Comparing and Developing Tools to Measure the Readability of Domain-Specific Texts

Elissa M. Redmiles[1], Lisa Maszkiewicz[1], Emily Hwang[1], Dhruv Kuchhal[2], Everest Liu[1], Miraida Morales[3], Denis Peskov[1], Sudha Rao[1], Rock Stevens[1], Kristina Gligorić[4], Sean Kross[5], Michelle L. Mazurek[1], and **Hal Daumé III[1,6]**
[1]University of Maryland {eredmiles,mmazurek,hal}@cs.umd.edu
[2]Georgia Tech [3]Rutgers University [4]EPFL [5]UCSD [6]Microsoft Research

Abstract

The readability of a digital text can influence people's ability to learn new things about a range topics from digital resources (e.g., Wikipedia, WebMD). Readability also impacts search rankings, and is used to evaluate the performance of NLP systems. Despite this, we lack a thorough understanding of how to validly measure readability at scale, especially for domain-specific texts.

In this work, we present a comparison of the validity of well-known readability measures and introduce a novel approach, *Smart Cloze*, which is designed to address shortcomings of existing measures. We compare these approaches across four different corpora: crowdworker-generated stories, Wikipedia articles, security and privacy advice, and health information. On these corpora, we evaluate the convergent and content validity of each measure, and detail tradeoffs in score precision, domain-specificity, and participant burden. These results provide a foundation for more accurate readability measurements and better evaluation of new natural-language-processing systems and tools.

1 Introduction

Readability metrics are used in a variety of computational contexts, including to evaluate the quality of novel natural language processing (NLP) systems (Sugawara et al., 2017; Kandula et al., 2010) or machine generated translations (House, 2014), or to rank search results (Slegg, 2018). A variety of readability metrics are available for assessing comprehensibility of texts: human-expert-written comprehension questions, automatically generated readability tests, and computed metrics requiring no human/agent input (Bormuth, 1968; Flesch, 1948; Taylor, 1953; Graesser et al., 2004).

Despite being used frequently in computational contexts, the majority of readability assessments were developed for grade-school texts and validated with grade-school readers. Online texts are generally targeted toward adult readers, leading to differences in text structure (e.g., bullet points), word abstraction (Graesser et al., 2004), and domain-specificity (e.g., medical advice, digital-security advice). Such differences may affect the accuracy of computed metrics and automatically generated readability tests, which are increasingly used to scale readability measurements in the digital world (Friedman and Hoffman-Goetz, 2006; Eysenbach et al., 2002; Bernstam et al., 2005). Despite this use, the validity of readability assessment techniques has rarely been re-evaluated for online contexts.

In this work, we make three contributions: **First**, we evaluate the most commonly used methods for measuring readability in terms of *content validity* (the degree to which different measures relate to theoretically-grounded linguistic components like text cohesion or syntactic complexity), *convergent validity* (the degree to which these measures correspond to each other), *redundancy* (the degree to which one measure is subsumed by another), and *score precision* (the shape of the distribution of score from a given measure, and how well it distinguishes among documents). **Second**, we identify a need for domain-specific automatically generated readability tests. To address this, we develop and evaluate a novel technique for automatically generating readability tests specifically for domain-specific texts: *Smart Cloze*. We find that Smart Cloze offers some benefits for domain-specific applications compared to existing measures. **Third**, we contribute two open-science resources: our open-source Smart Cloze tool, as well as a *Digital Readability* evaluation corpus of 100 documents, including 300 comprehension questions written by human experts, that we use in our evaluation.

2 Related Work

Human-written comprehension questions are the gold standard for measuring readability (Duke and Pearson, 2009; Sarroub and Pearson, 1998), but developing such questions is costly and difficult to scale. As such, prior work has explored various automated, scalable approaches to generating comprehension questions. One such approach is automatic reading test generation, typically using the Cloze (Taylor, 1953) procedure, which involves removing every *n*th word in a given document and requiring the reader to "fill-in-the-blank" with the correct word. The Cloze procedure was validated as a scalable method of comprehension assessment through comparison with expert-written comprehension questions for grade-school texts (Bormuth, 1967; Heilman, 2011; Rankin and Culhane, 1969; Oller et al., 1972).

Recently, researchers have explored approaches to adjusting the construction of Cloze tests: selecting particular key sentences or parts of speech to use as blanks, often to assess retention of factual knowledge or awareness of vocabulary (Goto et al., 2010; Chen et al., 2006; Gates, 2011; Lin and Ji, 2010; Lee and Seneff, 2007), and multiple-choice Cloze tests in which test-takers select from a set of distractors rather than filling in an open blank, avoiding potential scoring issues with typos and equally-correct synonyms (Goto et al., 2010; Narendra et al., 2013; Brown et al., 2005; Mostow and Jang, 2012; Gates, 2011; Pino et al., 2008; Hoshino and Nakagawa, 2007). Our Smart Cloze tool builds on this prior work by choosing distractors from a domain-specific rather than a general dictionary, answering the call from Collins-Thompson's 2014 review of readability measures for more domain-specific tool options.

