing legal texts accords with Salyzyn & Stroud's observation that "[n]otwithstanding ChatGPT's public release in November 2022, we did not identify any tools during our review [of direct-to-public digital legal tools available in Canada] that appear to be using generative AI."[75]

I understand that better-performing legal document summarization tools that use AI technology currently exist. However, if these tools are not freely-accessible, and cannot easily be found by non-expert users, then they are unlikely to represent an adequate substitute for a readable decision. The CanLII summary seems to be the only tool that meets these criteria.

My assessment is that the CanLII summary would be a poor substitute for the actual text of the Saskatchewan Court of Appeal decision for a reader who was specifically motivated to read about this case. Having reviewed several other CanLII summaries for other decisions, I believe that this assessment would apply equally to most other CanLII summaries. The summaries do not trace the decision-makers' reasoning processes in ways that allow a reader to logically follow the full chain of reasoning. These summaries do, however, effectively communicate the result of a decision and the basic context from which the decision was made. Based on the information in the summaries, they are probably helpful in providing people with enough information about a case to help them decide whether they would want to read the full decision, but they are probably

[75] Amy Salyzyn & Shannon Stroud, "Direct-to-Public Digital Legal Tools in Canada: A 2023 Update" (June 30, 2024). Ottawa Faculty of Law Working Paper No. 2024-15, online: <https://ssrn.com/abstract=4880902>, at 6.

inadequate substitutes for the full decision for people who want a greater degree of information about the facts of the case and the chain of reasoning that led to the conclusion.

Furthermore, at least according to the MADRS formula, the summaries are not consistently more readable than the full decisions – within a small pool of decisions and summaries that I assessed. Specifically, I compared MADRS results for 4 decisions (one each from Alberta, Saskatchewan, Manitoba, and Prince Edward Island)[76] and the accompanying AI-generated summaries that CanLII produced for these decisions.[77] The scores for each decision and the associated summaries are reproduced below in Table 7.1.

Table 7.1 – MADRS Results for Select SCC Opinions and Summaries

Case Name	**MADRS: Decision (# words)**	**MADRS: Summary (# words)**
R v Gallant	4.52 (4592)	4.79 (446)
Aalbers v Aalbers	4.79 (8802)	4.72 (416)
R v Cousins	3.99 (6874)	4.79 (327)
Freid v Air Canada	4.21 (5033)	4.44 (463)

As this table shows, the MADRS results for 3 of the 4 summaries are higher than the results for the associated cases. In other words, the summaries are less quantitatively readable in all cases except for *Aalbers v Aalbers* than the decisions that are summarized, and even in the *Aalbers* case the summary is only marginally (1.8%) lower-scoring on the MADRS scale than the decision.

[76] *R v Gallant*, 2024 ABKB 541; *Aalbers v Aalbers*, 2024 SKCA 83; *R v Cousins*, 2024 MBPC 69; and, *Freid v Air Canada*, 2024 PESC 39.
[77] The summaries were available online at the following website addresses:
<https://www.canlii.org/en/ab/abkb/doc/2024/2024abkb541/2024abkb541.html#summary>;
<https://www.canlii.org/en/sk/skca/doc/2024/2024skca83/2024skca83.html#summary>;
<https://www.canlii.org/en/mb/mbpc/doc/2024/2024mbpc69/2024mbpc69.html#summary>; and,
<https://www.canlii.org/en/pe/pesctd/doc/2024/2024pesc39/2024pesc39.html#summary>.

From the figures in parentheses in this table, it is clear that the summaries are much shorter than the decisions, so they could be useful in saving readers time (if, for instance, the information that they communicate is sufficient to help readers triage which cases of potential interest they should actually read in full, as opposed to passing over). But whatever else these summaries might do, it is not obvious that they would be helpful in consistently communicating the content of decisions to readers in language that is easier to understand than the language of the full decisions.

The above discussion is not intended to suggest that CanLII's summaries (or others) will never become suitable substitutes for full decisions. The extremely small and non-random[78] sample that I selected for study here may not fairly represent the relative readability levels of the full population of CanLII's summarized decisions and their associated summaries. And my qualitative assessment of these summaries is not necessarily applicable to other summaries that I have not seen. But my own review of these summaries accords with other research[79] that identifies weaknesses in the summaries which probably render them less than suitable as complete substitutes for the full texts of adjudicative decisions.

My review of these CanLII summaries does, however, suggest that they could have a useful role to play in helping non-lawyers to identify which cases are more relevant to these people for the purposes of reading the full decisions in more detail, and which cases are likely not relevant. The summaries are free, are easy to find on the internet within the largest freely-available caselaw database in Canada, and likely contain enough information to assist readers in determining whether

---

[78] I selected one decision from each jurisdiction that has CanLII's summaries available, from 2024, with a view to getting decisions from Appeal, Superior, and Inferior courts. Beyond these criteria, I scrolled randomly through the results and selected whichever decision my cursor landed upon. This is not a sound randomization methodology, but the goal was not to develop a random sample that would be representative of the full population of cases here – this would potentially be a useful research project for the future.

[79] Deroy, et al, "Expert Summaries of Legal Cases", *supra* note 69; and, Ash, et al, "Translating Legalese", *supra* note 69.

to read the full case. If nothing else, these summaries should help to reduce the volume of decisions that readers might need to sift through in order to find decisions that are most relevant to their own situations. And, since the summaries seem to be often written in language that is more difficult to understand than the language of the decisions, the summaries might help readers to more quickly assess whether they would be able to understand the full decision text: if the reader can easily understand the language of the summary, then (based on my admittedly very small and non-random sample of cases represented above in Table 7.1) we have some reason to believe that the reader would often also understand the full text of the decision.

7.6.2 Generative Text Tools to Help Authors Draft or Edit Decisions

AI tools can also be used to generate and revise – not just summarize – written texts. So it is worth considering how these tools might be helpful in creating decisions that are more understandable to a wider audience. Notwithstanding my observation earlier in this chapter (within section 7.5) that many writers use a process of intuitive simplification to rewrite their texts for particular audiences, I acknowledge that a weakness of this approach involves its dependence on individual intuitions – which may or may not lead to helpful revisions of a text. For instance, a writer who has no or limited familiarity with the needs of less-educated or less competent users of the English language might not have accurate intuitions about how to rewrite a text in order to make it more understandable to these groups of people. This situation represents one where AI tools might assist decision-makers in revising their language to better meet the needs of audience members.

Research to date suggests that AI can improve readability scores of a source document in the general direction of a target score, but cannot reliably achieve the target score.[80] In a study using ChatGPT and Llama-2 AI tools, the authors prompted the AI systems to paraphrase a set of source texts of different Flesch Reading Ease Score (FRES) readability levels for audiences ranging from Grade 5 students to university graduates.[81] The authors used both one-step ("zero-shot") and two-step approaches (wherein the output text from the first step was reprocessed according to the same prompt a second time).[82] They noted that, "often the outputs from the Llama-2 zero-shot solution for certain input passages are random incoherent string of tokens."[83] This output would not help authors to better calibrate their writing levels to match the needs of their audiences.

More generally, the authors found that the two-step processes pushed source text closer to the target levels than the one-step process.[84] But the tools were generally not successful in attaining the target levels, on average, particularly where there were larger differences between source and target levels of FRES values. For instance, when the target readability level was a FRES of 55 (suitable for readers in Grade 10-12) the AI tools produced average paraphrased FRES results of between 28.3 and 64.2 (depending on the starting FRES value for the source text).[85] In other words, AI tools can "steer" readability, but the resulting texts still display readability levels that are

---

[80] Asma Farajidizaji, Vatsal Raina & Mark Gales, "Is it Possible to Modify Text to a Target Readability Level? An Initial Investigation Using Zero-Shot Large Language Models" (2024) *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation* (LREC-COLING 2024) 9235, online: <https://aclanthology.org/2024.lrec-main.815.pdf> ["Modify Text to a Target Readability Level"].
[81] Farajidizaji, Raina & Gales, "Modify Text to a Target Readability Level", *supra* note 80 at 9329.
[82] Farajidizaji, Raina & Gales, "Modify Text to a Target Readability Level", *supra* note 80 at 9329.
[83] Farajidizaji, Raina & Gales, "Modify Text to a Target Readability Level", *supra* note 80 at 9329.
[84] Farajidizaji, Raina & Gales, "Modify Text to a Target Readability Level", *supra* note 80 at 9330.
[85] Farajidizaji, Raina & Gales, "Modify Text to a Target Readability Level", *supra* note 80 at 9332, Figure 4a.

correlated with, or anchored to a certain extent by, their source texts. They produce relative improvements, but do not achieve the intended absolute changes to readability levels.

In spite of this current limitation of generative AI tools, they could still be useful to decision-writers who lack an (or do not trust their) intuitive sense of how to rewrite a decision to make it more understandable. The above research suggests that most texts can be driven closer to a readability target through the use of AI tools. Thus, decision-makers may wish to experiment with, for instance, a two-step ChatGPT prompt asking the system to "[p]araphrase this document for 8th/9th grade school level (US). It should be plain English and easily understood by 13- to 15-year-old students."[86] If the decision-makers' original drafts are significantly less readable than this target, then the AI tools would seem to produce output that will come closer to the target in many cases.

However, two messages of caution are warranted here. First, Farajidizaji, Raina & Gales found that "more significant shifts in readability lead to reduced semantic and lexical similarity between source and target texts, highlighting the challenge of balancing readability control and content preservation."[87] In other words, output texts that succeeded in producing substantially different readability levels when compared to the source texts also may have communicated more different content in the ultimate meaning of the texts. So, authors would need to review the output texts closely to ensure that they still accurately communicated the authors' intended messages.

Second, decision-makers would likely want to assess any ethical and privacy-related considerations that arise when using AI tools for assistance in reviewing or producing decisions that have not been finalized, and are not in the public domain. Specifically, these decision-makers

---

[86] Farajidizaji, Raina & Gales, "Modify Text to a Target Readability Level", *supra* note 80 at 9329, Figure 2. This was the prompt used within the authors' study that was designed to generate texts with a FRES of 65.
[87] Farajidizaji, Raina & Gales, "Modify Text to a Target Readability Level", *supra* note 80 at 9333.

should perhaps consider whether it is appropriate from the perspective of deliberative secrecy (or other similar perspectives) to allow any draft decision text to be uploaded and processed by AI tools in ways that might place the draft texts beyond the control and 'safe-keeping' of the authoring decision-makers' own institution's secure information technology systems. I am not an expert in information technology security, and am not in an informed position to fully canvas the extent to which different AI systems might create these kinds of risks or concerns for decision-makers, but I raise them here as potentially important factors for decision-makers to contemplate for themselves before using AI tools on draft decisions.

In contrast, where an author uses the same AI tools on finalized decisions that are already in the public domain (e.g.: decisions that are published on CanLII), then no similar ethical, privacy, or deliberative secrecy concerns arise. The AI tools would arguably be less useful for readers of the decision in question – since it is already finalized, such that any suggested revisions to the decision would be moot for these audiences. But the AI tools could nonetheless be helpful to authors who use them in order to learn what changes to a finalized decision would likely have made it more understandable. Perhaps through these kinds of examples of how to revise their texts to make them more readable, authors could develop better intuitive simplification skills that they could apply in future decision-writing efforts.[88]

As this discussion suggests, current AI systems might offer decision-writers some assistance in making their texts more readable, subject to the above limitations and concerns. And

[88] As I noted above, at section 6.5, and as much research has illustrated, writers can improve their texts through study and emulation of examples of other documents that are at the standard these writers are striving to produce. *See, e.g.*, Zhiwei Wu, "Understanding Students' Mimicry, Emulation and Imitation of Genre Exemplars: An Exploratory Study" (2019) 54 English for Specific Purposes 127; 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; and, Martine AH Braaksma, et al, "What Observational Learning Entails: A Multiple Case Study" (2006) 6:1 L1 – Educational Studies in Language and Literature 31.

with future improvements to these AI tools, their potential for helping judges and tribunal members to produce understandable decisions is likely to only increase. As this evolution continues, there are opportunities for further research like the study conducted by Farajidizaji, Raina & Gales that could apply more specifically to the legal domain by using adjudicative decision texts as the source documents, and that would assess source and output documents for readability using the MADRS formula in order to measure the extent of any AI-generated improvements between the source and output documents.

