Readability of Water Reports in the U.S.

Disclosing information biases through Flesch readability ease scores

Renee Ana Aziz

Abstract

The average reader has an 8th-grade reading ability, and all papers written for the public are at or below this level. Unfortunately, water quality reports, which reflect the water toxicity levels of counties across the U.S., can be unreadable by the public; sometimes, they reflect scores at the university level. Every individual in the U.S. has the right to know their water quality and the health risks of dangerous heavy metal levels.

This paper analyzes more than 1,000 water reports in the U.S. using readability metrics. With Python, each PDF was parsed and text-extracted using optical character recognition. A dataframe displayed each water report's readability through multiple readability indices. All processes disclosed a grave information bias that needs resolution.

Why the Public Should Care

- Arsenic can cause various types of cancer within the body
- Hydraulic fracturing chemicals can injure the immune and reproductive systems
- Pesticides can generate neurological damage and Parkinson's disease

Why Water Utilities Should Care

- The health of their consumers should be a top priority
- Reports should help consumers know the health risks
- Can prevent water contamination lawsuits by clearly informing the public

Why State Governments Should Care

- The health of their citizens should be a top priority
- Should implement policies to encourage more readable reports
- Can send out PSAs to raise awareness about water contamination

Why Nonprofits Should Care

- Should rally and encourage utilities to use literacy writers
- Can raise money for campaigns for more readable water reports
- Find a way for the public to have their voices heard regarding water contamination

Introduction

Water Contamination

Water reports inform the public about hazardous water contaminants. Lead is a severe threat, especially in tap water. Most individuals exposed to lead do not show symptoms, and any lead exposure can accumulate in the human body over time. Lead in the blood can cause serious cardiovascular and kidney issues. For children, this can mean permanent developmental issues.

Consumer Confidence Reports

The best way for all households and institutions to know whether their water is at risk is via a Consumer Confidence Report (CCR), also known as a water quality report. These reports inform the public about the level of contaminants in tap water. However, these reports are only effective if the public can understand CCRs.

Tech and Readability

Analyzing a CCR is key to determining its effectiveness. The Flesch-Kincaid readability tests show how difficult a passage is to read. The two formulas are the Flesch reading ease and the Flesch-Kincaid grade level. This report harnesses the Flesch reading ease to analyze information biases. The Flesch reading ease formula uses the number of total syllables, words, and sentences along with coefficients to report a score.

The high-level programming language Python can allow us to use the library textstat to generate readability metrics without the need to write lengthy code. Textstat retrieves the sentence and word length and inputs them into formulas to return the complexity and readability of texts.

Difficulties

The tools used to extract the text from the CCRs in PDF form are not always accurate. PyPDF2 is another Python library that extracts text from PDFs; however, it can be highly inaccurate because it neglects spaces between words and specific characters. It also can add random characters to text. Pytesseract, a Python optical character recognition (OCR) library, proved more accurate than PyPDF2 in text extraction.

Counties should be using these tools to determine how readable their reports are. Health literacy writers should be hired to use these tools and rewrite CCRs to make them readable at an 8th-grade level or lower. Water contaminants are incredibly detrimental to every individual's health, and it is essential to understand what is in one's water.

Method

Data Collection and the Testbed

The project began with the collection of water reports. Each water report was placed in a folder according to the state name, and the PDF link was inputted into an Excel sheet. Data collection took the majority of the time.

During the data collection process, a testbed allowed for an understanding of PDF parsing and applying textstat. Jupyter Notebook, an open-source web application to share live code, was used for the testbed. The testbed consisted of small directories and subdirectories with example PDFs to practice text extraction before the actual project involving around 1000 PDFs.

Run 1

After searching and gathering the CCRs in folders, Jupyter Notebook was the central hub for the project. The actual coding process began with using PyPDF2 before OCR. Readability metrics were applied to each CCR and then organized in a dataframe. These include the Flesch reading ease and the Flesch-Kincaid grade level. Other metrics that were applied include the SMOG index, Coleman Liau index, Automated Readability index, Dale Chall readability grade, Linsear write formula, Gunning fog index, and the number of difficult words. This report's findings will only cover the Flesch reading ease, but each formula or index is unique.