The second scalable alternative consists of readability metrics that take no reader input. The first such metrics were readability formulae, the most popular of which is the Flesch reading ease score (FRES) (Tekfi, 1987; Flesch, 1948, 1943). FRES assumes that longer sentences and words — which often co-occur with complex syntax — indicate greater reading difficulty (Dale and Tyler, 1934; Feng et al., 2009). More recently, linguistic feature-based (McNamara et al., 2014; François and Miltsakaki, 2012; Collins-Thompson and Callan, 2004) and machine learning approaches have also been used to predict the readability of text (Kate et al., 2010; De Clercq and Hoste, 2016; Collins-Thompson, 2014). Finally, there has also been recent work that goes beyond readability to assess the overall *quality* of text, including factors such as topical interest, persuasiveness, or grammatical correctness (Louis and Nenkova, 2013; Pitler et al., 2010; Tan et al., 2016). In this work, we focus strictly on readability and exclude other quality measures from our comparative evaluation.

All but one of the new approaches to measuring readability (modified Cloze and linguistic metric/ML approaches) were evaluated only through comparison with annotator judgements of perceived readability or correlation with FRES-style formulae, rather than the gold standard of human comprehension questions (Benjamin, 2012), the exception being Mostow et al. (2004). Further, these evaluations were conducted strictly with grade-school texts, leaving a significant gap around online, adult texts (Benjamin, 2012). Our work fills this gap in evaluation (Benjamin, 2012; Collins-Thompson, 2014) by systematically evaluating the validity of currently used readability metrics through comparison with each other and with results from expert written comprehension tests on a wide set of both general and domain-specific documents.

3 Methods

In our evaluation, we compare readability scores from five sources: human-written comprehension questions; automatically generated readability tests including both traditional Cloze and our Smart Cloze domain-specific variant; annotator perceived ease (Sauro and Dumas, 2009; Rello et al., 2016), which has been used to evaluate readability metrics in the past; and the Flesch Reading Ease Score (FRES) (Flesch, 1948). We compared these metrics across our *Digital Readability* evaluation corpus. Here we describe our corpus, how we generated each of the readability metrics, and how we conducted our validity analysis.

3.1 Digital Readability Corpus

We draw our final evaluation corpus from four source corpora, as follows:

Story corpus. We drew our crowd-worker-created stories from the MCTest (Richardson et al., 2013) dataset, which consists of 500 simple stories created by Amazon Mechanical Turk crowdworkers and validated manually for quality.

Wikipedia corpus. We drew our Wikipedia articles from a corpus of 20,000 Wikipedia articles scraped from Wikipedia and cleaned for quality by Shaoul (2010). We selected Wikipedia articles as a baseline of adult texts against which to compare the domain-specific texts. Wikipedia articles have a mean FRES similar to our domain-specific texts (mean FRES for the wikipedia sample = 47.9; for the health documents = 53.7; and for the security documents = 48.7), suggesting that, at least by one measure, the texts should be similar in readability.

Health corpus. We drew health articles from the 500-document Health Text Readability Corpus (Morales and Wacholder, 2018). This corpus includes consumer health-information documents made available for public use by the CDC, NIH, American Heart Association, American Diabetes Association, and the National Library of Medicine's Medline Plus resource. Worksheets, posters, infographics, and websites are not included. More than half (N=293) of the documents were found in "Easy to Read" collections; that is, the document has been designated by its source agency as appropriate for adults who read at or below a 7th-8th grade reading level.

Security corpus. We collected security advice documents through two methods: (a) asking MTurk workers to create Google search queries for computer security advice, then scraping the top 20 Google results of each query, using the DiffBot API[1] to parse and sanitize HTML body elements within each identified site, and (b) by asking 10 security experts and librarians to recommend digital security advice sources and scraping those websites. These two approaches, along with a manual cleaning process in which we performed spot checks and also manually reviewed 144 documents identified as outliers by FRES or length, generated 1,878 security advice documents.

The last two corpora – health and security advice – are domain-specific: focused on a singular domain and often containing jargon or topics not typically encountered in daily life.

Final evaluation corpus. To ensure comparability of results, we used a standardized subsampling procedure to select 25 documents from each corpus. To ensure that our evaluation captured some variance in documents, we subsampled by length. We first remove the shortest and longest 5% of documents, then we then divide the documents into five bins by length, based on how many standard deviations the length of a given document is from the mean length for that corpus. We manually reviewed all selected documents to ensure that they were on-topic and appropriately clean.[2]

3.2 Readability Metrics

We created three **comprehension questions** for each of the documents in our evaluation corpus: one True/False question and two multiple choice questions with four answer options each, per comprehension question best practices (Anderson, 1972; Day and Park, 2005). Domain-specific questions were written by three co-authors who were domain experts in digital security or in health; the general questions were written by two other co-authors. All 300 comprehension questions were reviewed and edited by a paid comprehension question specialist, who had experience writing and evaluating comprehension questions for the SAT, Discovery Science, and similar organizations; the specialist spent more than 10 hours editing and refining the questions.[2]

We selected the FRES as our **computed measure**, as it is the most-used by number of citations, and anecdotally, by wide-spread application. We computed the FRES for each document using the Python `textstat` package [3].

For our **annotator perception of ease** measurement, we use a single-item question "How easy is this document to read?" with 5-point Likert-item response choices ranging from "Very Easy" to "Very Hard."

Finally, for our **automatically generated readability tests** we used both the traditional Cloze Procedure and our Smart Cloze procedure. Prior work suggests that the frequency of blanks does not significantly affect results (Taylor, 1953). We select set $n = 5$, up to a maximum of 35 target words, for both our traditional Cloze implementation and our Smart Cloze tests, as was done in the original Cloze implementation (Taylor, 1956).