7.6.3 Other Ways of Communicating the Content of Adjudicative Decisions

Although this dissertation has studied the readability of adjudicative decisions, and has suggested instances where these decisions are likely not written in ways that many audience members for the decisions could easily understand, it is not necessarily the case that the decisions need to become more readable. Perhaps there are other ways to communicate the content of the decisions without imposing this burden on the decision-makers.

One example of such a mechanism for communicating the important information about decisions to the general public is the SCC's "cases in brief" summaries.[89] As the court's website indicates, "Cases in Brief are short summaries of the Court's written decisions drafted in reader-friendly language, so that anyone interested can learn about the decisions that affect their lives. They are prepared by communications staff of the Supreme Court of Canada. They do not form part of the Court's reasons for judgment and are not for use in legal proceedings."[90] This approach seems to shift some of the work of writing about the case from the judges of the court, to communications professionals who may have more trained or calibrated understandings of how to

[89] Supreme Court of Canada, "Cases in Brief", online: <https://www.scc-csc.ca/case-dossier/cb/index-eng.aspx>.
[90] Supreme Court of Canada, "Cases in Brief", online: <https://www.scc-csc.ca/case-dossier/cb/index-eng.aspx>.

intuitively simplify the language of the decision. It does not require judges to produce the summaries. It uses real people to write the summaries instead of machines. And the summary does not have any legal effect on the decision; for instance, any legal imprecision in the summary does not undermine the legal precision of the decision.

In theory, this approach is sound. In practice, however, it is not clear that the summaries would have the intended effects. To begin with, the summaries that I have read (a large sample, but not all) only describe the controlling judgement from a decision. They do not describe any dissenting or concurring opinions. This approach is justifiable if the function of the summary is simply to instruct readers about the disposition of the case and the reasons for this disposition, but it is probably inadequate if the summaries are intended to instruct readers more broadly "about the decisions that affect their lives,"[91] because it may often be important to readers to know the extent to which consensus was achieved by the court in a particular case.

Furthermore, my impressionistic reading of the summaries suggested to me that they may not actually be written in language that is any more readable than the language of the decisions that they summarize. I again selected a small sample of cases and summaries to measure (using the MADRS formula) how readable the summaries are in comparison with the decisions. Specifically, I selected 3 cases from 2024 without using any particular selection criteria.[92] By coincidence, all 3 cases involved at least one concurring or dissenting opinion. I then calculated MADRS scores for the majority opinions from each of these 3 cases, and the corresponding "case in brief" summaries of these majority opinions.[93]

---

[91] Supreme Court of Canada, "Cases in Brief", online: <https://www.scc-csc.ca/case-dossier/cb/index-eng.aspx>.
[92] *Scott v Golden Oaks Enterprises Inc*, 2024 SCC 32 [*Scott*]; *Iris Technologies Inc v Canada (Attorney General)*, 2024 SCC 24 [*Iris*]; and, *R v Bykovets*, 2024 SCC 6 [*Bykovets*].
[93] The summaries are available at the following link, online: <https://www.scc-csc.ca/case-dossier/cb/index-eng.aspx>.

The results of this small study were revealing. The MADRS results for each majority opinion and summary were as follows (majority opinion / summary): *Scott* (4.97 / 4.64); *Iris* (5.49 / 5.34); and, *Bykovets* (5.10 / 4.98). So, in each case, the summary scored lower, or was more readable, than the majority opinion that the summary described. But the readability differences between the summaries and the opinions were very small, at less than 1/3 of a point on the 7-point MADRS scale for each pairing. In none of the 3 cases was the summary written at a level that a person with only a completed high school education would be likely to easily understand. These results suggest that the summaries are not as useful as we might want them to be in terms of helping members of the general public to understand SCC decisions.

A closer look at the component variable scores for the MADRS results also exposes some interesting considerations. The component variable scores, and MADRS results, are shown below in Table 7.2 for both the majority opinions (MAJ) and the case in brief (CIB) summaries.

Table 7.2 – Component Variable Scores for Select SCC Opinions and Summaries

Document	# Words	MRC_Imageability_AW	MRC_Concreteness_FW	Kuperman_AoA_CW	COCA_spoken_tri_prop_30k	Legal_GI	nominal_deps_NN_stdev	MADRS
Iris MAJ	5535	298.98	229.27	8.436	0.06073	0.07501	1.2111	5.49
Iris CIB	540	303.78	224.88	8.801	0.06691	0.10000	0.9885	5.34
Scott MAJ	12929	303.01	232.37	7.988	0.05593	0.05616	1.1700	4.97
Scott CIB	617	306.57	231.35	7.743	0.07317	0.06483	1.0707	4.64
Bykovets MAJ	9415	296.77	227.61	7.744	0.05397	0.03425	1.2291	5.10
Bykovets CIB	621	292.94	229.00	7.560	0.07108	0.06763	1.0991	4.98

As these results show, the summaries are successful in dramatically reducing the length of the text that is to be read in order to learn about the decision – cutting down the number of words by at least 90%. Three other patterns within these component variable results are also worth noting.

First, with respect to syntax (*Nominal_Deps_NN_StDev*), the summaries score lower than the majority opinions in all three cases. In other words, the summaries consistently use simpler sentences with less dependent clauses than the majority opinions. This makes sense, since almost every element of plain language- or readability-related advice suggests that using shorter sentences

is helpful. A second, and similar pattern exists for the use of more common three-word phrases (*COCA_Spoken_Tri_Prop_30k*): the scores for this variable are higher in the summaries than in the majority opinions for all 3 cases, which indicates that the summaries tend to use word combinations that are likely more familiar to readers in higher proportions than the majority opinions use these familiar word combinations. This reality is also consistent with general writing advice to use simpler and more common language instead of more obscure or novel words and phrases. Both the syntax and three-word phrase variables' scores help to explain why the summaries have lower (more readable) MADRS scores than the majority opinions.

Finally, with respect to the density of legalese (*Legal_GI_Nouns*), the scores are higher for the summaries than for the majority opinions in all 3 cases – and higher scores for this variable contribute to higher (less readable) MADRS values. This result is perhaps because the summaries attempt to communicate the most important content from the majority opinion (which is law-heavy), and omit discussion of certain other topics (that are perhaps less law-heavy). If this approach to the summaries is actually taken by the writers, then the product seems to become one with far fewer total words, but a more legally-dense text that is not necessarily much easier for readers to understand. Although the sample size of SCC case summaries discussed here is small, the tendency for supposedly 'plain language' summaries to be largely unreadable to the general population because of a high density of technical jargon has been documented by researchers in other contexts (i.e.: plain language summaries of health care research).[94] These legal and healthcare results raise an interesting question, for which I do not have a definitive answer: is

[94] See, generally, Iain A Lang, et al, "Jargon and Readability in Plain Language Summaries of Health Research: Cross-Sectional Observational Study" (2025) 27 Journal of Medical Internet Research 1, online: <https://www.jmir.org/2025/1/e50862>.

comprehensibility better served by producing a shorter, but more technically dense document that summarizes the original, or by a longer original with less dense technical language?

As the above discussion illustrates, there is probably a role to be played by documents other than the original decisions in terms of communicating the decisions' contents to the general public. But even the SCC's cases in brief are not necessarily effective in accomplishing this role, yet. To be sure, the above discussion focused on a very small sample of only 3 summary/opinion pairings, so the points that emerge from this small study are not necessarily applicable to the full spectrum of case in brief summaries. The above results do show that the studied human-written summaries achieve small MADRS readability gains when compared to the opinions that they summarize, but the summaries seem to incorporate older theory (related more to sentence length and word or phrase frequency variables) into their simplification approaches, without necessarily considering concepts like the density of technical language in the shorter documents. There seems to be room for further readability improvements in the SCC summaries, and in any other similar products that are designed to reach widespread audiences with information about what was said and done within an adjudicative decision.

If nothing else, the summaries do represent a small step toward improved readability of documents that deal with the SCC's decisions, and they highlight how other texts beyond just the decisions themselves could be useful in communicating the substance of adjudicative decisions to different audiences. And the above discussion also highlights how a qualitative approach to summarizing the court's decisions in more understandable ways can be assessed using a quantitative approach (the MADRS formula) in a way that may help to identify how successful the former approach will be. In this sense, the discussion reiterates an idea that was raised earlier

in this dissertation,[95] namely that quantitative and qualitative approaches will often function best when used together with one another as a combined means of improving the comprehensibility of a person's writing.

7.7 Concluding Thoughts on the Implementation of Recommendations

Almost everyone is busy, or feels busy, in their professional lives already. I appreciate that every good idea from an academic researcher about how others ought to do things in their work usually comes at both a time and a monetary cost to the implementing organization, and that these costs often make ideas that are attractive in theory unworkable in practice. The workability levels for the different recommendations in this chapter are therefore somewhat unusual. The software that is needed to calculate MADRS results is freely available. The work of building, validating, and demonstrating how to use the formula is all already done within this dissertation. And, while it is true that a small amount of person-hours would be needed (likely less than 80 hours – or about two weeks worth of work – per year) from a data-entry clerk or other similar employee in order to produce MADRS values for a medium-to-large sized court or tribunal consisting of about 50 decision-makers who each write 45 decisions per year, this time cost is perhaps reasonable in light of the insights and feedback that could follow from MADRS results for all decision-makers and all decisions in a given year.

The point here is that it is practically within reach for many key stakeholders to now start measuring and considering the readability scores of adjudicative decisions in Canada. As multiple examples from within Chapters 4-6 of this dissertation have demonstrated, we can learn new things about adjudicative decisions and the ways in which these legal texts are likely to be understood (or

[95] Above, at section 6.5.

not) by their audiences through use of the MADRS formula. No statistical, linguistic, or other specialized knowledge is needed to either calculate or understand the meaning of MADRS readability results. The tool is ready for immediate use.

8. Conclusion: Legal Readability's Coming of Age in Canada

In Chapter 1, I suggested that Canada may be witnessing the beginning of a movement toward increased interest in quantitative readability studies of the written products that emerge from courts and tribunals,[1] based on efforts made by certain tribunals to set and deliver upon readability targets.[2] The Federal Court also recently endorsed a tribunal member's use of simple and effective language that non-lawyers can understand within the tribunal member's decision.[3] These developments, together with statements from serving and prospective judges,[4] and guidance documents for tribunal members,[5] all suggest that judges and tribunal members are generally expected, and want, to write readable decisions.

As this dissertation has shown, however, there is a disconnect between expectation and reality in terms of the readability levels of Canadian adjudicative decisions. The average court or tribunal decision – regardless of subject area, court level, jurisdiction, author's gender, or other special contexts – is written at a level that requires a reader to have more than a completed high school diploma in order to easily understand the decision, when millions of Canadians do not have this level of education. There are several possible explanations for this disconnect, and several

---

[1] At section 1.2.
[2] Social Security Tribunal of Canada, "An Evaluation of How Easy it is to Read Decisions of the Social Security Tribunal" (last modified 10 May 2022), online: <https://sst-tss.gc.ca/en/our-work-our-people/evaluation-easy-it-read-decisions-social-security-tribunal> ["An Evaluation"]; and, Shannon Salter, "RDO Speak to Shannon Shalter [SIC] About All Things ODR" (November 2019), online: <https://resolvedisputes.online/blog_shannon_salter.html>.
[3] *Bashir v Canada (Citizenship and Immigration)*, 2025 FC 19, at paras 36-37.
[4] Canada, Department of Justice, *The Honourable Mahmud Jamal's Questionnaire (Questionnaire for the Supreme Court of Canada Judicial Appointment Process)* (17 Jun 2021), online: <https://fja-cmf.gc.ca/scc-csc/2021/nominee-candidat-eng.html>; Beverly McLachlin, "Remarks of the Rt Hon Beverley McLachlin, PC, Chief Justice of Canada" (May 16, 2010), *NJI and CIAJ Advanced Judgment Writing Seminar* (Perth, Ontario), online: <https://ciaj-icaj.ca/wp-content/uploads/podcasts/2020/11/podcast1b_16may10-nji-ciaj-judgment-writing-course-copy.pdf>, at 11.
[5] Ombudsman Saskatchewan, *Practice Essentials for Administrative Tribunals* (Saskatchewan: Ministry of Justice and Attorney General, 2020) at 68-69; Tribunals Ontario, "Guidelines for Reasons Review and Decision Quality Standards" (20 May 2022), online: <https://tribunalsontario.ca/en/practice-directions/>.

related ways in which the contents of this dissertation can help courts and tribunals to overcome readability deficits within their decisions.

