Validating Readability and Complexity Metrics: A New Dataset of Before-and-After Laws

Wolfgang Alschner

MERCATUS WORKING PAPER

All studies in the Mercatus Working Paper series have followed a rigorous process of academic evaluation, including (except where otherwise noted) at least one double-blind peer review. Working Papers present an author's provisional findings, which, upon further consideration and revision, are likely to be republished in an academic journal. The opinions expressed in Mercatus Working Papers are the authors' and do not represent official positions of the Mercatus Center or George Mason University.

MERCATUS CENTER
George Mason University
3434 Washington Blvd., 4th Floor, Arlington, Virginia 22201
www.mercatus.org

Wolfgang Alschner, "Validating Readability and Complexity Metrics: A New Dataset of Before-and-After Laws," Mercatus Working Paper, Mercatus Center at George Mason University, Arlington, VA, March 2021.

Abstract

If algorithms are to be the policy analysts of the future, the policy metrics they produce will require careful validation. This paper introduces a new dataset that assists in the creation and validation of automated policy metrics. It presents a corpus of laws that have been redrafted to improve readability without changing content. The dataset has a number of use cases. First, it provides a benchmark of how expert legislative drafters render texts more readable. It thereby helps test whether off-the-shelf readability metrics such as Flesch-Kincaid pick up readability improvements in legal texts. It can also spur the development of new readability metrics tailored to the legal domain. Second, the dataset helps train policy metrics that can distinguish policy form from policy substance. A policy text can be complex because it is poorly drafted or because it deals with a complicated substance. Separating form and substance creates more reliable algorithmic descriptors of both.

JEL codes: K2, L5, N4, Y1

Keywords: Readability, Complexity, Flesch-Kincaid, Policy Analytics, Law, Statutes, Plain Language

Author Affiliation and Contact Information

Wolfgang Alschner
Associate Professor, Common Law Section University of Ottawa
Wolfgang.alschner@uottawa.ca

Acknowledgment

This paper is based on a research project on operationalizing plain legal language. The author is grateful to Sophie Gadbois, Giovanni Cesario Giuga, and Daniel D'Alimonte for their research assistance.

This paper can be accessed at https://www.mercatus.org/publications/regulation/validating-readability-and-complexity-metrics-new-dataset-and-after-laws.

Validating Readability and Complexity Metrics:

A New Dataset of Before-and-After Laws

Wolfgang Alschner

The Need for Benchmarking and Validation in Policy Analytics

Poorly written laws make it challenging for individuals to understand their rights and obligations, just as overly complex regulations make it cumbersome for businesses to comply with regulatory requirements. Finding reliable and efficient ways to assess the readability of laws or to measure the complexity of regulations can therefore facilitate reforms that improve access to justice and lighten regulatory burden. The emerging field of policy analytics leverages state-of-the-art computational methods to render such policy analysis scalable—that is, to automatically investigate large amounts of policy texts efficiently and effectively to solve challenges of fundamental importance to governments, citizens, and businesses. By using natural language processing and artificial intelligence, policy analysts can treat policy documents as data and use data science tools to mine them. The ability to efficiently investigate thousands of laws, regulations, or other policy texts can, in turn, help inform, improve, and accelerate evidence-based legal and regulatory reform.[1]

For algorithms to become the policy analysts of tomorrow, policy analytics needs to develop scalable metrics that meet two criteria. First, policy metrics must reliably capture policy-relevant attributes of policy texts. If algorithms can generate insights that manual analysis would reveal as well but faster and cheaper, then the case for the use of automated policy analytics becomes

[1] Omar Al-Ubaydli and Patrick A. McLaughlin, "RegData: A Numerical Database on Industry-Specific Regulations for All United States Industries and Federal Regulations, 1997–2012," *Regulation & Governance* 11, no. 1 (2017): 109–23.

highly persuasive. Ensuring accuracy and reliability of policy metrics requires validation, which in social and computer sciences typically means rigorously testing automated results against how a human would have performed the same task. Second, policy metrics must be useful in guiding policy reform. That means metrics must be explainable and detailed enough to inspire policy action. In both instances, computer-generated policy metrics need to be assessed against human-generated policy analysis.