Smart Cloze tool. Prior work to improve Cloze tests offered a multiple-choice variant of the traditional Cloze procedure in which distractors (incorrect answer choices) are randomly drawn from a general dictionary containing other words with the

[1] https://www.diffbot.com

[2] You can find the documents in our corpus, the 300 comprehension questions, and the code for generating traditional Cloze and Smart Cloze tests at: https://github.com/SP2-MC2/Readability-Resources.

[3] https://pypi.org/project/textstat/

same part of speech. While such multiple-choice variants offer improvements in test-taker time, they are potentially inappropriate for domain-specific applications. For example, replacing the word "encryption" in a cybersecurity text with "dog" creates a very easy test. As such, we implemented a novel approach that we call Smart Cloze: we construct a domain-specific dictionary from the same corpus for which we are generating tests and draw distractors from it. The goal is to offer relevant alternatives such as "antivirus" and "key" as distractors for "encryption."

To construct a Smart Cloze test for some document d selected from a domain-specific corpus c, our tool follows the following procedure. First, we bin all of the words in c by part of speech (tagged using Spacy[4]) to create a domain-specific dictionary. We then construct a similar part-of-speech-tagged *document-specific dictionary* using only the words in d. Third, we identify *target words* in d to be replaced by multiple-choice questions. Fourth, we generate distractors for each target. We randomly select up to 14 potential distractors with the same part of speech as the target word from each of the domain-specific and document-specific dictionaries. We then process these distractors in random order, optimizing to obtain two from each dictionary, until we have found four satisfactory distractors.

We measure whether a potential distractor is satisfactory by examining how probable it is that the distractor might substitute for the target word within d. To do this, we first look up the bigram probabilities of the target word (w_c) with its preceding (w_{c-1}) and following (w_{c+1}) words in Google's n-gram corpus. This gives us a baseline for how probable the correct answer is. We then look up bigram probabilities of the potential distractor (say w_d) in combination with the same preceding (w_{c-1}) and following (w_{c+1}) words. Satisfactory distractors have both preceding-distractor and distractor-following bigram probabilities within two orders of magnitude of those for the correct target word. [5] More precisely, a distractor w_d will be accepted if:

$$[P(w_d|w_{c+1}) \geq P(w_c|w_{c+1})] \wedge [P(w_{c-1}|w_d) >= P(w_{c-1}|w_c)]$$

If we do not find four satisfactory distractors (by this definition) within the candidate 28, we instead select the potential distractors with the highest bigram probabilities until we obtain the desired four distractors. Finally, to avoid very small lists of distractor options for certain part of speech (e.g., `TO` only contains to'), we merge parts of speech with small wordlists with larger, related parts of speech until enough unique distractors can be found.

3.3 Validity Evaluation

To evaluate the validity of these readability metrics and compare them, we needed readers to answer the comprehension questions, Cloze tests, and ease question for our documents. We recruited U.S. Amazon Mechanical Turk workers (MTurkers) with a 95% approval rating or above to complete these tasks. Each worker completed one randomly selected readability measure for four documents, including one randomly selected from each of the four corpora. MTurkers were compensated with $1.50 for completing the task. We recruited at least five distinct MTurkers for each type of measure and each document (n=841).

We compare our five readability metrics by examining their construct validity (Cronbach and Meehl, 1955): the degree to which it appears that the measures are accurately measuring readability. To do so, we examine:

- *Content validity:* the degree to which the measures relate to concepts that have been theorized to be relevant to readability; and
- *Convergent validity:* the degree to which related measures (e.g., multiple measures of the same construct) are correlated.

We also explore three factors that are relevant to selecting an appropriate readability measure:

- *Redundancy:* the degree to which any measure is fully, and redundantly, covered by another measure;
- *Score precision:* the precision with which the measure distinguishes between different documents; and
- *Participant burden:* the cost of the measure to the participant (and the researcher) in time to complete.

To assess **content validity**, we examine the degree to which five core linguistic components (narrativity, syntactic simplicity, word concreteness, referential cohesion, and deep cohesion) theorized to be related to readability (Graesser et al., 2004) can *explain the variance* in the measure scores. We

[4] https://spacy.io

[5] We selected two orders of magnitude heuristically to narrow the search space for faster computation while obtaining an appropriate difficulty for the test. Future work could explore alternative heuristics in more detail.

measure these components using the Cohmetrix tool (Graesser et al., 2004). We construct linear regression models, in which the mean measure score for a document is the outcome variable and the input variables are the five linguistic components.

As we wish to understand *which* components are related to which measures, we seek to ensure that we construct a model of best fit. To do so, we perform feature selection via stepwise backward selection, minimizing AIC (Bursac et al., 2008). We further measure applicability to domain-specific texts by including the source corpora of the document as a sixth covariate in the regression model. We set Wikipedia as the baseline for corpora source, as it represents a broad set of non-domain-specific documents with similar FRES to the domain-specific documents.

To assess **convergent validity**, we compute the Pearson correlation between the scores for each readability method in our evaluation dataset. We report the ρ value (strength of the correlation) for correlations significant at $\alpha < 0.05$; Holm-Bonferonni (Abdi, 2010) correction is applied to account for multiple testing.