To begin with, courts and tribunals (with very few exceptions), do not publicly measure and report on readability levels for their decisions. However, this dissertation has begun the work of publicly reporting on the topic. Perhaps courts and tribunals, or individual judges and tribunal members, who see the results of this dissertation will carry the project forward by measuring readability levels for their future decisions. For those who had never considered whether it would be possible to quantitatively measure readability levels, this dissertation has shown that such measurements can be obtained.

However, even for institutions like the British Columbia Civil Resolution Tribunal (BCCRT), and the Social Security Tribunal of Canada (SST) – that did establish and keep track of readability objectives – the results of this dissertation suggest that the goals set by these institutions are not being met, regardless of what their internal analyses might conclude. For instance, the results above (at Chart 4.7) show that SST decisions score an average of 3.93 on the MADRS scale. This means that a reader would need more than a completed high school diploma (that is, more than grade 12 education) in order to easily understand the SST's average decision. This result is inconsistent with the SST's target of a grade 9 reading level for its decisions,[6] and undermines the tribunal's internal analysis which showed that approximately 65% of its decisions were written at or below a grade 10 reading level (using the Flesch-Kincaid Grade Level (FKGL) formula as the basis of measurement).[7]

---

[6] Social Security Tribunal of Canada, "An Evaluation", *supra* note 2.
[7] Social Security Tribunal of Canada, "An Evaluation", *supra* note 2.

To be clear, the SST's average decision was more readable than any other subclass of decisions that have been reported on within this dissertation, and the tribunal seems to be at the forefront of efforts to improve decision readability levels. But where the tribunal has relied upon a readability formula from the pre-computer era to assess its performance, the results are not necessarily reliable. The MADRS formula – that was developed from recent Canadian adjudicative decision files and that was statistically found to be a better predictor of human expert-assigned readability scores than the FKGL formula – indicates that more work needs to be done by the SST if the tribunal hopes to make its decisions easily understandable to people with only a grade 9 level of education.

Similarly, the BCCRT is far from attaining the tribunal's target of writing decisions at a grade 6 reading level. Within the user-searchable database of MADRS results that is described in Chapter 6, there were 34 decisions from the BCCRT.[8] The average MADRS result for these decisions was 4.09. The most readable decision scored 3.10 on the MADRS scale.[9] In other words, none of the tribunal's decisions scored less than 3 (completed high school education), or 2 (some high school education). The tribunal's target of a grade 6 readability level would actually require scores in the range of 1.0 (no high school education, or less than fluent in English) on the MADRS scale. None of the decisions that were studied in this dissertation, from the BCCRT or any court or tribunal, scored less than 2.

Again, the BCCRT's average score was excellent in comparison with the average scores from other court levels and in other contexts that were studied in this dissertation. Furthermore,

[8] If one sorts the database by "CourtLevel", then by "Jurisdiction", then by "FileName", the BCCRT decisions appear in Rows 1178 to 1211 of the database.
[9] *Smith v Elise*, 2022 BCCRT 903. This case was about a commercial landlord-tenant dispute, and was written by Tribunal Member Leah Volkers.

the number of decisions from the BCCRT that are included in the user-searchable database described in Chapter 6 (and the manner in which these decisions were non-randomly selected for inclusion) do not allow us to draw statistical inferences from these decisions about the full population of BCCRT decisions. It is therefore statistically possible that the average BCCRT decision does score in the 1.0 range on the MADRS scale, and that the 34 decisions for which scores were calculated in the Chapter 6 database are coincidentally among the least readable decisions that the tribunal produced in 2022. In light of what is now known about the MADRS results for thousands of other decisions that were scored within this dissertation, though, it does not seem plausible that the BCCRT is producing decisions that are, on average, approximately 3 points lower on the MADRS scale than any other court or tribunal's average decision.

The point here is not to embark on a new case study of readability levels for the BCCRT. Rather, the point is to demonstrate how even two of the most readability-conscious tribunals in Canada (the BCCRT and the SST) have perhaps developed unrealistic readability targets for their decisions, and/or relied on outdated and inaccurate general-purpose readability metrics that do not perform particularly well in the specialized environment of legal English. Ideally, the work within this dissertation will motivate leaders within these tribunals to reconsider whether or where they want to set readability targets, and how they assess the extent to which tribunal members are achieving any established targets. Other courts and tribunals that had not previously contemplated the idea of measuring readability levels for decisions may also find inspiration here about how they could perform this task. For all these courts and tribunals, the MADRS formula is likely their best tool for the purpose.

Benjamin, in her review of relatively recent developments in readability theory, practice, and technology, proposes the following practical framework for evaluating different readability approaches:

- Is the method intended for / tested on a particular population?
- Is the method intended for / tested with particular texts?
- Has the method demonstrated predictive validity—especially in predicting readers' comprehension of texts?
- How widely available for immediate use is the method?
- How much special training is required to use the method?[10]

These guided questions imply that domain-specific readability formulas will tend to be superior to general-purpose formulas, and that valuable formulas must be capable of being immediately and easily available for stakeholders to use.

The MADRS formula seems to fare rather well when compared against Benjamin's evaluation framework:

- The formula was developed from expert assessments of texts based on the reading needs and abilities of the Canadian adult population;
- It was developed from, and intended for use on, contemporary Canadian adjudicative decision texts;
- It was statistically validated using an independent data set, and outperformed 6 other general-purpose formulas in terms of its predictive power;

[10] Rebekah George Benjamin, "Reconstructing Readability: Recent Developments and Recommendations in the Analysis of Text Difficulty" (2012) 24 Educational Psychology Review 63 at 65.

- It relies on freely available linguistic software to compute the variable scores that contribute to the formula; and,
- The formula could be used by any person who has basic data-entry and spreadsheet competencies, with a small amount of training about how to calculate MADRS values. No specialized understanding of linguistics or law is needed in order to compute MADRS values.

In other words, the MADRS formula could be immediately and efficiently put into action by courts and tribunals, scholars, advocacy groups, or other interested stakeholders, with relative ease.

This dissertation has created the new MADRS tool that can reliably tell us how readable court and tribunal decisions are, and it has demonstrated through analysis of thousands of files that these decisions (in 2022) were generally not readable enough for many Canadians who will inevitably need to interact with courts or tribunals at some point in the future to be able to understand the decisions. If the goal is to produce decisions that the public and the parties can understand – as many judges affirm – then the goal is not being met in a significant number of circumstances. Having put numbers to words within this dissertation, and having identified the disconnect that exists between the aspirational and actual readability levels for Canada's adjudicative decisions, the more challenging task now lies ahead: it is time for deeper, more regular, and more goal-focused attention to be given to the idea of communicating understandably through the decisions of judges and tribunal members. If leaders within our courts and tribunals can succeed in effectively creating and sustaining this attention on improved decision readability levels, then one would expect to see a drop in MADRS scores within any future studies that use the MADRS formula. Such a change would be welcome to any of us who have struggled to

understand past decisions, and would reflect positively on Canada in terms of our commitment to access to justice.

Appendices

Appendix 1 – Office of Research Ethics Certificate of Approval

Université d'Ottawa
Bureau d'éthique et d'intégrité de la recherche

University of Ottawa
Office of Research Ethics and Integrity

S-06-23-9339 - REG-9339 - Certificat d'approbation éthique / Certificate of Ethics Approval

(English message follows)

Cher/Chère Michael Madden,

Veuillez trouver ci-joint le certificat d'approbation éthique pour le projet intitulé «Putting Numbers to Words: Quantitatively Measuring the Readability of Court and Administrative Tribunal Decisions in Canada».

Le certificat est valide jusqu'au : 13-06-2024

Recherche financée : veuillez faire suivre une copie du certificat au Service de gestion de la recherche.

Si vous avez des questions, n'hésitez pas à communiquer avec le Bureau d'éthique à ethique@uottawa.ca ou en composant le 613-562-5387.

Vous pouvez voir votre demande en vous connectant à votre compte eReviews.

Cordialement,

Riana Marcotte
Responsable d'éthique en recherche

Ceci est une réponse automatisée, merci de ne pas répondre à ce courriel.

Dear Michael Madden,

Please find attached the certificate of ethics approval for your research project titled "Putting Numbers to Words: Quantitatively Measuring the Readability of Court and Administrative Tribunal Decisions in Canada".

This certificate is valid until: 13-06-2024

Funded research: A reminder that you must provide a copy of this certificate to Research Management Services.

If you have any questions, please contact the Ethics Office at ethics@uottawa.ca or by telephone at 613-562-5387.

You can view your project at any time by logging into eReviews.

Best regards,

Riana Marcotte
Protocol Officer

This is an automated message. Please do not reply directly to this email.

Attachement(s) / Attachment(s)
approvalLetter1686771229478.pdf

550, rue Cumberland, pièce 154
Ottawa (Ontario) K1N 6N5 Canada

550 Cumberland Street, Room 154
Ottawa, Ontario K1N 6N5 Canada

613-562-5387 • 613-562-5338 • ethique@uOttawa.ca / ethics@uOttawa.ca
www.recherche.uottawa.ca/deontologie I www.recherche.uottawa.ca/ethics

APPENDIX 2 – BIOGRAPHIES OF EXPERT RATERS

Johana Bearden is a Staff Lawyer with Nova Scotia Legal Aid. She received her LL.B. from the University of New Brunswick in 2009 and was called to the bar in 2010. She has practiced criminal defense since then and specialized in representing youth since 2015.

James Beaton received a B.A. in International Relations (2013) and a Juris Doctor (2016) from the University of British Columbia. He was called to the bar of Alberta in 2017. James practised law in the private and public sectors before becoming a federal tribunal member in 2021. He is an award-winning author, speaker, and researcher who enjoys volunteering in his community to improve access to justice.

Ryan Fritsch received a LL.M. (McGill 2010), LL.B. (Windsor 2005) and B.A. (Victoria 2002). He was called to the bar of British Columbia in 2006 and Ontario in 2007. Ryan served as in-house counsel to the Psychiatric Patient Advocate Office (2008-2011); developed the Legal Aid Ontario Mental Health Strategy (2012-2016); and now leads law reform projects related to consumer protection, artificial intelligence and Indigenous health law at the Law Commission of Ontario (2016-present).

Seamus Murphy received his LL.B. from Dalhousie in 2012, following the completion of his BA (Honours) from Queen's University. He has been a member of the Nova Scotia Barrister's Society since 2018, having previously been called to the bar in Ontario in 2014. He practised immigration law when in Ontario and for the past six years has been practicing family law. He is currently the managing lawyer of the Yarmouth office of Nova Scotia Legal Aid.

Sarah Orr received her Bachelor of Arts (Honours) degree in International Development Studies from McGill University in 2007, and her Juris Doctor from the University of Victoria in 2011. She was called to the British Columbia bar in 2012. She started her career in private practice before working as an Industrial Relations Officer at BC's Employment Standards Branch where she mediated and adjudicated employment disputes. Since 2018 she has been a Tribunal Member at BC's Civil Resolution Tribunal where she adjudicates small claims disputes under $5,000, strata disputes, and motor vehicle disputes.

Bibliography

LEGISLATION

Bill C-27*, An Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to make consequential and related amendments to other Acts,* 1st Sess, 44th Parl, cl 39, 2022 (died on the order paper on January 6, 2025).

Canadian Charter of Rights and Freedoms, Part I of the *Constitution Act, 1982*, being Schedule B to the *Canada Act 1982 (UK)*, 1982, c 11.

Canadian Human Rights Act, RSC 1985, c H-6.

The Constitution Act, 1867, 30 & 31 Vict, c 3.

Department of Employment and Social Development Act, SC 2005, c 34.

Federal Courts Act, RSC, 1985, c F-7.

Immigration and Refugee Protection Act, SC 2001, c 27.

Judges Act, RSC, 1985, c J-1.

National Defence Act, RSC 1985 c N-5.

Residential Tenancies Act, 2006, SO 2006, c 17.