The initial database reflected inaccurate scores. The Flesch reading ease only returns values from 0 to 100: the higher the number, the more readable. More than half of the scores were a negative integer or greater than 100. Those scores did not make sense, and after printing the texts the code had extracted, it became apparent that PyPDF2 was not performing properly. The extracted text sometimes contained strings of characters with no spaces or random characters.

Run 2

To run OCR, optical character recognition, the code was rewritten to use Pytesseract, another Python library. After implementing OCR instead of PyPDF2, the dataframe showed mostly reasonable results and very few errors.

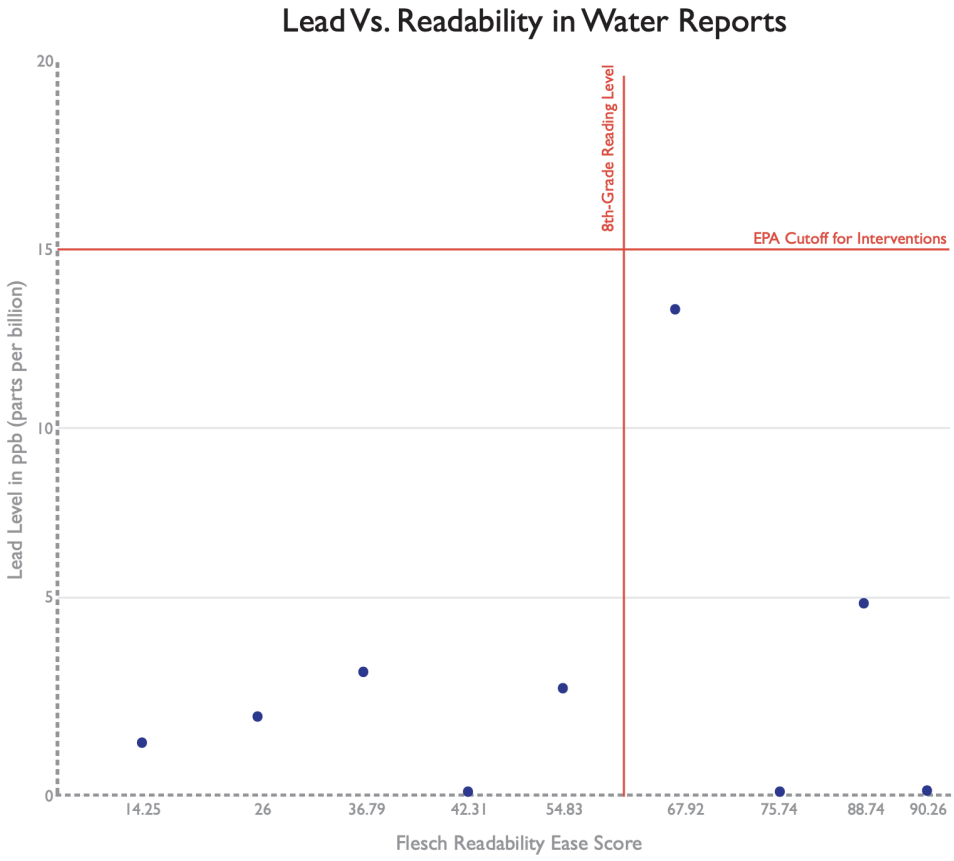
OCR with Pytesseract is more accurate than PyPDF2 because it returned scores that made more sense. On occasion, however, OCR can misinterpret the text or extract a series of

jumbled characters. One way to increase accuracy is to interpret scores that seem inaccurate, print the extracted text, and compare it with the PDF.
Moreover, the Flesch reading ease may not be entirely sound since the formula only considers long words and sentences when not all of them are necessarily hard to read. For example, "beginnings" has the same number of syllables as "irately," which is a more complex word.

Findings

The main takeaway was that sophisticatedly written CCRs create decision barriers. It is reasonable to anticipate all water reports to be relatively readable, but this assertion is incorrect.

Below is a graph of 9 water utility reports spread across the U.S. with their lead levels versus their readability ease scores. These examples were purely selected for illustrative purposes. The blue dots represent the reports, and the orange line shows the EPA's cutoff for lead levels before intervening.



All but three reports from the sample have a lead level greater than 0ppb (parts per billion). Any lead level greater than 0ppb can seriously impact one's health over time if consumed regularly.