We also assess **redundancy**, which is not strictly a property of convergent validity, but is relevant when comparing multiple measures that attempt to assess the same construct. Demonstrating that two related measures are correlated establishes convergent validity, but if they are perfectly correlated, then it is unlikely both are needed (Quinn et al., 2010). For this analysis, we construct linear regression models in which the mean score from a given measure for a given document is the outcome variable and the input variables are the three other types of measures (note that we do not include both Cloze measures in any model, but instead construct separate, three-variable models, each with FRES, comprehension questions, ease, and one of the Cloze measures). We consider the degree of redundancy to be the proportion of variance in measure scores explained by the other measures (that is, the R^2 value of this regression model).

To assess **score precision**, we examine the shape of the distribution of scores for a given measure. Per best practice for observing distributions, we do so both through visual inspection and by measuring kurtosis (a statistical measure of the 'tailness' of a distribution) (DeCarlo, 1997).

Finally, we assess **participant burden** in terms of time to complete the task (which also proxies for researcher cost). We compare time by bootstrapping confidence intervals for the mean time for completion of a readability assessment for a given document. Non-overlapping confidence intervals indicate a significant difference in completion time.

3.4 Limitations

Our work is subject to four primary limitations. First, automatic selection of distractors means that there may be differences in the difficulty of different distractors (or variances in difficulty of tests generated by the method when used repeatedly). Based on a manual review of the Cloze tests we conducted before deployment, we did not find trivial distractors to be highly prevalent, given the breadth of words available in each dictionary. However, future work may wish to explore methods for measuring and ensuring consistency in distractor difficulty. Second, MTurk respondents are known to be more educated than the general population, and thus the results of our work may not generalize to low-literacy populations, second-language learners, and others (Redmiles et al., 2019; Ipeirotis, 2010). Third, while we attempted to cover a relatively broad space of online documents, other types of documents (e.g., news articles, Facebook posts) may perform differently. Finally, it is possible that MTurkers were inattentive to our tasks, limiting the validity of our data. We mitigate this possibility by restricting our sample to workers with 95% approval rates on past tasks, as shown in prior work to ensure participant attention to surveys as well as gold-standard 'test' questions (Peer et al., 2014).

4 Results

In this section we summarize our results for content validity (including domain sensitivity of measurements), convergent validity, redundancy, score precision, and participant burden (Table 1).

4.1 Content Validity

We find that comprehension question scores are significantly related to the narrativity ($p = 0.003$) and syntactic complexity ($p = 0.035$) of the document, while performance on comprehension questions is not significantly related to the other three linguistic factors we examined (word concreteness, referential cohesion, deep cohesion) or to

	Linguistic Components (Content Validity)					Additional Considerations			
	Narrativity	Syntactic Simplicity	Word Concreteness	Referential Cohesion	Deep Cohesion	Burden (Mean Time)	Mean Score	Score Precision (Distribution Trend)	Domain Sensitivity
Comprehension	✓	✓				2.86 min	75.7%	exponential	
Traditional Cloze	✓			✓		5.05 min	34.1%	normal	
Smart Cloze	✓	✓		✓		4.55 min	52.4%	normal	✓
Ease			✓			1.67 min	67.1%	uniform	✓
FRES	✓	✓	✓			—	61.0%	uniform	✓

Table 1: Summary of our results on content validity (significant relationships between readability measure and linguistic components theorized to explain comprehension) and other considerations for selecting a readability measure (time for participants to complete a test for a given measure on an average document, average score achieved across documents, trend in the shape of the distribution of scores achieved with a measure, and whether the measure exhibits variation by document domain.

type of document (source corpus).

Traditional and Smart Cloze scores are significantly related to the narrativity (Traditional: $p < 0.001$; Smart: $p = 0.040$) and referential cohesion (Traditional: $p = 0.035$; Smart: $p = 0.008$) of the document. Smart Cloze scores were significantly related to the syntactic complexity ($p = 0.005$) of the document; traditional Cloze scores were not significantly related to syntactic complexity. Finally, neither type of Cloze score was significantly related to deep cohesion or to word concreteness. Smart Cloze scores vary significantly by document domain, while traditional Cloze scores do not. Specifically, Smart Cloze scores are significantly higher for domain-specific documents: those from the health ($p < 0.001$) and security (0.031) source corpora, than for Wikipedia documents. We hypothesize that this is the case because the topics of domain-specific documents are narrower — there are fewer reasonable options for any given blank space — than in the Wikipedia documents, resulting in easier multiple-choice questions. (Anecdotal observation of the generated questions seems to align with this theory.)