Supreme Court Act, RSC, 1985, c S-26.

Youth Criminal Justice Act, SC 2002, c 1.

JURISPRUDENCE

Baker v Canada (Minister of Citizenship and Immigration), [1999] 2 SCR 817.

Bashir v Canada (Citizenship and Immigration), 2025 FC 19.

City of Medicine Hat v Medicine Hat Fire Fighters Association, IAFF Local 263, 2014 CanLII 50000 (AB GAA).

CN v Canada (Canadian Human Rights Commission), [1987] 1 SCR 1114.

Donoghue v Stevenson, [1932] UKHL 100 (UK).

IWA v Consolidated-Bathurst Packaging Ltd, [1990] 1 SCR 282.

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

Palsgraf v Long Island Railroad Co, 298 NY 339 (App Ct 1928) (USA).

R v Fearon, 2014 SCC 77.

R v Grant, 2009 SCC 32.

R v Hape, 2007 SCC 26.

R v Oakes, [1986] 1 SCR 103.

R v Shafia, 2016 ONCA 812.

R v Sheppard, 2002 SCC 26.

R v Sullivan, 2022 SCC 19

Re Onischuk, 2017 ABQB 663.

Reference re Remuneration of Judges of the Provincial Court (PEI), [1997] 3 SCR 3.

Snell v United Special Insurance Company, No 22-12581, United States Court of Appeals, Eleventh Circuit (May 28, 2024).

South West Terminal Ltd v Achter Land, 2023 SKKB 116.

Springs of Living Water Centre Inc. v The Government of Manitoba, 2020 MBQB 185.

Terra Marine Industries Ltd v N & N Electric Motor Specialists Ltd, 1991 CanLII 13663 (NS SC).

Winnipeg School Division No 1 v Craton, [1985] 2 SCR 150.

SECONDARY SOURCES AND OTHER MATERIALS

Adalberto Jordan. “Imagery, Humor, and the Judicial Opinion” (1987) 41:3 University of Miami Law Review 693.

Adler, Mark. “The Plain Language Movement” in Lawrence M Solan & Peter M Tiersma, eds, *The Oxford Handbook of Language and Law* (London: Oxford UP, 2018) 67.

Amendum, Steven J, Kristin Conradi & Elfrieda Hiebert. “Does Text Complexity Matter in the Elementary Grades? A Research Synthesis of Text Difficulty and Elementary Students’ Reading Fluency and Comprehension” (2018) 30 Educational Psychology Review 121.

Anderson, Elizabeth S. “What Is the Point of Equality?” (1999) 109 Ethics 287.

Anleu, Sharyn Roach, Kathy Mack & Jordan Tutton. “Judicial Humour in the Australian Courtroom” (2014) 38:2 Melbourne University Law Review 621.

Arnon, Inbal & Neal Snider. “More Than Words: Frequency Effects for Multi-Word Phrases” (2010) 62 Journal of Memory and Language 67.

Ash, Elliott, et al. “Translating Legalese: Enhancing Public Understanding of Court Opinions with Legal Summarizers” in *CSLAW ’24: Proceedings of the Symposium on Computer Science and Law* (New York: Association for Computing Machinery, 2024) 136.

Assy, Rabeea. “Can the Law Speak Directly to Its Subjects? The Limitation of Plain Language” (2011) 38:3 Journal of Law & Society 376.

Azuelos-Atias, Sol. “On the Incoherence of Legal Language to the General Public” (2011) 24 International Journal for the Semiotics of Law 41.

Bailin, Alan & Ann Grafstein. *Readability: Text and Context* (New York: Palgrave Macmillan, 2016).

Balyan, Renu, Kathryn S McCarthy & Danielle S McNamara. “Applying Natural Language Processing and Hierarchical Machine Learning Approaches to Text Difficulty Classification” (2020) 30 International Journal of Artificial Intelligence in Education 337.

Bardovi-Harlig, Kathleen. “A Second Look at T-Unit Analysis: Reconsidering the Sentence” (1992) 26:2 TESOL Quarterly 390.

Barnes, Jeffrey. “When Plain Language Legislation is Ambiguous – Sources of Doubt and Lessons for the Plain Language Movement” (2010) 34:3 Melbourne University Law Review 671.

Basch, Corey H, et al. “Public Health Communication in Time of Crisis: Readability of On-Line COVID-19 Information” (2020) 14:5 Disaster Medicine and Public Health Preparedness 635.

Baud, Patrick F. “The Crown’s Prerogatives and the Constitution of Canada” (2021) 3 Journal of Commonwealth Law 219.

Beach, Kristen L. *The Effect of Media, Text Length, and Reading Rates on College Student Reading Comprehension Levels*, (unpublished Doctor of Business Administration Dissertation, University of Phoenix, 2008).

Benjamin, Rebekah George. "Reconstructing Readability: Recent Developments and Recommendations in the Analysis of Text Difficulty" (2012) 24 Educational Psychology Review 63.

Benoliel, Uri & Shmuel I Becher. "The Duty to Read the Unreadable" (2019) 60:8 Boston College Law Review 2255.

Benson, Robert W & Joan B Kessler. "Legalese v. Plain English: An Empirical Study of Persuasion and Credibility in Appellate Brief Writing" (1987) 20:2 Loyola of Los Angeles Law Review 301.

Berk, Kenneth N. "Validating Regression Procedures With New Data" (1984) 26:4 Technometrics 331.

Bildfell, Connor. "The Role of Chief Justices in British Columbia's Superior Courts" (2018) 76:1 The Advocate 25.

Binder, Guyora & Robert Weisberg. *Literary Criticisms of Law*. (Princeton: Princeton UP, 2000).

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

Bizzoni, Yuri, et al. "Good Reads and Easy Novels: Readability and Literary Quality in a Corpus of US-published Fiction" in Tanel Alumäe & Mark Fishel, eds, *Proceedings of the 24th Nordic Conference on Computational Linguistics* (Estonia: University of Tartu Library, 2023) 42, online: <https://aclanthology.org/2023.nodalida-1.5.pdf>.

Blaisie, Michael A. "Regulating Plain Language" [2023] Wisconsin Law Review 687.

Bormuth, John R. "Cloze Test Readability: Criterion Reference Scores" (1968) 5:3 Journal of Educational Measurement 189.

Bouck, John C. "Civil Jury Trials – Assessing Non-Pecuniary Damages – Civil Jury Reform" (2002) 81:3 Canadian Bar Review 493.

Bowie, Jennifer & Elisha Carol Savchak. "State Court Influence on US Supreme Court Opinions" (2022) 10:1 Journal of Law & Courts 139.

Braaksma, Martine AH, et al. "What Observational Learning Entails: A Multiple Case Study" (2006) 6:1 L1 – Educational Studies in Language and Literature 31.

Brown, David M. "The Five Habits of Highly Effective and Accessible Civil Courts: Transparency; Perspicacity; Creativity; Simplicity; and Accountability. Reflections at the End of a Judicial Career" in *International Access to Justice Forum* (Toronto: 25 Oct 2024).

Bruggink, Marian, et al. *Putting PIRLS to Use in Classrooms Across the Globe: Evidence-Based Contributions for Teaching Reading Comprehension in a Multilingual Context* (Nijmegen, The Netherlands: Springer, 2022).

Brysbaert, Marc. "How Many Words do we Read per Minute? A Review and Meta-analysis of Reading Rate" (2019) 109 Journal of Memory and Language 104047.

Budak, Tamer, Simon James & Adrian Sawyer. "The Complexity of Tax Simplification: Experiences From Around the World" in Tamer Budak, Simon James & Adrian Sawyer (eds), *The Complexity of Tax Simplification: Experiences From Around the World* (New York: Palgrave Macmillan, 2016), 1.

Burrows, Andrew Stephen. "Judgment-Writing: A Personal Perspective", Annual Conference of Judges of the Superior Courts in Ireland (20 May 2021), online: <https://www.supremecourt.uk/docs/judgment-writing-a-personal-perspective-lord-burrows.pdf>.

Burton, Gene E. "The Readability of Consumer-Oriented Bank Brochures: An Empirical Investigation" (1991) 30:1 Business & Society 21.

Canada, Department of Justice. *The Honourable April D Grosse's Questionnaire (Questionnaire for the Alberta Court of Queen's Bench Judicial Appointment Process)* (11 July 2018), online: <https://www.canada.ca/en/department-justice/news/2018/07/the-honourable-april-d-grosses-questionnaire.html>.

---. *The Honourable Anne Louise Kirker's Questionnaire (Questionnaire for the Alberta Court of Queen's Bench Judicial Appointment Process)* (13 Nov 2018), online: <https://www.canada.ca/en/department-justice/news/2018/11/the-honourable-anne-louise-kirkers-questionnaire.html>.

---. *The Honourable Barbara J Norell's Questionnaire (Questionnaire for the Supreme Court of British Columbia Judicial Appointment Process)* (13 Apr 2018), online: <https://www.canada.ca/en/department-justice/news/2018/04/the-honourable-barbara-j-norells-questionnaire.html>.

---. *The Honourable Cynthia Petersen's Questionnaire (Questionnaire for the Ontario Superior Court of Justice Judicial Appointment Process)* (last modified on 30 May 2022), online: <https://www.canada.ca/en/department-justice/news/2022/05/the-honourable-cynthia-petersens-questionnaire.html>.

---. *The Honourable Jill Miriam Copeland's Questionnaire (Questionnaire for the Court of Appeal for Ontario's Judicial Appointment Process)* (30 May 2022), online: <https://www.canada.ca/en/department-justice/news/2022/05/the-honourable-jill-miriam-copelands-questionnaire.html>.

---. *The Honourable Mahmud Jamal's Questionnaire (Questionnaire for the Supreme Court of Canada Judicial Appointment Process)* (17 Jun 2021), online: <https://fja-cmf.gc.ca/scc-csc/2021/nominee-candidat-eng.html>.

---. *The Honourable Malcolm Rowe's Questionnaire, (Questionnaire for the Supreme Court of Canada Judicial Appointment Process)* (4 July 2017), online: <https://www.fja-cmf.gc.ca/scc-csc/2016-MalcolmRowe/nominee-candidat-eng.html>.

---. *The Honourable Mary Moreau's Questionnaire (Questionnaire for the Supreme Court of Canada Judicial Appointment Process)* (26 Oct 2023) online: <https://fja-cmf.gc.ca/scc-csc/2023/nominee-candidate-eng.html>.

---. *The Honourable Michael Brundrett's Questionnaire (Questionnaire for the British Columbia Supreme Court Judicial Appointment Process)* (2 Nov 2017), online: <https://www.canada.ca/en/department-justice/news/2017/11/the_honourable_michaeljbrundrettsquestionnaire.html>.

---. *The Honourable Michele Hollins' Questionnaire (Questionnaire for the Alberta Court of Queen's Bench Judicial Appointment Process)* (31 Mar 2017), online: <https://www.canada.ca/en/department-justice/news/2017/03/the_honourable_justicemichelehhollinssquestionnaire.html>.

---. *The Honourable Michelle O'Bonsawin's Questionnaire (Questionnaire for the Supreme Court of Canada Judicial Appointment Process)* (22 Aug 2022), online: <https://fja-cmf.gc.ca/scc-csc/2022/nominee-candidat-eng.html>.

---. *The Honourable Nicholas Kasirer's Questionnaire (Questionnaire for the Supreme Court of Canada Judicial Appointment Process)* (18 April 2019), online: <https://www.fja.gc.ca/scc-csc/2019/nominee-candidat-eng.html>.

---. *The Honourable Renu Mandhane's Questionnaire (Questionnaire for the Ontario Superior Court of Justice Judicial Appointment Process)* (28 Mar 2022), online: <https://www.canada.ca/en/department-justice/news/2022/03/the-honourable-renu-mandhanes-questionnaire.html>.

---. *The Honourable Sally A Gomery's Questionnaire (Questionnaire for the Ontario Superior Court of Justice Judicial Appointment Process)* (3 Nov 2017), online: <https://www.canada.ca/en/department-justice/news/2017/11/the_honourable_sallyagomerysquestionnaire.html>.

---. *The Honourable Sebastian Grammond's Questionnaire (Questionnaire for the Federal Court Judicial Appointment Process)* (8 Jun 2021), online: <https://www.canada.ca/en/department-justice/news/2021/06/the-honourable-sebastien-grammond-questionnaire.html>.