This paper introduces and showcases a before-and-after dataset of laws from five common law jurisdictions that have been revised by plain language drafting experts so as to enhance their accessibility and readability. The dataset supports the design, benchmarking, and validation of automated policy metrics in two ways. First, it establishes a benchmark on how human experts make legislative statutes more readable. This helps to validate automated metrics that seek to assess the readability of texts and allows the creation of new automated measures on the basis of an inductive comparison of before-and-after texts. Second, the dataset allows for developing metrics that distinguish stylistic changes from policy changes. That matters, for example, when analysts evaluate the complexity of policy texts. They may ask, is a regulation concerning nuclear power plants complex because it is drafted in a wordy manner or because the underlying policy area necessitates a more detailed regulatory approach? Because the dataset comprises the same laws in an original and a plain language revision without changes to the underlying substance, it becomes possible to distinguish between readability (that changes) and policy complexity (that remains constant).

Promises and Limitations of Existing Policy Metrics

In an ideal world, an abundance of policy texts carefully annotated by expert analysts would allow scaling policy analysis through supervised machine learning, a form of artificial intelligence whereby algorithms learn relationships from manually classified data in order to recognize identical patterns in new documents.[2] But because machine learning models need to be trained anew for each task and require large amounts of training data, their employment in policy analysis remains in its infancy.[3] Simpler but equally scalable rules-based metrics that count specific textual features to describe attributes of policy texts are a promising second-best solution.

The perhaps best-known such metric is the Flesch-Kincaid readability score that assesses how readable a text is—that is, how easy it is to understand. Flesch-Kincaid calculates readability based on syllables-per-word counts and words-per-sentence counts. Since Flesch-Kincaid scores can be calculated easily once text is available in digital format and is not domain specific, it is widely used to benchmark policy documents. For example, North Carolina, Florida, and Oregon have recently enacted legislation that requires government documents to meet a defined Flesch-Kincaid readability threshold.[4]

Scholars have proposed similar rules-based metrics to describe other dimensions of policy texts. Katz and Bommarito, for instance, proposed word-based Shannon entropy scores to measure the United States Code's complexity.[5] Shannon entropy is a metric developed in information theory to describe the information content of a signal. The concept has since been applied to new domains—for example, to assess the quality of literary translations or to compare

[2] Justin Grimmer and Brandon M. Stewart, "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Text," *Political Analysis* 21, no. 3 (2013): 267–97.

[3] For a small-scale pilot study, see Michael Curtotti et al., "Machine Learning for Readability of Legislative Sentences," in *ICAIL'15: Proceedings of the 15th International Conference on Artificial Intelligence and Law* (New York: Association for Computing Machinery, 2015), 53–62.

[4] NC Gen Stat § 58-66-30(b) (2013); FL Stat § 627.4145 (2019); OR Rev Stat § 316.364 (2019).

[5] Daniel M. Katz and M. J. Bommarito II, "Measuring the Complexity of the Law: the United States Code," *Artificial Intelligence and Law* 22 (2014): 337–74.

the complexity of natural language to computer code.[6] Shannon entropy helps to quantify the variance of words in a text and can serve as a proxy for complexity under the assumption that texts with more diverse words are more complex than texts with more homogenous terms. It is increasingly used to assess the complexity of legal and policy documents.[7]

What makes these generic measures so attractive is that they are easy to quantify and can be applied across domains. However, they also come with important limitations—especially when applied to policy texts. First, they are only rough approximations of the text characteristics they seek to measure. Consider, for example, legal texts that extensively use Latin legalese. Plain language drafting experts universally agree that such legalese makes legal texts less readable and more complex, but Latin legalese may produce shorter words and sentences, and thus result in better Flesch-Kincaid scores. Similarly, if used consistently, frequent legalese will not significantly affect Shannon entropy scores. Second, it is not always clear what policy dimension rules-based metrics are evaluating. For example, verbose drafting and a more complicated subject matter can both increase word variance resulting in higher entropy scores. It is then unclear how a document's complexity can be reduced: should the text be redrafted or does the underlying policy field require reform?

These shortcomings of generic policy metrics risk making results confusing, misleading, and ambiguous, and thus ultimately unhelpful in guiding meaningful policy reform. To ensure the best of both worlds—simple implementation and meaningful guidance for policy reform—

[6] Gerardo Febres, Klaus Jaffé, and Carlos Gershenson, "Complexity Measurement of Natural and Artificial Languages," *Complexity* 20 (2015): 25–48; Marcin Lawnik, "Shannon's Entropy in Literary Works and Their Translations," *Journal of Computer Science* 1, no. 3 (2012).

[7] Patrick A. McLaughlin, Oliver Sherouse, Mark Febrizio, and M. Scott King, "Is Dodd-Frank the Biggest Law Ever?" (Mercatus Working Paper, Mercatus Center at George Mason University, Arlington, VA, June 2020).

greater attention needs to be paid to validating, benchmarking, and fine-tuning rules-based policy metrics. To accomplish that, a benchmark of human-generated policy reform is needed.