Ease perceptions are significantly related only to word concreteness ($p = 0.015$) and document domain: stories ($p = 0.027$) and security ($p = 0.015$) documents are perceived as significantly easier to read than Wikipedia articles. The relationship between ease perceptions and concreteness (and lack of relationship with the other linguistic features we examined) is worth remark. Concreteness of words appears to be easy for readers to assess with a quick glance at an article. This assessment, and their overall perception of ease, may in turn determine whether readers are willing to further read a document they encounter "in the wild," at which point other readability factors may become more relevant. We therefore hypothesize that ease and other measures may complement

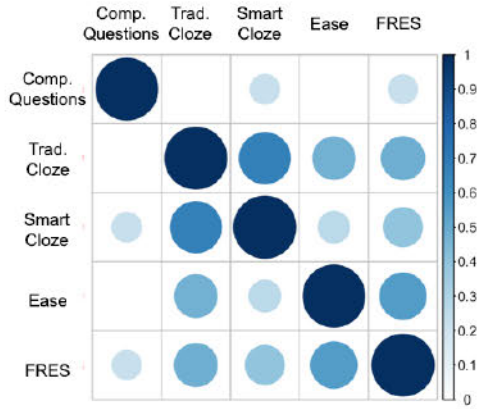

Figure 1: Correlation matrix showing the convergent validity of the measures. That is, the correlation between readability measurement methods. Non-significant correlations ($p > 0.05$) are not shown.

each other. Finally, FRES scores are significantly related to narrativity ($p < 0.001$), word concreteness ($p < 0.001$), and syntactic complexity ($p < 0.001$); but not to either referential or deep cohesion. Perhaps unsurprisingly, FRES scores were significantly higher for stories than for Wikipedia ($p < 0.001$). FRES scores were also higher for security than for Wikipedia ($p = 0.015$), but the health and Wikipedia documents in our sample did not differ in FRES.

While the regression models we constructed explained a significant portion of the variance in scores for ease [6] ($R^2 = 0.504$), FRES ($R^2 = 0.758$), Smart Cloze ($R^2 = 0.389$) and traditional Cloze ($R^2 = 0.334$), these factors explained much less of the variance for comprehension question scores ($R^2 = 0.132$).

[6] This result closely parallels prior work, which predicted perceived ease of Wall Street Journal articles using discourse, vocabulary and length, resulting in an R^2 of 0.503 (Pitler and Nenkova, 2008).

4.2 Convergent Validity

To examine **convergent validity**, we examine the correlation between scores from different measures (Figure 1). Comprehension question scores have the least correlation with scores from the other methods: no correlation with traditional Cloze or ease ratings, and small correlation with FRES ($\rho = 0.22$) and Smart Cloze ($\rho = 0.23$).

This low correlation between comprehension questions and the other methods of measuring readability, together with the low explanation of variance noted above, suggest that comprehension questions assess a combination of the readability of the text and the reader's cognitive abilities, different from the other metrics, which may be more specific to just the text itself (Sarroub and Pearson, 1998). Traditional Cloze, on the other hand, correlates relatively well with all other methods. Perhaps unsurprisingly, there is high correlation ($\rho = 0.71$) between traditional and Smart Cloze scores. Traditional Cloze also correlates well with ease ($\rho = 0.47$) and FRES ($\rho = 0.48$). Smart Cloze correlates less with ease than does traditional Cloze (ease: $\rho = 0.264$, FRES: $\rho = 0.44$). Finally, ease and FRES correlate relatively strongly with each other ($\rho = 0.56$).

4.3 Redundancy

By constructing regression models with the mean score from a given measure on a given document as the outcome variable, and the other measures as the input variables, we find that 4.02% of the variance in the comprehension question scores can be explained by ease perception, FRES, and traditional Cloze (7.92% with Smart Cloze). 20.1% of the variance in traditional Cloze is explained by the other measures, while 22.1% of the variance in Smart Cloze is explained by these measures. 36.0% of the variance in ease perception is explained by mean comprehension question scores, FRES, and traditional Cloze (31.8% with Smart Cloze), while 35.8% of the variance in FRES measurements is explained by scores on comprehension questions, ease perception, and traditional Cloze (37.8% Smart Cloze). Thus, none of the measures are redundant, as the variance in no measure is fully (or even more than 50%) explained by the others.

4.4 Score Precision

Researchers selecting a readability measurement method may also wish to consider the **score precision**: that is, are you trying to find a few bad outliers in a corpus of highly readable documents, or are you expecting a relatively normal distribution of document quality? Figure 2 shows the score distributions by method across all documents and for each document type.

Across domains, the Cloze tests provide the most normal distributions of scores (average traditional Cloze kurtosis = 2.34, average Smart Cloze kurtosis = 3.08)[7]. Cloze scores are thus useful in cases where the relative readability of documents is of interest and where you hypothesize that a normal distribution of readability may be appropriate. The distribution of traditional Cloze scores is transposed left, with a mean of 0.341 (95% confidence interval: [0.329, 0.353]), while the Smart Cloze distribution is centered, with a mean of 0.524 (95% confidence interval: [0.510, 0.537]). Traditional Cloze scores may thus need to be scaled (considered relative to each other rather than as absolute values) to account for this observed ceiling effect.

Ease ratings and FRES, on the other hand, have a more platykurtic distribution (ease: average kurtosis 1.91; FRES: average kurtosis 1.94; fully uniform or platykurtic distribution is 1). A platykurtic distribution has fewer outliers than a normal distribution. Thus, these methods may be more useful in corpora where you expect few readability outliers. Further, ease ratings and FRES both have means higher than 0.5: ease has a mean across domains of 0.671 (95% CI: [0.657, 0.685]) and FRES has a mean of 0.610 (95% CI: [0.594, 0.625]). Given these relatively high means, these methods may also need to be scaled, or may be most useful in cases where you anticipate that an average document in your corpus will be fairly readable. Comprehension questions provide a similarly platykurtic distribution (average kurtosis: 2.06), but with a very high mean (0.757, 95% CI:[0.739, 0.778]).