---. *The Honourable Shaun S. Nakatsuru's Questionnaire (Questionnaire for the Ontario Superior Court of Justice Judicial Appointment Process)* (11 May 2017), online: <https://www.canada.ca/en/department-justice/news/2017/05/the_honourable_justiceshaunsnakatsurusquestionnaire.html>.

---. *The Honourable Sheilah Martin's Questionnaire (Questionnaire for the Supreme Court of Canada Judicial Appointment Process)* (21 December 2021), online: <https://www.fja.gc.ca/scc-csc/2017-SheilahMartin/nominee-candidat-eng.html>.

---. *The Honourable Steven Wilson's Questionnaire (Questionnaire for the British Columbia Supreme Court Judicial Appointment Process)* (1 Mar 2017), online: <https://www.canada.ca/en/department-justice/news/2019/02/the-honourable-steven-wilsons-questionnaire.html>.

---. *The Honourable Warren B Milman's Questionnaire (Questionnaire for the British Columbia Supreme Court Judicial Appointment Process)* (3 Nov 2017), online: <https://www.canada.ca/en/department-justice/news/2017/11/the_honourable_warrenbmilmansquestionnaire.html>.

Canada, Prime Minister's Office. "Prime Minister announces appointment of new Associate Chief Justice of the Court of King's Bench of Alberta" (October 10, 2023), online: <https://www.pm.gc.ca/en/news/news-releases/2023/10/10/prime-minister-announces-appointment-new-associate-chief-justice>.

Canadian Human Rights Tribunal. *Canadian Human Rights Tribunal, Annual Report 2023* (Minister of Public Works and Government Services, 2024), online: <https://www.chrt-tcdp.gc.ca/sites/default/files/2024-04/CHRT-AnnualReport-2023-EN.pdf>.

---. "Practice Direction: Use of Artificial Intelligence (AI) in Tribunal Proceedings", online: <https://chrt-tcdp.gc.ca/en/about-us/practice-direction-use-artificial-intelligence-ai-tribunal-proceedings>.

---. "Welcome to the Canadian Human Rights Tribunal", online: <https://www.chrt-tcdp.gc.ca/en>.

Canadian Institute for the Administration of Justice. "Mastering the Skill of Judgment Writing (2024 Edition): Joint program with the NJI", online: <https://ciaj-icaj.ca/en/upcoming-programs/mastering-judgment-writing-2024/>.

---. *Report on the Results of a Survey of Canadian Judges Concerning Technology in Judging Conducted by the Canadian Institute for the Administration of Justice and Athabasca University in 2022*, online: <https://ciaj-icaj.ca/wp-content/uploads/page/2023/08/ciaj-athabasca_2023-report-on-canadian-judges-and-technology_en.pdf>.

Canadian Judicial Council. "Statement of Principles on Self-represented Litigants and Accused Persons" (September 2006), online: <https://cjc-ccm.ca/sites/default/files/documents/2020/Final-Statement-of-Principles-SRL.pdf>.

CanLII. “CanLII Launches Nationwide Survey to Assess AI-Generated Legal Summaries” (1 Oct 2024), online: <https://blog.canlii.org/2024/10/01/canlii-launches-nationwide-survey-to-assess-ai-generated-legal-summaries/>.

Cappelletti, Mauro. “Alternative Dispute Resolution Processes within the Framework of the World-Wide Access-to-Justice Movement” (1993) 56 Modern Law Review 282.

Cardozo, Benjamin Nathan. “Law and Literature” in Margaret Hall, ed, *Selected Writings of Benjamin Nathan Cardozo* (New York: Matthew Bender & Co, 1947).

Carpenter, Anna E. “Active Judging and Access to Justice” (2017) 93 Notre Dame Law Review 647.

Chall, Jeanne S. “Varying Approaches to Readability Measurement” (1996) 25:1 Revue québécoise de linguistique 23.

Chall, Jeanne S & Edgar Dale. *Readability Revisited: The New Dale-Chall Readability Formula*. (Cambridge, MA: Brookline Books, 1995)

Charrow, Robert P & Veda R Charrow. “Making Legal Language Understandable: A Psycholinguistic Study of Jury Instructions” (1979) 79:7 Columbia Law Review 1306.

Chatman, Seymour. “Pre-adjectivals in the English Nominal Phrase” (1960) 35:2 American Speech 83.

Chebat, Jean-Charles. “Testing Consumers’ Motivation and Linguistic Ability as Moderators of Advertising Readability” (2003) 20:7 Psychology & Marketing 599.

Chen, Chih-Ming & Yu-Ju Lin. “Effects of Different Text Display Types on Reading Comprehension, Sustained Attention and Cognitive Load in Mobile Reading Contexts” (2016) 24:3 Interactive Learning Environments 553.

Chen, Xiaobin & Detmar Meurers. “Word Frequency and Readability: Predicting the Text-Level Readability with a Lexical-Level Attribute” (2018) 41:3 Journal of Research in Reading 486.

Choi, Joon Suh & Scott A Crossley. “Advances in Readability Research: A New Readability Web App for English” in *Proceedings of the Twenty-second IEEE International Conference on Advanced Learning Technologies (ICALT 2022)*, 6.

Cohen, Jacob. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed (New York: Lawrence Erlbaum Associates, 1988).

Coleman, Brady. “Lord Denning & Justice Cardozo: The Judge as Poet-Philosopher” (2001) 32:2 Rutgers Law Journal 485.

Collins-Thompson, Kevyn. “Computational Assessment of Text Readability: A Survey of Current and Future Research” (2014) 165:2 International Journal of Applied Linguistics 97.

Coltheart, Max. “The MRC Psycholinguistic Database” (1981) 33A Quarterly Journal of Experimental Psychology 497.

Cotterrell, Roger. “Access to Justice, Moral Distance and Changing Demands on Law” (2019) 36 Windsor Yearbook of Access to Justice 193.

Council of Canadian Administrative Tribunals. “Decision Writing Seminar 2024”, online: <https://www.ccat-ctac.org/national-administrative-law-week/decision-writing-seminar-june-2024/>.

Craig, Elaine. “Judicial Audiences: A Case Study of Justice David Watt’s Literary Judgments” (2018) 64:2 McGill Law Journal 309.

Crawford, Kate. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence* (New Haven: Yale UP, 2021).

Crossley, Scott A. “Limitations of Natural Language Processing Tools for Data Mining: Text Length and Grade Level Differences” in Kristy Elizabeth Boyer & Michael Yudelson, eds, *Proceedings of the 11th International Conference on Educational Data Mining* (Buffalo, NY: International Conference on Educational Data Mining, 2018), 630.

Crossley, Scott A, David B Allen & Danielle S McNamara. “Text Readability and Intuitive Simplification: A Comparison of Readability Formulas” (2011) 23:1 Reading in a Foreign Language 84.

Crossley, Scott A, et al. “A Large-Scaled Corpus for Assessing Text Readability” (2022) Behaviour Research Methods, online: <https://link.springer.com/article/10.3758/s13428-022-01802-x>.

---. “Predicting the Readability of Physicians’ Secure Messages to Improve Health Communication Using Novel Linguistic Features: Findings from the ECLIPPSE Study” (2020) 13:4 Journal of Communication in Healthcare 344.

Crossley, Scott A, 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.

Crossley, Scott A, Stephen Skalicky & Mihai Dascalu. “Moving Beyond Classic Readability Formulas: New Methods and New Models” (2019) 42 Journal of Research in Reading 541.

Crossley, SA, K Kyle & DS McNamara. “The Tool for the Automatic Analysis of Text Cohesion (TAACO): Automatic Assessment of Local, Global, and Text Cohesion” (2016) 48 Behavior Research Methods 1227.

Cunningham, James W, Elfrieda H Hiebert & Heidi Anne Mesmer. “Investigating the Validity of Two Widely Used Quantitative Text Tools” (2018) 31 Reading and Writing 813.

Curtotti, Michael & Eric McCreath. “A Right to Access Implies a Right to Know: An Open Online Platform for Research on the Readability of Law” (2013) 1:1 Journal of Open Access Law 1.

Dale, Edgar & Jeanne S Chall. “A Formula for Predicting Readability” (1948) 27:1 Educational Research Bulletin 11.

Dale, Edgar & Ralph W Tyler. “A Study of the Factors Influencing the Difficulty of Reading Materials for Adults of Limited Reading Ability” (1934) 4:1 The Library Quarterly 384.

Danet, Brenda. “Language in the Legal Process” (1980) 14 Law & Society Review 445.

Danielson, Wayne A & Dominic L Lasorsa. “A New Readability Formula Based on the Stylistic Age of Novels” (1989) 33:3 Journal of Reading 194.

Darvil, Kathleen. “Increasing Access to Justice by Improving Usability of Statutory Code Websites” (2023) 115 Law Library Journal 123.

Dattalo, Patrick. *Determining Sample Size: Balancing Power, Precision, and Practicality* (New York: Oxford UP, 2008).

Davies, Breese. “Going Broke in Legal Aid: The High Cost of Cheap Justice,” The Walrus (Sept 2016), online: <https://thewalrus.ca/going-broke-in-legal-aid/>.

Davies, Mark. “The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English” (2010) 25 Literary & Linguistic Computing 447.

Davis, Frederick B. “Psychometric Research on Comprehension in Reading” (1972) 7:4 Reading Research Quarterly 628.

Davis, Scott W, et al. “Say What? How the Interplay of Tweet Readability and Brand Hedonism Affects Consumer Engagement” (2019) 100 Journal of Business Research 150.

Davison, Alice & Robert N Kantor. “On the Failure of Readability Formulas to Define Readable Texts: A Case Study from Adaptations” (1982) 17:2 Reading Research Quarterly 187.

DeClercq, Orphé, et al. “Using the Crowd for Readability Prediction” (2012) 20:3 Natural Language Engineering 293.

DeFriez, Brian M. *Toward a Clearer Democracy: The Readability of Idaho Supreme Court Opinions as a Measure of the Court's Democratic Legitimacy* (2017) (Unpublished Ph.D. Dissertation, University of Idaho's College of Graduate Studies).

Delage, Hélène & Ulrich Hans Frauenfelder. "Syntax and Working Memory in Typically-Developing Children: Focus on Syntactic Complexity" (2019) 10:2 Language, Interaction and Acquisition 141.

Delgado, Pablo, et al. "Don't Throw Away Your Printed Books: A Meta-Analysis on the Effects of Reading Media on Reading Comprehension" (2018) 25 Educational Research Review 23.

Deroy, Aniket, et al. "An Analytical Study of Algorithmic and Expert Summaries of Legal Cases" in Erich Schweighofer, ed, *Legal Knowledge and Information Systems* (Amsterdam: IOS Press, 2021) 90.

Deslongchamps, Alexandre. "Readability and the Bank of Canada," Bank of Canada Staff Analytical Note 2018-20, June 2018, online: <https://www.bankofcanada.ca/2018/06/staff-analytical-note-2018-20>.

Douven, Igor. "A Bayesian Perspective on Likert Scales and Central Tendency" (2018) 25:3 Psychonomic Bulletin & Review 1203.

Dubay, William H. *The Principles of Readability* (Impact Information: California, 2004).

Duckworth, Mark. "Clarity and the Rule of Law: The Role of Plain Judicial Language" in Ruth Sheard, ed, *A Matter of Judgment: Judicial Decision-Making and Judgment Writing* (Sydney: Judicial Commission of New South Wales, 2003) 91.

Duggan, Anthony & Iain Ramsay. "Front-end Strategies for Improving Consumer Access to Justice" in Michael J Trebilcock, Lorne Mitchell Sossin & Anthony J Duggan, eds, *Middle Income Access to Justice* (Toronto: U of T Press, 2012) 95

Dyer, Charles R, et al. "Improving Access to Justice: Plain Language Family Law Court Forms in Washington State" (2013) 11:3 Seattle Journal of Social Justice 1065.

Economides, Kim, Aaron Timoshanko & Leslie S Ferraz. "Justice at the Edge: Hearing the Sound of Silence" (2020) 41 Adelaide Law Review 39.

Edmunds, Matthew R, Robert J Barry & Alastair K Denniston. "Readability Assessment of Online Ophthalmic Patient Information" (2013) 131:12 JAMA Ophthalmology 1610.