A Plain Language Gold Standard: A New Dataset of Before-and-After Legislation

For decades, proponents of the "Plain Legal Language Movement" have sought to render legal texts more accessible and readable for nonlawyers.[8] The movement has not only helped to produce official plain language guidelines, such as the Canadian *Legistics* guidebook used in federal legislative drafting or the United States' 2010 Plain Writing Act, but has also inspired plain-language rewrites of existing statutes across several common law jurisdictions. These rewritten texts are a unique benchmark against which to test policy metrics. They embody how human experts have restyled policy texts to make them more readable without changing the substance of the original law. Put differently, these plain language revisions hold the policy substance constant and only change the language in which that substance is being communicated.

My research assistants and I have engaged in a comprehensive effort to identify instances of such plain language rewrites and have succeeded in compiling a dataset of originally enacted legislation ("before") and their plain-language rewritten versions ("after") in five Anglo-American jurisdictions (table 1). Three of the rewrites in the dataset are drafted by academics but not enacted (Equality Act,[9] Takeover Codes,[10] Timeshares Act[11]); the others are officially

[8] Mark Adler, "The Plain Language Movement," in *The Oxford Handbook of Language and Law*, ed. Lawrence M. Solan and Peter M. Tiersma (Oxford, UK: Oxford University Press, 2012).

[9] Annelize Nienaber. "Search for Clarity in South Africa's New Equality Legislation" *Clarity* 46, no. 11 (2001).

[10] Martin Cutts, "Clearer Timeshare Act 1993" (1993), https://irp-cdn.multiscreensite.com/aaf9e928/files/uploaded/LucidLawClearerTimeshar.pdf.

[11] Law Reform Commission of Victoria, *Plain English and the Law,* Report No. 9, June 30, 1987, Melbourne, Victoria.

enacted texts to replace the "before" legislation (Minneapolis City Charter,[12] New Zealand's Contract and Commercial Law Bill[13]).

We preprocessed the before-and-after texts to facilitate their analysis. First, we digitized each of the documents from pdf. Second, we converted each document into a structured text format (XML). The three longer rewrites included changes in the document structure and were accompanied by equivalence tables that matched original sections with rewritten sections. We used the provided equivalence tables to align sections on the same subject matter in the XML. This allows us and other researchers to compare individual sections on the same issues in addition to a comparison of the full texts.

As noted in table 1, the rewritten documents are consistently shorter than the original documents. This is not surprising given that plain language guidelines go beyond a substitution of legalese with natural language equivalents and provide for a range of techniques to simplify and consolidate texts. The particularly large decrease in length of the Minneapolis City Charter is additionally due to the fact that some sections of the original Charter were outsourced to separate ordinances and not included in the revised Charter. All other revisions, however, fully preserve the scope of the original texts.

In general, all the rewrites seek to leave the substance of the original laws unchanged and merely aim to make the statutory text more accessible. In relation to the New Zealand Contract and Commercial Law Bill, for example, the bill's commentary clarifies that the purpose of the revision "is to re-enact laws in a modern, accessible format with-out changing the substance of

[12] See Minneapolis Charter Commission, "Side-by-Side Comparison: Source Provisions to Successor Provisions," May 2013; Minneapolis Charter Commission, "Plain Language Charter Revision" (on file with the authors).
[13] Contract and Commercial Law Bill 2016 (134–2), New Zealand, https://www.legislation.govt.nz/bill/government/2016/0134/latest/d56e2.html.

the existing law."[14] Similarly, the materials accompanying the revision of the Minneapolis City Charter explain the rationale of the rewrite as follows: "The revision simplifies the Charter, redrafts it for clarity, removes inconsistencies and organizes it in a logical way. At the same time, the new Charter preserves the way Minneapolis is governed."[15] In short, the rewritten laws included in the dataset provide a unique benchmark of plain language redrafting unaffected by parallel substantive changes.

Table 1: Before-and-After Dataset

Legislation	Jurisdiction	Word Count (Before)	Word Count (After)
Promotion of Equality and Prevention of Unfair Discrimination Act ("Equality Act") Section 12	South Africa	73	40
Timeshares Act	United Kingdom	3,531	2,600
Contract and Commercial Law Bill	New Zealand	35,066	33,523
Minneapolis City Charter	United States	65,554	12,865
Takeover Codes	Australia	31,635	13,764

Source: Author's tabulation.