4.5 Participant Burden

Finally, research is often constrained by resources, including time and budget, and ethically we must be mindful of the burden we impose on our partici-

[7]The kurtosis of a normal distribution is 3; the kurtosis of a unifom distribution is 1.

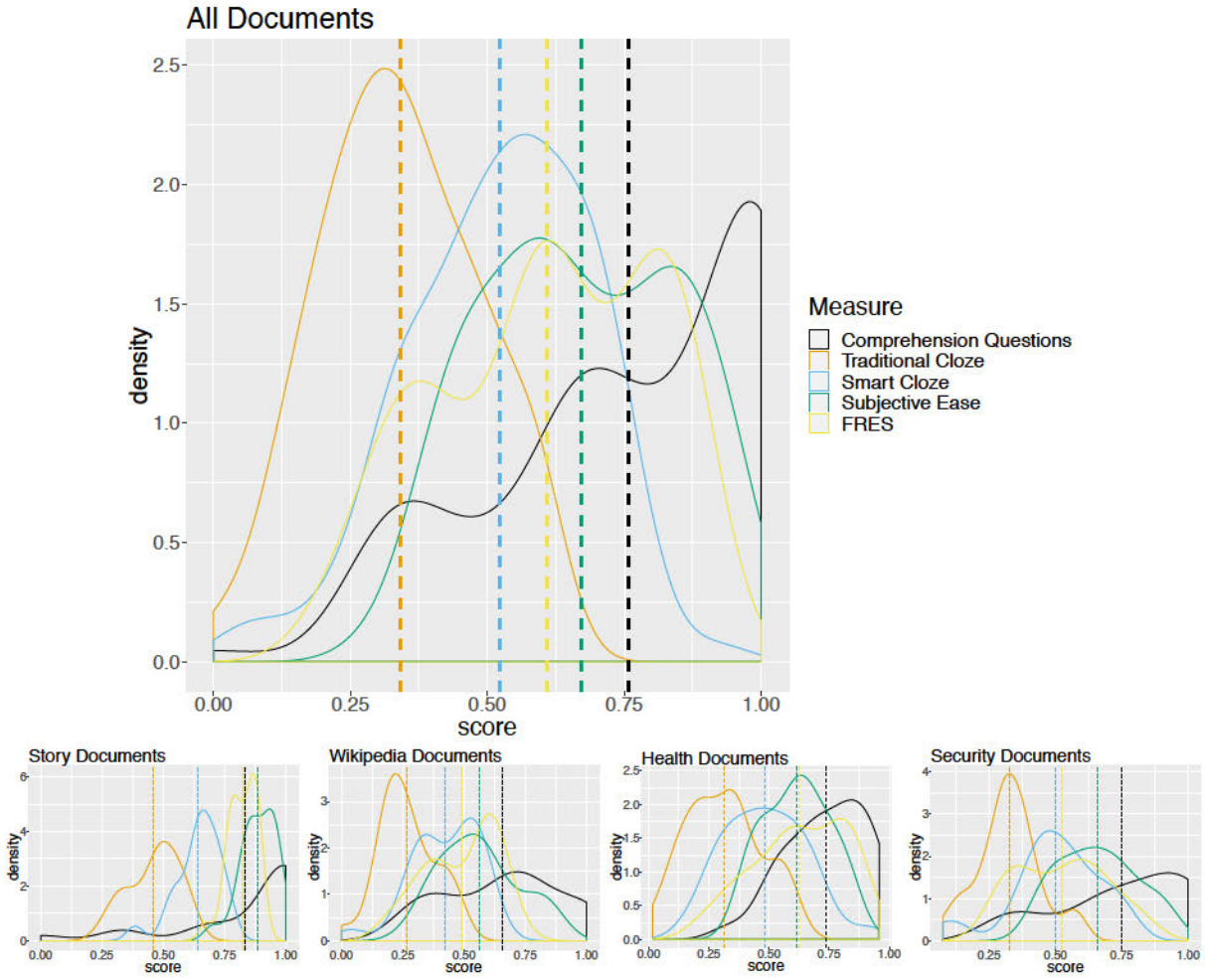

Figure 2: Score distributions by method, across all corpora (top) and by corpus (bottom).

pants. Ease perception (one question) is the fastest test for a worker to complete, with participants spending an average of 1.67 minutes (95% CI: [1.56, 1.78]) per document. Comprehension questions (three questions) took a significantly longer period of time, averaging 2.86 minutes ([2.64, 3.12]) per document, followed by Smart Cloze with an average of 4.55 minutes ([4.08, 4.60]) per document. Finally, traditional Cloze took significantly longer than Smart Cloze, averaging 5.05 minutes per document ([4.72, 5.42]).

5 Moving Forward

In sum, no single readability metric outperformed all the others. Each metric offers different benefits and tradeoffs, and human-written comprehension questions differ the most from the other metrics. We summarize the relevant considerations for selecting a readability metric in Figure 3 and encourage the use of multiple metrics in cases where creating comprehension questions is not scalable.