Enarsson, Therese, Lena Enqvist & Markus Naarttijärvi. "Approaching the Human in the Loop – Legal Perspectives on Hybrid Human/Algorithmic Decision-Making in Three Contexts" (2022) 31:1 Information & Communications Technology Law 123.

Engler, Russell. “Opportunities and Challenges: Non-lawyer Forms of Assistance in Providing Access to Justice for Middle-income Earners” in Michael J Trebilcock, Lorne Mitchell Sossin & Anthony J Duggan, eds, *Middle Income Access to Justice* (Toronto: U of T Press, 2012) 145.

Farajidizaji, Asma, Vatsal Raina & Mark Gales. “Is it Possible to Modify Text to a Target Readability Level? An Initial Investigation Using Zero-Shot Large Language Models” (2024) *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation* (LREC-COLING 2024) 9235, online: <https://aclanthology.org/2024.lrec-main.815.pdf>.

Federal Judicial Centre. *Judicial Writing Manual: A Pocket Guide for Judges*, 2nd ed (Federal Judicial Center, 2013).

Federation of Law Societies of Canada. *Model Code of Professional Conduct, as Amended April 2024*, online: <https://flsc.ca/wp-content/uploads/2024/05/2024-Model-Code-of-Professional-Conduct.pdf>.

Feng, C-X J, Z-G S Yu & J-H J Wang. “Validation and Data Splitting in Predictive Regression Modeling of Honing Surface Roughness Data” (2005) 43:8 International Journal of Production Research 1555.

Ferguson, Robert A. “The Judicial Opinion as Literary Genre” (1990) 2 Yale Journal of Law & Humanities 201.

Finegan, Edward & Benjamin T Lee. “Legal Writing: Attitude and Emphasis – Corpus Linguistic Approaches to ‘Legal Language’: Adverbial Expression of Attitude and Emphasis in Supreme Court Opinions” in Malcolm Coulthard, Alison May & Rui Sousa-Silva, eds, *The Routledge Handbook of Forensic Linguistics*, 2nd ed (New York: Routledge, 2020) 48.

Finkelstein, Michael O & Bruce Levin. *Statistics for Lawyers*, 3rd ed (New York: Springer, 2015).

Finnis, John. *Natural Law and Natural Rights* (Oxford: Oxford UP, 2011).

Flammer, Sean. “An Empirical Analysis of Writing Style, Persuasion, and the Use of Plain Language” (2010) 16 Journal of the Legal Writing Institute 183.

Flesch, Rudolph. “A New Readability Yardstick” (1948) 32 Journal of Applied Psychology 221.

Flor, Michael & Beata Beigman Klebanov. “Associative Lexical Cohesion as a Factor in Text Complexity” (2014) 165:2 International Journal of Applied Linguistics 223.

François, Thomas & Eleni Miltsakaki. “Do NLP and machine learning improve traditional readability formulas?” in *Proceedings of the First Workshop on Predicting and Improving Text Readability for Target Reader Populations* (Montreal: Association for Computational Linguistics, 2012) 49.

Franklin, Kris. "*See Erie*: Critical Study of Legal Authority" (2008) 31:1 University of Arkansas at Little Rock Law Review 109.

Frohock, Christina M. "Schrödinger's Dissent: The Hybrid Authority of a Dissenting Opinion" (2024) 107:4 Marquette Law Review 963.

Frost, Elizabeth Ruiz. "Overcoming 'Illusory Superiority': Mentoring Legal Writers" (2013) 73:9 Oregon State Bar Bulletin 13.

Fuller, Lon L. *The Morality of Law*. Revised ed (New Haven: Yale UP, 1971).

García-Crespo, Ángel, et al. "A Qualitative Study of Hard Decision Making in Managing Global Software Development Teams" (2010) 27 Information Systems Management 247.

Garner, James, Scott Crossley & Kristopher Kyle. "N-Gram Measures and L2 Writing Proficiency" (2019) 80 System 176.

Goerlandt, Floris & Huiyan Liu. "Readability of Maritime Accident Reports: A Comparative Analysis" (2023) Maritime Policy & Management 1, online: <https://doi.org/10.1080/03088839.2023.2166685>.

Goldstein, Jospeh. *The Intelligible Constitution: The Supreme Court's Obligation to Maintain the Constitution as Something We the People Can Understand* (New York: Oxford UP, 1992).

Goriely, Tamara. "The English Legal Aid White Paper and the LAG Conference" (1996) 3:3 International Journal of the Legal Profession 353.

Gorman, Wayne K. "Access to Justice in Canada: The Unrepresented Accused and the Role of the Trial Judge" (2020) 56:1 Court Review 48.

Gotti, Maurizio. "Text and Genre" in Lawrence Solan & Peter Meijes Tiersma, eds, *The Oxford Handbook of Language and Law* (Oxford: Oxford University Press, 2012) 52.

Government of Ontario. "Attorney General Announces Four Judicial Appointments" (January 18, 2008), online: <https://news.ontario.ca/en/release/87966/attorney-general-announces-four-judicial-appointments>.

Gowder, Paul. "Is Legal Cognition Computational? (When will DeepVehicle replace Judge Hercules?)" in Ryan Whalen, ed, *Computational Legal Studies: The Promise and Challenge of Data-Driven Research* (Cheltenham, UK: Elgar, 2020) 215.

Goździ-Roszkowski, Stanisław. "Communicating Dissent in Judicial Opinions: A Comparative, Genre-Based Analysis" (2020) 33 International Journal for the Semiotics of Law 381.

Graesser, Arthur C, et al. “Coh-Metrix: Analysis of Text on Cohesion and Language” (2004) 36 Behavior Research Methods, Instruments, & Computers 193.

Grey, Alexandra & Laura Smith-Khan. “Bringing Linguistic Research into Legal Scholarship and Practice” (2021) 46 Alternative Law Journal 64.

Griffin, David. “Considering Legal English: Consensus and Complications” (2024) 31:1 International Journal of Speech, Language and the Law 77.

Grossman, Maura R, et al. “The GPTJudge: Justice in a Generative AI World” (2023) 23 Duke Law & Technology Review 1.

Gunning, Robert. *The Technique of Clear Writing* (New York: McGraw Hill, 1952).

Habbig, Ann-Katrin & Ingrid Robeyns. “Legal Capabilities” (2022) 23:4 Journal of Human Development and Capabilities 611.

Hadden, Kristie B, et al. “Improving Readability of Informed Consents for Research at an Academic Medical Institution” (2017) 1 Journal of Clinical and Translational Science 361.

Hahn, Gerald J. “The Hazards of Extrapolation in Regression Analysis” (1977) 9 Journal of Quarterly Technology 159.

Hallgren, Kevin A. “Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial” (2012) 8:1 Tutorials in Quantitative Methods for Psychology 23.

Hamann, Hanjo & Friedemann Vogel. “Evidence-Based Jurisprudence Meets Legal Linguistics: Unlikely Blends Made in Germany” [2017] Brigham Young University Law Review 1473.

Helis, John. *Quasi-Constitutional Laws of Canada* (Toronto: Irwin Law, 2018).

Horwitz, Rosalind & S Jay Samuels. “Reading and Listening to Expository Text” (1985) 17:3 Journal of Reading Behavior 185.

Hughes, Patricia. “Advancing Access to Justice through Generic Solutions: The Risk of Perpetuating Exclusion” (2013) 31 Windsor Yearbook of Access to Justice 1.

Jacobson, Milton D, C Eric Kirkland & Ramsay W Selden. “An Examination of the McCall-Crabbs Standard Test Lessons in Reading” (1978) 22:3 Journal of Reading 224.

Jindal, Pranay & Joy C MacDermid. “Assessing Reading Levels of Health Information: Uses and Limitations of Flesch Formula” (2017) 30:1 Education for Health 84.

Johnson, Richard. “How to Apply the Flesch Kincaid Readability Formula to Your Content,” February 23, 2021, online: <https://www.optimonk.com/how-to-apply-the-flesch-kincaid-readability-formula-to-your-content/>.

Johnson, Stephen M. “The Changing Discourse of the Supreme Court” (2014) 12:1 University of New Hampshire Law Review 29.

Kandula, Sasikiran & Qing Zeng-Treitler. “Creating a Gold Standard for the Readability Measurement of Health Texts” [2008] American Medical Informatics Association Symposium Proceedings 353.

Kaplan, David & Gloria S Waters. “Verbal Working Memory and Sentence Comprehension” (1999) 22 Behavioral & Brain Sciences 77.

Kayam, Orly. “The Readability and Simplicity of Donald Trump’s Language” (2018) 16:1 Political Studies Review 73.

Kelly, David St L. “Legislative Drafting and Plain English” (1986) 10 Adelaide Law Review 409.

Khan, Jon. *‘The Life of a Reserve’: How Might We Improve the Structure, Content, Accessibility, Length & Timeliness of Judicial Decisions?* (unpublished LL.M. Thesis, University of Toronto, 2019), online: <https://tspace.library.utoronto.ca/bitstream/1807/98120/1/Khan_Jon_%20_201911_LLM_thesis.pdf>.

Khanna, Maya M & Michael J Cortese. “How Well Imageability, Concreteness, Perceptual Strength, and Action Strength Predict Recognition Memory, Lexical Decision, and Reading Aloud Performance” (2021) 29 Memory 622.

Kincaid, J P, et al. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel* (Memphis, TN: US Naval Technical Training Command - Research Branch, 1975).

Kintsch, Walter & Douglas Vipond. “Reading Comprehension and Readability in Educational Practice and Psychological Theory” in Lars-Goran Nilsson (ed), *Perspectives on Memory Research* (London: Psychology Press, 1979) 329.

Kirchner, Jesse Saba, Justin Nuger & Yi Zhang. “An Extensible Crosslinguistic Readability Framework” in *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora* (Singapore: ACL Anthology, 2009) 11.

Kitto, Frank. “Why Write Judgments?” in Ruth Sheard, ed, *A Matter of Judgment: Judicial Decision-Making and Judgment Writing* (Sydney: Judicial Commission of New South Wales, 2003) 69.

Klare, George R. “The Measurement of Readability: Useful Information for Communicators” (2000) 24:3 ACM Journal of Computer Documentation 11.

Kowalski, Mitch. “When Judges Make Readable Decisions” Financial Post (5 Apr 2013), online: <https://financialpost.com/legal-post/when-judges-make-readable-decisions>.

Kraska-Miller, M. *Nonparametric Statistics for Social and Behavioral Sciences* (Boca Raton, FL: Taylor & Francis, 2014).

Kruse, Jessica, et al. “Readability, Content, and Quality of COVID-19 Patient Education Materials from Academic Medical Centers in the United States” (2021) 49 American Journal of Infection Control 690.

Kuperman, Victor, Hans Stadthagen-Gonzalez & Marc Brysbaert. “Age-of-Acquisition Ratings for 30,000 English Words” (2012) 44 Behavior Research Methods 978.

Kyle, K & SA Crossley. “Measuring Syntactic Complexity in L2 Writing Using Fine-Grained Clausal and Phrasal Indices” (2018) 102:2 Modern Language Journal 333.

Kyle, Kristopher. *Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication* (unpublished Ph.D. Dissertation, Georgia State University, 2016), online: <http://scholarworks.gsu.edu/alesl_diss/35>.

Kyle, Kristopher, Scott Crossley & Cynthia Berger. “The Tool for the Automatic Analysis of Lexical Sophistication (TAALES): Version 2.0” (2018) 50 Behavior Research Methods 1030.

Lacey, Nicola. “Theories of Justice and the Welfare State” (1992) 1 Social & Legal Studies 323.

Landis, J Richard & Gary G Koch. “The Measurement of Observer Agreement for Categorical Data” (1977) 33:1 Biometrics 159.

Lang, Iain A, et al. “Jargon and Readability in Plain Language Summaries of Health Research: Cross-Sectional Observational Study” (2025) 27 Journal of Medical Internet Research 1.

Lebovits, Gerald, Alifya V Curtin & Lisa Solomon. “Ethical Judicial Opinion Writing” (2008) 21 Georgetown Journal of Legal Ethics 237.

Lee, Thomas R & Stephen C Mouritsen. “The Corpus and the Critics” (2021) 88:2 University of Chicago Law Review 275.