Using the Dataset to Validate Existing Readability and Complexity Metrics

By epitomizing how human drafters free from other policy considerations have rewritten legal texts to make them more accessible to nonlawyers, the laws of the dataset can represent a new gold standard against which to validate regularly used policy metrics of readability and complexity. We calculated the Flesch-Kincaid scores as a proxy for readability and Shannon

[14] Contract and Commercial Law Bill 2016.

[15] Minneapolis Charter Commission, "Proposed Minneapolis Charter Amendments, November 2013, Frequently Asked Questions," http://www2.ci.minneapolis.mn.us/www/groups/public/@clerk/documents/webcontent/wcms1p-115129.pdf. A similar text was provided to the electorate during a referendum on the adoption of the Charter (https://vote.minneapolismn.gov/results-data/election-results/2013/ballot-questions/). As noted, however, some elements of the original texts were outsourced to ordinances rather than included in the revised Charter.

word-based entropy as a proxy for complexity for each document in our corpus. We minimally preprocessed our texts by streamlining punctuation, because Flesch-Kincaid necessitates clear detection of sentence boundaries and can be confused by the odd punctuation conventions in legal texts (such as frequent use of the semicolon in lieu of full stops or cross-references with punctuation marks).

Table 2 displays the results of these popular policy metrics applied to our before-and-after data. The Flesch-Kincaid measures behave as expected: all rewritten documents have a substantially lower score, which indicates they are easier to read on the basis of that metric. In spite of its simplicity, the Flesch-Kincaid score, based on syllable-per-word and word-per-sentence counts, captures the readability difference between the “before” and “after” versions. This validates the use of Flesch-Kincaid scores as proxy for the readability of texts.

Table 2: Flesch-Kincaid and Shannon Entropy Metrics

Statue	Original FK Score	Plain L. FK Score	FK Diff.	Original Entropy Score	Plain L. Entropy Score	Entropy Diff.
Minneapolis City Charter	24.91	12.60	−12.31	0.81	0.85	.04
New Zealand’s Commercial Bill	29.21	19.80	−9.41	0.83	0.82	−.01
South Africa’s Equality Act S. 12	37.77	11.74	−22.03	0.98	0.97	−.01
Australia’s Takeovers Code	47.33	23.53	−23.8	0.81	0.84	.03
Timeshares Act	23.32	15.68	−7.74	0.87	0.88	.01

Source: Author’s tabulation.

Note: FK = Flesh-Kincaid; L = language.

The entropy results as proxy for complexity are more varied. The change in entropy is marginal; at times positive and at times negative. Entropy thus does not consistently register changes introduced by human drafters to make texts more accessible. This could mean a number of things. Entropy—that is the variance or predictability of words—could be independent from the stylistic complexity of text (which changed in the data) and could rather be linked to the complexity of the policy domain (which was the same in both the "before" and "after" texts). Or entropy could be a poor proxy for measuring complexity in legal texts altogether. In either case, our findings suggest that complexity as proxied by entropy remains unaffected by plain language rewrites and that more research is therefore required to validate entropy as a reliable measure for complexity.

Using the Dataset to Create New Policy-Oriented Metrics

Our dataset can also be used to generate and validate new policy metrics. Whereas Flesch-Kincaid scores, for example, are easily quantifiable and seem to correlate with readability of legal texts, they provide little guidance to drafters apart from reducing the number of syllabi per word or the words per sentence. In contrast, the Plain Language Movement has developed a detailed set of guidelines on how to render legal texts more readable and accessible. Our before-and-after dataset allows the creation and testing of new metrics that are inspired by these guidelines and tailored to legal drafting realities. For example, plain language principles routinely require drafters to avoid unnecessary legalese and to substitute legalistic terms with their ordinary-use equivalent—for example, replacing "shall" with "must."[16]

We thus test how well these formal recommendations are reflected in our rewrites. To operationalize the detection of legalese, we used *Black's Law Dictionary*[17] to create a subset of terms that are considered legalese. Specifically, we considered each term in *Black's Law*

[16] Peter Butt and Richard Castle, *Modern Legal Drafting: A Guide to Using Clearer Language*, 2nd ed. (Melbourne, Australia: Cambridge University Press, 2006), 170.

[17] Bryan A. Garner, *Black's Law Dictionary,* 11th edition (Toronto: Thomson Reuters, 2019).

Dictionary that was not present in the Hunspell open-source English dictionary as legalese. Our list of legalese included primarily Latin terms and unusual English terms such as “offeree.” We then looked for the occurrence of such legalese as well as counts of “shall” versus “must” in our dataset.