We find that comprehension questions and Smart Cloze both relate significantly to syntactic complexity, perhaps because they require selection among different possible answer choices. Traditional and Smart Cloze relate to referential cohesion, which makes logical sense, as filling-in-the-blank questions require context from prior sentences. Finally, ease and FRES relate to word concreteness, potentially providing relevant assessments of "first glance" readability reactions. The readability metrics examined also exhibit convergent validity, with the three traditional methods (traditional Cloze, subjective ease, and FRES exhibiting the strongest correlation in scores. Finally, the measures are not redundant: a significant portion of the variance in each remains unexplained by the others.

These different methods offer different levels of precision: the Cloze methods trend toward normal distributions with low (traditional) and centered (Smart) means. On the other hand, ease

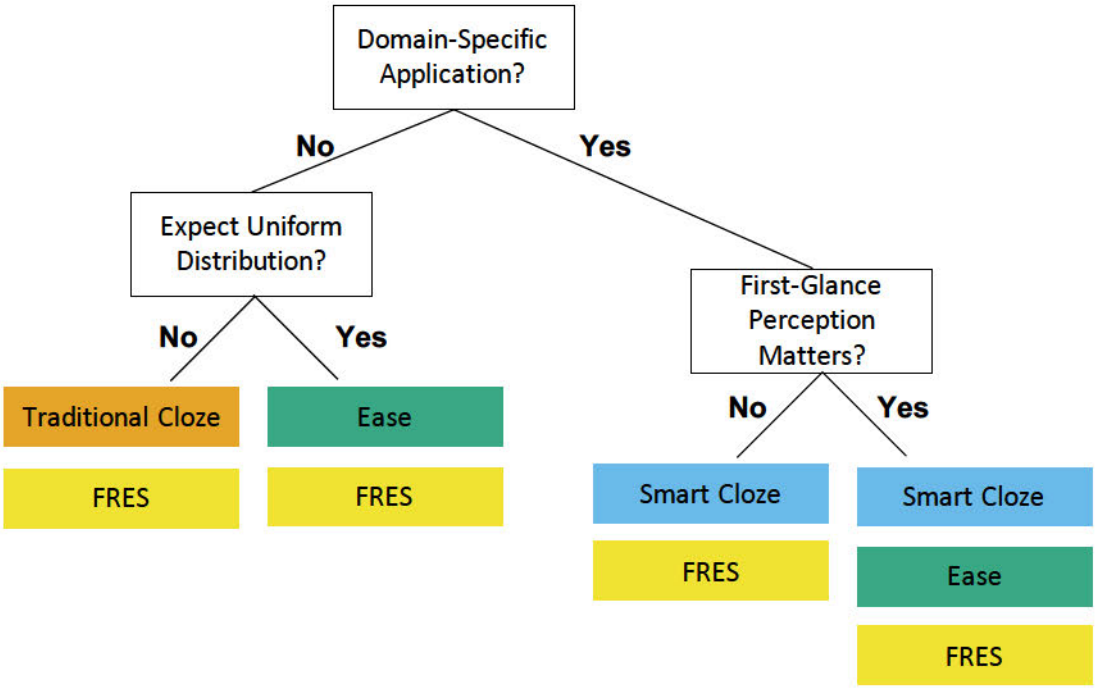

Figure 3: Flow chart for selecting readability measures.

and FRES assessments are more uniformly distributed, with higher means (near 60 and 70%, respectively). Further, Smart Cloze, FRES, and ease measurements all significantly co-varied with document type: Smart Cloze scores were significantly higher for the domain-specific documents (health, security) than for Wikipedia articles, while FRES and ease scores were significantly higher for the story and security documents than for Wikipedia.

While it may be tempting to exclusively use linguistic features because they are cheap and easy to obtain, we find that for the five linguistic factors we explored in this work, these factors explain only 30-50% of the variance in the reader-input readability metrics. Future work may wish to explore additional linguistic factors (Tham, 1987; Hancke et al., 2012; Vajjala and Meurers, 2013; Kate et al., 2010), beyond those covered in this work. In the mean time, our results suggest that when possible, researchers should still consider augmenting these factors with a human-input method. The Smart Cloze tool we propose offers improvements in participant burden, especially for domain-specific documents: scores are higher on average than for traditional Cloze, and tests are 30 seconds faster on average (54 seconds faster for domain-specific documents). However, Smart Cloze is less correlated with perceived ease than traditional Cloze, possibly because the multiple choice option makes the test easier to complete, lessening the chance that participants will "give up." Thus, Smart Cloze is best used in cases where cursory or first glance assessment of readability is less relevant, or in combination with an ease assessment.

Acknowledgements

We are grateful to all reviewers for their thoughtful feedback. We also thank Hanna Wallach introducing us to the ideas of measurement modeling, as well as in depth discussions on the topic, as well as sharing her work with Abigail Jacobs on this topic in their in-preparation paper (Jacobs and Wallach).

This material is based upon work supported by a UMIACS contract under the partnership between the University of Maryland and DoD. Elissa M. Redmiles additionally wishes to acknowledge support from the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 1322106 and a Facebook Fellowship.

References

Hervé Abdi. 2010. Holm's sequential bonferroni procedure. *Encyclopedia of research design*, 1(8):1–8.

Richard C Anderson. 1972. How to construct achievement tests to assess comprehension. *Review of educational research*, 42(2):145–170.

Rebekah George Benjamin. 2012. Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24(1):63–88.

Elmer V Bernstam, Dawn M Shelton, Muhammad Walji, and Funda Meric-Bernstam. 2005. Instruments to assess the quality of health information on the world wide web: what can our patients actually use? *International journal of medical informatics*, 74(1):13–19.