Leung, Marlene. “‘Transformative’ decision: How a judge’s decision is turning legal heads,” CTV News (9 Mar 2015), online: <https://www.ctvnews.ca/canada/transformative-decision-how-a-judge-s-decision-is-turning-legal-heads-1.2270358>.

LexUM. “AI-Powered Case Analysis Added on CanLII for All Case Law from Alberta, Saskatchewan, Manitoba, and Prince Edward Island” (3 Jun 2024), online:

<https://lexum.com/en/blog/ai-powered-case-analysis-added-on-canlii-for-all-case-law-from-alberta-saskatchewan-manitoba-and-prince-edward-island/>.

L'Heureux-Dube, Claire. "The Dissenting Opinion: Voice of the Future" (2000) 38:3 Osgoode Hall Law Journal 495.

Lin, Zhiquing. "A Methodological Review of Machine Learning in Applied Linguistics" (2021) 14:1 English Language Teaching 74.

Livermore, Michael A. "Rule by Rules" in Ryan Whalen, ed, *Computational Legal Studies: The Promise and Challenge of Data-Driven Research* (Cheltenham, UK: Elgar, 2020) 238.

Livermore, Michael A & Bao Kham Chau. "Studying Judicial Behaviour with Text Analysis" in Lee Epstein et al (eds), *The Oxford Handbook of Comparative Judicial Behaviour* (Oxford: Oxford UP, 2024).

Lochner, Lance. "Education and Crime" in Penelope Peterson, Eva Baker & Barry McGaw (eds), *International Encyclopedia of Education* (3rd ed) (London: Elsevier Science, 2010), 239.

Lorge, Irving. "Predicting Readability" (1944) 45:6 Teachers College Record 404.

Ma, Lilian, et al. *A National Survey of Tribunal Responsiveness to Self-Represented Parties: Measuring Access to Justice for Canadian Administrative Tribunals* (Council of Canadian Administrative Tribunals, 2015), online: <https://www.ccat-ctac.org/wp-content/uploads/2021/11/AJCpaper-FinalPOSTAugust112015.pdf>.

MacDonald, Sean. "Balance and Independence: The Judicial Process in the Charter Era" (2007) 16 Dalhousie Journal of Legal Studies 161.

Macdonald, Roderick A. "Access to Civil Justice", in Peter Cane & Herbert M Kritzer, eds, *The Oxford Handbook of Empirical Legal Research* (Oxford: Oxford UP, 2010) 493.

Macfarlane, Julie. *The National Self-Represented Litigants Project: Identifying and Meeting the Needs of Self-Represented Litigants – Final Report (May 2013)*, online: <https://representingyourselfcanada.com/wp-content/uploads/2016/09/srlreportfinal.pdf>.

Macfarlane, Julie & Charlotte Sullivan. "Tracking the Trends of the Self-Represented Litigant Phenomenon: Data from the National Self-Represented Litigants Project, 2019-2021" (2021), at 12, online: <https://scholar.uwindsor.ca/lawnsrlppubs/12>.

Mackey, Mary Alton & Marilyn Metz. "Ease of Reading of Mandatory Information on Canadian Food Product Labels" (2009) 33 International Journal of Consumer Studies 369.

Madden, Mike. "Judging Readability: A Study of Opinions by Apex-Court Judges from Australia, Canada, South Africa, the United Kingdom, and the United States" (2024) 21 Scribes Journal of Legal Writing 15.

Maldonado, Daniel Bonilla. “The Right to Access to Justice: Its Conceptual Architecture” (2020) 27 Indiana Journal of Global Legal Studies 15.

Mangen, Anne, Bente R Walgermo & Kolbjørn Brønnick. “Reading Linear Texts on Paper Versus Computer Screen: Effects on Reading Comprehension” (2013) 58 International Journal of Educational Research 61.

Mankayi, Sinethemba, Frank Ranganai Matenda & Mabutho Sibanda. “An Analysis of the Relationship Between Readability of the Chairman’s Statement and Firm Financial Performance in South Africa” (2023) 12:8 International Journal of Research in Business & Social Science 205.

Martínez, Eric, Francis Mollica & Edward Gibson. “Even Laypeople Use Legalese” (2024) 121:35 Proceedings of the National Academy of Sciences e2405564121.

McLachlin, Beverly. “Remarks of the Rt Hon Beverley McLachlin, PC, Chief Justice of Canada” (May 16, 2010), NJI and CIAJ Advanced Judgment Writing Seminar (Perth, Ontario), online: <https://ciaj-icaj.ca/wp-content/uploads/podcasts/2020/11/podcast1b_16may10-nji-ciaj-judgment-writing-course-copy.pdf>.

McLaughlin, G Harry. “SMOG Grading – A New Readability Formula” (1969) 12 Journal of Reading 639.

McMullin, Jess, et al. “Effects of Surrounding Information and Line Length on Text Comprehension from the Web” (2002) 28:1 Canadian Journal of Learning and Technology 1.

McNamara, Danielle & Arthur C Graesser. “Coh-Metrix: An Automated Tool for Theoretical and Applied Natural Language Processing” in Philip M McCarthy & Chutima Boonthum-Denecke, eds, *Applied Natural Language Processing: Identification, Investigation, and Resolution* (Hershey, PA: IGI Global, 2012) 188.

McNamara, Danielle S, et al. *Automated Evaluation of Text and Discourse with Coh-Metrix* (New York, Cambridge UP, 2014)

Meade, Cathy D & Cyrus F Smith. “Readability Formulas: Cautions and Criteria” (1991) 17 Patient Education and Counseling 153.

Medizabal, Victoria. “Courts in Buenos Aires are Using ChatGPT to Draft Rulings,” Rest of World (22 Nov 2024), online: <https://restofworld.org/2024/buenos-aires-courts-adopt-chatgpt-draft-rulings/>.

Mehrpour, Saeed & Abdolmehdi Riazi. “The Impact of Text Length on EFL Students’ Reading Comprehension” (2004) 6:3 Asian EFL Journal.

Mertz, Elizabeth & Jothie Rajah. "Language-and-Law Scholarship: An Interdisciplinary Conversation and a Post-9/11 Example" (2014) 10 Annual Review of Law and Social Science 169.

Miller, David. "Two Ways to Think About Justice" (2002) 1 Politics, Philosophy & Economics 5.

Milne, George R, Mary J Culnan & Henry Greene. "A Longitudinal Assessment of Online Privacy Notice Readability" (2006) 25:2 Journal of Public Policy & Marketing 238.

Mindlin, Maria. "Is Plain Language Better – A Comparative Readability Study of Court Forms" (2006) 10 Scribes Journal of Legal Writing 55.

Mohammadi, Hamid & Seyed Hossein Khasteh. "Text as Environment: A Deep Reinforcement Learning Text Readability Assessment Model" (2019), online: <https://arxiv.org/pdf/1912.05957.pdf>.

Morison, John & Adam Harkens. "Re-engineering Justice? Robot Judges, Computerised Courts and (Semi) Automated Legal Decision-Making" (2019) 39:4 Legal Studies 618

Morris, Craig D & Christine Stilwell. "Getting the Write Message Right: Review of Guidelines for Producing Readable Print Agricultural Information Materials" (2003) 69:1 South African Journal of Libraries and Information Science 71.

Moustafa, Khaled. "Improving PDF Readability of Scientific Papers on Computer Screens" (2016) 35 Behaviour & Information Technology 319.

Nanavati, Anuj A & Randolph G Bias. "Optimal Line Length in Reading – A Literature Review" (2005) 39:2 Visible Language 120.

National Judicial Institute. "Course Descriptions", online: <https://www.nji-inm.ca/index.cfm/publications/public-resources/nji-course-descriptions/>.

New Zealand Law Commission. *Presentation of New Zealand Statute Law* (Wellington: Office of Parliamentary Counsel, 2007), online: <https://www.lawcom.govt.nz/sites/default/files/projectAvailableFormats/NZLC%20IP2.pdf?>.

Ngabut, Maria Novary. "Reading Theories and Reading Comprehension" (2015) 5:1 Journal on English as a Foreign Language 25.

Nguyen, Justin Hung. "Tax Avoidance and Financial Statement Readability" (2021) 30:5 European Accounting Review 1043.

Nussbaum, Martha Craven. *Sex & Social Justice* (Oxford: Oxford UP, 1999).

Office of the Commissioner for Federal Judicial Affairs Canada. "Demographic Statistics on Diversity in the Judiciary" (6 Mar 2024), online: <https://www.fja.gc.ca/appointments-nominations/StatisticsCandidate-StatistiquesCandidat-2023-eng.html>.

---. "Number of Federally Appointed Judges as of May 1, 2024" (May 1, 2024), online: <https://www.fja.gc.ca/appointments-nominations/judges-juges-eng.aspx>.

Okin, Susan Moller. "Justice and Gender: An Unfinished Debate" (2004) 72 Fordham Law Review 1537.

Ombudsman Saskatchewan. *Practice Essentials for Administrative Tribunals* (Saskatchewan: Ministry of Justice and Attorney General, 2020).

Organisation for Economic Cooperation and Development. "Education at a Glance – 2022 Indicators – Canada", online: <https://www.oecd-ilibrary.org/sites/5ea1485d-en/index.html?itemId=/content/component/5ea1485d-en>.

O'Reilly, Tenaha & Danielle S McNamara. "Reversing the Reverse Cohesion Effect: Good Texts Can Be Better for Strategic, High-Knowledge Readers" (2007) 43:2 Discourse Processes 121.

Orwell, George. "Politics and the English Language," (1968), online: <http://www.orwell.ru/library/essays/politics/english/e_polit/>.

Pantula, Muralidhar, & KS Kuppusamy. "A Machine Learning-Based Model to Evaluate Readability and Assess Grade Level for the Web Pages" (2020) 65:4 The Computer Journal 831.

Perfetti, Charles & Joseph Stafura. "Word Knowledge in a Theory of Reading Comprehension" (2014) 18:1 Scientific Studies of Reading 22.

Pishghadam, Reza & Hannaneh Abbasnejad. "Emotioncy: A Potential Measure of Readability" (2016) 9:1 International Electronic Journal of Elementary Education 109.

Pons, Emilia da Silva, et al. "Users' Preferences and Perceptions of the Comprehensibility and Readability of Medication Labels" (2019) 14:2 PLoS ONE e0212173, online: <https://doi.org/10.1371/journal.pone.0212173>.

Popkin, William D. "Law and Linguistics: Is There Common Ground" (1995) 73 Washington University Law Quarterly 1043.

Posner, Richard A. "Judges' Writing Styles (And Do They Matter?)" (1995) 62 University of Chicago Law Review 1420.

Rakedzon, Tzipora & 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.

Rawls, John. *A Theory of Justice* (Cambridge, MA: Harvard UP, 1999).

Raz, Joseph. *The Authority of Law* (Oxford: Oxford UP, 1979).

Re, Richard M. "Artificial Authorship and Judicial Opinions" (2024) 92:4 George Washington Law Review 1558.

Rhode, Deborah L. *Access to Justice* (New York: Oxford UP, 2004).

Robbins, Ruth Anne. "Painting with Print: Incorporating Concepts of Typographic and Layout Design into the Text of Legal Writing Documents" (2004) 2 Journal of the Association of Legal Writing Directors 108.

Rogowsky, Beth A, Barbara M Calhoun & Paula Tallal. "Matching Learning Style to Instructional Method: Effects on Comprehension" (2015) 107:1 Journal of Educational Psychology 64.

Rosenberg, Joseph A. "Why Lawyers Should Write for the Digital Reader" (2023) 27 Journal of the Legal Writing Institute 111.

Russo, Riccardo. *Statistics for the Behavioural Sciences: An Introduction* (New York: Taylor & Francis, 2003).

Sableman, Mark. "Typographic Legibility: Delivering Your Message Effectively" (2017) 17 Scribes Journal of Legal Writing 9.

Sadoski, Mark, Erin M McTigue & Allan Paivio. "A Dual Coding Theoretical Model of Decoding in Reading: Subsuming the Laberge and Samuels Model" (2012) 33:5 Reading Psychology 465.

Salter, Shannon. "Online Dispute Resolution and Justice System Integration: British Columbia's Civil Resolution Tribunal" (2017) 34 Windsor Yearbook of Access to Justice 112.