An 8th-grade reading level is 60-70 on the Flesch Readability Ease scale, and the four points to the right of the 8th-grade reading level orange line are at and below 8th grade. All points to the left of the vertical orange line are at a high school reading level or above. This is concerning since all but one of these points has a lead level above 0ppb. Even more concerning is that most CCRs are written at the 11th-14th grade level.

When a water report is not easily readable and boasts lead contamination, those who do not understand the CCR are clueless about the health risks of their water. For the CCR to the left of the orange line with a score of 42.31, that person may not even need to pay for water filtration, but because their CCR is not easily legible, they may not know that.

CCRs that are not easily understandable but have a low Flesch readability ease score create an information bias. Those who cannot read at an 8th-grade reading level are at risk for more health conditions than those who can.

Practical Insights and Applications

Utilities should hire literacy writers to make their CCRs more readable. Not only is it a right for the public to know how contaminated their water is, but it is also a danger to the public not to know if their water contains a concerning lead level. Here are some actions that can solve this issue:

1. State governments should require every water utility to hire literacy writers.
2. State governments should ensure that each CCR has a readability score of 60 or higher (8th-grade reading level).
3. Water utilities should invite a few public volunteers to read the water reports and report the lead/copper levels to see if the public has an accurate understanding of contaminant levels.
4. If the CCRs are above an 8th-grade level, the literacy writers should shorten sentences and word lengths within the report. They should also avoid sophisticated language.
5. Each water utility should offer a place, in-person or virtual, for the public to complain about water report readability.
6. General healthcare professionals should inform their patients and the public about the dangers of contaminated water.
7. Physicians should educate their patients on the symptoms of lead/copper poisoning.

8. There should be more influencers/organizations raising awareness about CRR readability and the dangers of water contamination.
9. Environmental organizations and nonprofits should advocate for the negative effects of bad water quality on plants and animals.

These are just a few actions that can advocate for more readable CCRs and better water quality.

Conclusion and Next Steps

The readability scores of CCRs vary, and hard-to-read reports and high lead levels create a grave information bias. Due to the dangers of water contaminants on the public's health, CCRs should be at the 8th-grade reading level or higher.

The Flesch reading ease is a great formula to determine the readability of water reports, but some questions go unanswered. How much do sentence and word length affect the readability of the reports, and how can we make text extraction from PDFs more accurate?

Answering these questions will take more theorizing and statistical analysis but resolving them is imperative. Utilities should consider sending out a readability quiz for a sample of the public to complete. Alternatively, each utility should ensure literacy writers run their text through readability indices before publishing. Whatever the solution, the public, nonprofits, state governments, and water utilities should join forces to ensure readable CCRs.

Works Cited

Bansal, Shivam, and Chaitanya Aggarwal. "Textstat." *PyPI*, Python Software Foundation, 2022, https://pypi.org/project/textstat/.

"Center for Writing Excellence." *Center For Writing Excellence - Montclair State University*, Montclair State University, 2022, https://www.montclair.edu/center-for-writing-excellence/.

Fenniak, Mathieu. "PYPDF2." *PyPI*, Python Software Foundation, 2022, https://pypi.org/project/PyPDF2/.

Flesch, Rudolf. "How to Write Plain English." *Guide to Academic Writing Article - Management - University of Canterbury - New Zealand*, The University of Canterbury, https://web.archive.org/web/20160712094308/http://www.mang.canterbury.ac.nz/writing_guide/writing/flesch.shtml.

Hoffstaetter, Samuel. "Pytesseract." *PyPI*, Python Software Foundation, 2022, https://pypi.org/project/pytesseract/.

"Lead in Drinking Water." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 16 Aug. 2022, https://www.cdc.gov/nceh/lead/prevention/sources/water.htm.

Roy, Siddartha, et al. "An Evaluation of the Readability of Drinking Water Quality Reports: A National Assessment." *Journal of Water and Health*, U.S. National Library of Medicine, Sept. 2015, https://pubmed.ncbi.nlm.nih.gov/26322750/.

"Tip 6. Be Cautious about Using Readability Formulas." *AHRQ*, Agency for Healthcare Research and Quality, May 2015, https://www.ahrq.gov/talkingquality/resources/writing/tip6.html.