The results of our analysis are displayed in table 3. Consistent with plain language guidelines, we find that rewritten texts omit the word “shall,” use “must” more frequently, and employ much less legalese. Our metrics thus successfully track specific revisions that plain language drafters implement in their rewrites. These metrics could now be scaled and applied to other datasets to evaluate the use of legalese in other policy texts.

The example underscores the use of the before-and-after dataset to develop, pilot, and validate tailored metrics to quantify readability-relevant document attributes before they are rolled out. The before-and-after dataset thereby remedies an important gap in the literature. While scholars continue to develop and propose new fine-tuned metrics for assessing the readability and complexity of documents, these metrics are typically applied directly to new datasets.[18] Adding a new validation step will help test such metrics and render them comparable before they are applied elsewhere.

Table 3: Legalese and Shall/Must Counts

	Minneapolis City Charter		New Zealand’s Commercial Bill		South Africa’s Equality Act S. 12		Australia’s Takeover Codes		UK’s Timeshares Act	
	Old	PL	Old	PL	Old	PL	Old	PL	Old	PL
Shall	1,763	0	283	0	0	0	155	0	32	0
Must	8	121	25	124	0	0	0	106	7	14

[18] Bernhard Waltl and F. Matthes, “Towards Measures of Complexity: Applying Structural and Linguistic Metrics to German Laws,” in *Legal Knowledge and Information Systems: JURIX 2014: The Twenty-Seventh Annual Conference,* ed. Rinke Hoekstre (Amsterdam: IOS Press BV, 2014), 153–62.

Legalese	186	46	159	125	1	0	317	216	50	3

Source: Author’s tabulation.

Note: PL = plain language.

Uses of the Dataset to Validate Substantive Rather Than Stylistic Policy Metrics

Finally, the dataset can help assess policy metrics that track substantive rather than stylistic characteristics of policy texts. The RegData project developed by Al-Ubaydli and McLaughlin, for example, proxies the substantive restrictiveness of a regulation by counting the occurrence of constraining signaling words such as “shall,” “must,” “may not,” “prohibited,” and “required” in a policy text.[19] Since, as noted above, the plain language rewrites in our dataset changed the form but not substance of the underlying law, it follows that metrics designed to track substantive policy text attributes, such as the restrictiveness of a legal text, should register few if any changes between the “before” and the “after” texts.

However, as table 3 showed, the plain language drafts contain fewer aggregated mentions of “must” and “shall” than the original texts. Although there are instances where “must” replaces “shall” (the New Zealand Commercial Bill, for example, changed “The District Court *shall* not approve a contract . . .” to “The District Court *must* not approve a contract . . .”), the rewrites often simply omit “shall,” as in the example of table 4. RegData’s restrictiveness metric would thus suggest a reduction in restrictions between before-and-after texts although the reworded text has remained substantively unchanged.

[19] Omar Al-Ubaydli and Patrick A. McLaughlin, “Regdata: A Numerical Database on Industry-Specific Regulations for All United States Industries and Federal Regulations, 1997–2012*,” Regulation & Governance* 11, no. 1 (2017): 109–23.

Table 4: Plain Language Legal Rewrites Often Omit Rather Than Replace Instances of Legalese Like "Shall"

Original Timeshare Act Article 5(4)	Plain Language Timeshare Act Article 3(4)
The offeree's giving, within the time allowed under this section, notice of cancellation of the agreement to the offeror at a time when the agreement has been entered into *shall* have the effect of cancelling the agreement.	An agreement is cancelled if the customer gives the seller notice of cancellation within the time this section allows.

Our findings therefore also underscore the limits of an approach based on signaling words to distinguish form and substance. Legal drafters can frame legal constraints and commands in a variety of textual guises. Policy metrics that track regulatory restrictiveness through legalistic signaling terms alone risk to overestimate and underestimate the actual restrictiveness of a regulation, respectively, by capturing signaling words that, in fact, do not embody constraints and by failing to spot constraints that are wrapped in a different textual guise. Using the dataset of before-and-after laws as benchmark and validation tool, the next-generation policy metrics could seek to differentiate form and substance more clearly in order to accurately track the substantive attributes of policy documents.

Conclusion

The before-and-after dataset of plain language rewrites presented showcases how existing policy analytics can be validated and new metrics can be developed. By functioning as a gold standard, the dataset can benchmark policy metrics before they are rolled out on new corpora and will allow the design of new metrics that capture what human drafters actually do when they render texts more readable. Furthermore, the plain language rewrites help differentiating between substantive policy reform and stylistic redrafting, and thus promise to disambiguate policy metrics and provide more targeted guidance for policy reform.