---. "RDO Speak to Shannon Shalter [SIC] About All Things ODR" (November 2019), online: <https://resolvedisputes.online/blog_shannon_salter.html>.

Salter, Shannon & Darin Thompson. "Public-Centered Civil Justice Redesign: A Case Study of the British Columbia Civil Resolution Tribunal" (2017) 3 McGill Journal of Dispute Resolution 113.

Salyzyn, Amy, et al. "Literacy Requirements of Court Documents: An Under-Explored Barrier to Access to Justice" (2016) 33 Windsor Yearbook of Access to Justice 216.

Salyzyn, Amy & Shannon Stroud. "Direct-to-Public Digital Legal Tools in Canada: A 2023 Update" (June 30, 2024). Ottawa Faculty of Law Working Paper No. 2024-15, online: <https://ssrn.com/abstract=4880902>

Samples, Tim R, Katherine Ireland & Caroline Kraczon. “TL;DR: The Law and Linguistics of Social Platform Terms-of-Use” (2024) 39 Berkeley Technology Law Journal 47.

Satz, Debra. “Amartya Sen’s the Idea of Justice: What Approach, Which Capabilities” (2012) 43 Rutgers Law Journal 277.

Saunders, Jamie WS. “Reflections on the Art and Science of Decision-Making and Decision-Writing” (Nov 2017), Professional Development Conference: Maritime Provinces’ Law Societies (Halifax, NS), online: <https://www.courts.ns.ca/sites/default/files/editor-uploads/From%20the%20Bench/Reflections_on_the_Art_and_Science_of_Decision_Making_11_09_17.pdf>.

Schmidt, Roy, et al. “Identifying Software Project Risks: An International Delphi Study” (2001) 17:4 Journal of Management Information Systems 5.

Schmitz, Benedikt & Charlotte Pavillon. “Measuring Transparency in Consumer Contracts: The Usefulness of Readability Formulas Empirically Assessed” (2020) 9:5 Journal of European Consumer and Market Law 191.

Schubert, Lenhart. “Computational Linguistics” in Edward N Zalta, ed, *Stanford Encyclopedia of Philosophy* (Stanford: Metaphysics Research Lab, 2020).

Schwabe, Annika, et al. “No Negative Effects of Reading on Screen on Comprehension of Narrative Texts Compared to Print: A Meta-analysis” (2022) 25:6 Media Psychology 779.

Segalowitz, Sidney J & Korri C Lane. “Lexical Access of Function versus Content Words” (2000) 75 Brain & Language 376.

Semple, Noel. “Tribunals for Access to Justice in Canada”, 2024 CanLIIDocs 2368, online: <https://canlii.ca/t/7ng33>.

Serry, Tanya, et al. “Improving Access to COVID-19 Information by Ensuring the Readability of Government Websites” (2022) 34:2 Health Promotion Journal of Australia 595.

Sherman, L A. *The Analytics of Literature* (Boston, Ginn & Company, 1893).

Sirico, Louis J Jr. “Readability Studies: How Technocentrism Can Compromise Research and Legal Determinations” (2007) 26:1 Quinnipiac Law Review 147.

Slocum, Brian G. Ed, *The Nature of Legal Interpretation: What Jurists Can Learn about Legal Interpretation from Linguistics and Philosophy* (Chicago: Chicago UP, 2017) 14.

Smeuninx, Nils, Bernard De Clerck & Walter Aerts. “Measuring the Readability of Sustainability Reports: A Corpus-Based Analysis Through Standard Formulae and NLP” (2020) 57:1 International Journal of Business Communication 52.

Smit, Tim, Max van Haastrecht & Marco Spruit. “The Effect of Countermeasure Readability on Security Intentions” (2021) 1 Journal of Cybersecurity & Privacy 675.

Smith, E A & R J Senter. *Automated Readability Index. AMRL-TR-66-22* (Wright-Patterson AFB, OH: Aerospace Medical Division, 1967).

Snee, Ronald D. “Validation of Regression Models: Methods and Examples” (1977) 19:4 Technometrics 415.

Social Security Tribunal of Canada. “An Evaluation of How Easy it is to Read Decisions of the Social Security Tribunal” (last modified 10 May 2022), online: <https://sst-tss.gc.ca/en/our-work-our-people/evaluation-easy-it-read-decisions-social-security-tribunal>.

---. “Evolution of Plain Language Decision Writing” (February 28, 2023), online: <https://www.sst-tss.gc.ca/en/our-work-our-people/evaluation-plain-language-decision-writing>.

---. “Style Guide: Social Security Tribunal of Canada Decisions” (15 May 2024), online: <https://www.sst-tss.gc.ca/en/our-work-our-people/style-guide-social-security-tribunal-canada-decisions>.

Soleimani, Hassan & Elham Mohammadi. “The Effect of Text Typographical Features on Legibility, Comprehension, and Retrieval of EFL Learners” (2012) 5:8 English Language Teaching 207.

Solum, Lawrence B. “The Interpretation-Construction Distinction” (2010) 27 Constitutional Commentary 95.

Sossin, Lorne. “Designing Administrative Justice” (2017) 24:1 Windsor Yearbook of Access to Justice 87.

Sparks, Jessica F & Jay D Hmielowski. “At the Extremes: Assessing Readability, Grade Level, Sentiment, and Tone in US Media Outlets” (2023) 24:1 Journalism Studies 24.

Spencer, Shaun B & Adam Feldman. “Words Count: The Empirical Relationship between Brief Writing and Summary Judgment Success” (2018) 22 Journal of the Legal Writing Institute 61.

Statistics Canada. *Statistics: Power from Data!* (Government of Canada: Ottawa, September 2, 2021), online: <https://www150.statcan.gc.ca/n1/edu/power-pouvoir/toc-tdm/5214718-eng.htm>.

---. “Table 2 – Language spoken at least regularly at home, provinces and territories, 2021” (Aug 17, 2022), online: <https://www150.statcan.gc.ca/n1/daily-quotidien/220817/t002a-eng.htm>.

---. “Table 98-10-0384-01: Highest level of education by census year: Canada, provinces and territories, census metropolitan areas and census agglomerations” (December 9, 2022), online: <https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=9810038401>.

---. “Table 98-10-0385-01 – High school completion by census year: Canada, provinces and territories, census metropolitan areas and census agglomerations” (Oct 4, 2023), online: <https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=9810038501>/

Steckel, Joel H. & Wilfried R Vanhonacker. “Cross-Validating Regression Models in Marketing Research” (1993) 12:2 Marketing Science 415.

Steenhuis, Quinten, Bryce Willey & David Colarusso. “Beyond Readability with RateMyPDF: A Combined Rule-based and Machine Learning Approach to Improving Court Forms,” in *Proceedings of International Conference on Artificial Intelligence and Law (ICAIL 2023)* (Association for Computing Machinery: New York, 2023), online: <https://doi.org/10.1145/3594536.3595146>.

Strauss, David A. “The Plain Language Court” (2016) 38:2 Cardozo Law Review 651.

Sullivan, Barry & Ramon Feldbrin. “The Supreme Court and the People: Communicating Decisions to the Public” (2022) 24:1 University of Pennsylvania Journal of Constitutional Law 1.

Sullivan, Ruth. “The Promise of Plain Language Drafting” (2001) 47:1 McGill Law Journal 97.

Supreme Court of Canada. “Cases in Brief” (last updated 29 Jul 2022), online: <https://www.scc-csc.ca/case-dossier/cb/index-eng.aspx>.

---. “The Honourable Sheilah L. Martin” (15 May 2018), online: <https://www.scc-csc.ca/judges-juges/bio-eng.aspx?id=sheilah-l-martin>.

---. “News Release” (8 Oct 2020), online: <https://decisions.scc-csc.ca/scc-csc/news/en/item/6964/index.do>.

Thomas, Jafrā D, Brian R Flay & Bradley J Cardinal. “Are Physical Activity Resources Understandable as Disseminated? A Meta-Analysis of Readability Studies” (2018) 70:4 Quest 492.

Thorndike, Edward L. *The Teacher’s Word Book* (New York: Columbia University Teachers College, 1921).

Tiersma, Peter Meijes. “A History of the Languages of Law” in Lawrence Solan & Peter Meijes Tiersma, eds, *The Oxford Handbook of Language and Law* (Oxford: Oxford University Press, 2012) 13

Tribunals Ontario. “Guidelines for Reasons Review and Decision Quality Standards” (20 May 2022), online: <https://tribunalsontario.ca/en/practice-directions/>.

---. “Practice Direction on the Use of Artificial Intelligence (AI) in Tribunal Proceedings”, April 2025, online: <https://tribunalsontario.ca/documents/TO/Practice-Direction-on-AI_EN.html>.

Tschopp, Daniel, Doug Barney & Passard C Dean. “The Readability of Financial Statement Note Disclosures” (2018) 12:1 International Journal of Business, Accounting, and Finance 63.

Turnbull, I. “Legislative Drafting in Plain Language and Statements of General Principle” (1997) 18:1 Statute Law Review 21.

Vaillancourt, François & Richard Bird. “Tax Simplification in Canada: A Journey Not Yet Mapped” in Tamer Budak, Simon James & Adrian Sawyer (eds), *The Complexity of Tax Simplification: Experiences from Around the World* (New York: Palgrave Macmillan, 2016), 70.

van Domselaar, Iris. “‘Plain’ Legal Language by Courts: Mere Clarity, An Expression of Civic Friendship or a Masquerade of Violence?” (2022) 10 Theory & Practice of Legislation 93.

Varsava, Nina. “Elements of Judicial Style: A Quantitative Guide to Neil Gorsuch’s Opinion Writing” (2018) 93 New York University Law Review Online 75.

---. “Professional Irresponsibility and Judicial Opinions” (2021) 59:1 Houston Law Review 103.

Vogel, Mabel & Carleton Washburne. “An Objective Method of Determining Grade Placement of Children’s Reading Material” (1928) 28:5 The Elementary School Journal 373.

Vos, Geoffroy. “The Future for Dispute Resolution: Horizon Scanning” (Sir Brian Neill Lecture, The Society of Computers and Law, 17 March 2022), online: <https://www.judiciary.uk/wp-content/uploads/2022/03/MR-to-SCL-Sir-Brain-Neill-Lecture-2022-The-Future-for-Dispute-Resolution-Horizon-Scannings-.pdf>.

Wald, Patricia M. “How I Write” (1993) 4 Scribes Journal of Legal Writing 55.

Weiler, Paul. “Two Models of Judicial Decision-making” (1968) 46 Canadian Bar Review 406.

Whalen, Ryan. “Judicial Gobbledygook: The Readability of Supreme Court Writing” (2016) 125 Yale Law Journal Forum 200.

Wilson, Michael. “MRC Psycholinguistic Database: Machine-Usable Dictionary, Version 2.00” (1988) 20:1 Behavior Research Methods, Instruments, & Computers 6.

Wilson, Theodore. “The Promise of Behavioral Economics for Understanding Decision-Making in the Court” (2019) 18:4 Criminology & Public Policy 785.

Wolf, MC, et al. “The Relationship Between Reading and Listening Comprehension: Shared and Modality-Specific Components” (2019) 32 Reading and Writing 1747.

Wood, Sarah G, et al. “Does Use of Text-to-Speech and Related Read-Aloud Tools Improve Reading Comprehension for Students With Reading Disabilities? A Meta-Analysis” (2018) 51:1 Journal of Learning Disabilities 73.

Wu, Zhiwei. “Understanding Students’ Mimicry, Emulation and Imitation of Genre Exemplars: An Exploratory Study” (2019) 54 English for Specific Purposes 127.

Young, Iris Marion. *Justice and the Politics of Difference* (Princeton: Princeton UP, 2011).

Young, Robert Q. “A Comparison of Reading and Listening Comprehension with Rate of Presentation Controlled” (1973) 21:3 AV Communication Review 327.

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

Zhang, Ting & Josh Hoddenbagh. *The Costs of the Youth Criminal Justice System 2010* (Ottawa: Department of Justice, 2013).

Zheng, Jiaping & Hong Yu. “Readability Formulas and User Perceptions of Electronic Health Records Difficulty: A Corpus Study” (2017) 19:3 Journal of Medical Internet Research.