John R Bormuth. 1967. Comparable cloze and multiple-choice comprehension test scores. *Journal of Reading*, 10(5):291–299.

John R Bormuth. 1968. Cloze test readability: Criterion reference scores. *Journal of educational measurement*, 5(3):189–196.

Jonathan C Brown, Gwen A Frishkoff, and Maxine Eskenazi. 2005. Automatic question generation for vocabulary assessment. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. ACL.

Zoran Bursac, C Heath Gauss, David Keith Williams, and David W Hosmer. 2008. Purposeful selection of variables in logistic regression. *Source code for biology and medicine*, 3(1):17.

Chia-Yin Chen, Hsien-Chin Liou, and Jason S Chang. 2006. Fast: an automatic generation system for grammar tests. In *Proceedings of the COLING/ACL on Interactive presentation sessions*. ACL.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135.

Kevyn Collins-Thompson and James P Callan. 2004. A language modeling approach to predicting reading difficulty. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*.

Lee J Cronbach and Paul E Meehl. 1955. Construct validity in psychological tests. *Psychological bulletin*, 52(4):281.

Edgar Dale and Ralph W Tyler. 1934. A study of the factors influencing the difficulty of reading materials for adults of limited reading ability. *The Library Quarterly*, 4(3):384–412.

Richard R Day and Jeong-suk Park. 2005. Developing reading comprehension questions. *Reading in a foreign language*, 17(1):60–73.

Orphée De Clercq and Véronique Hoste. 2016. All mixed up? finding the optimal feature set for general readability prediction and its application to english and dutch. *Computational Linguistics*, 42(3):457–490.

Lawrence T DeCarlo. 1997. On the meaning and use of kurtosis. *Psychological methods*, 2(3):292.

Nell K Duke and P David Pearson. 2009. Effective practices for developing reading comprehension. *Journal of education*, 189(1-2):107–122.

Gunther Eysenbach, John Powell, Oliver Kuss, and Eun-Ryoung Sa. 2002. Empirical studies assessing the quality of health information for consumers on the world wide web: a systematic review. *Jama*, 287(20):2691–2700.

Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. 2009. Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 229–237. ACL.

Rudolf Flesch. 1943. Marks of readable style; a study in adult education. *Teachers College Contributions to Education*.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Thomas François and Eleni Miltsakaki. 2012. 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*, pages 49–57. Association for Computational Linguistics.

Daniela B Friedman and Laurie Hoffman-Goetz. 2006. A systematic review of readability and comprehension instruments used for print and web-based cancer information. *Health Education & Behavior*, 33(3):352–373.

Donna Marie Gates. 2011. How to generate cloze questions from definitions: A syntactic approach. In *AAAI Fall Symposium Series*.

Takuya Goto, Tomoko Kojiri, Toyohide Watanabe, Tomoharu Iwata, and Takeshi Yamada. 2010. Automatic generation system of multiple-choice cloze questions and its evaluation. *Knowledge Management & E-Learning*, 2(3):210.

Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. 2004. Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36(2):193–202.

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability classification for german using lexical, syntactic, and morphological features. In *Proceedings of COLING 2012*, pages 1063–1080.

Michael Heilman. 2011. *Automatic factual question generation from text*. Ph.D. thesis, Carnegie Mellon University.

Ayako Hoshino and Hiroshi Nakagawa. 2007. Assisting cloze test making with a web application. In *Society for Information Technology & Teacher Education International Conference*. AACE.

Juliane House. 2014. Translation quality assessment: Past and present. In *Translation: A multidisciplinary approach*, pages 241–264. Springer.

Panagiotis G Ipeirotis. 2010. Demographics of mechanical turk.

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In preparation.

Sasikiran Kandula, Dorothy Curtis, and Qing Zeng-Treitler. 2010. A semantic and syntactic text simplification tool for health content. In *Annual symposium proceedings*. AMIA.

Rohit J Kate, Xiaoqiang Luo, Siddharth Patwardhan, Martin Franz, Radu Florian, Raymond J Mooney, Salim Roukos, and Chris Welty. 2010. Learning to predict readability using diverse linguistic features. In *Proceedings of the 23rd international conference on computational linguistics*, pages 546–554. Association for Computational Linguistics.

John Lee and Stephanie Seneff. 2007. Automatic generation of cloze items for prepositions. In *Eighth Annual Conference of the International Speech Communication Association*.

Wen-Pin Lin and Heng Ji. 2010. Automatic cloze generation based on cross-document information extraction. In *Asian Conference on Education*.

Annie Louis and Ani Nenkova. 2013. What makes writing great? first experiments on article quality prediction in the science journalism domain. *Transactions of the ACL*, 1:341–352.

Danielle S McNamara, Arthur C Graesser, Philip M McCarthy, and Zhiqiang Cai. 2014. *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

Miraida Morales and Nina Wacholder. 2018. Conceptualizing the role of reading and literacy in health information practices. In *International Conference on Information*. Springer.

Jack Mostow, Joseph Beck, Juliet Bey, Andrew Cuneo, June Sison, Brian Tobin, Joseph Valeri, et al. 2004. Using automated questions to assess reading comprehension, vocabulary, and effects of tutorial interventions. *Technology Instruction Cognition and Learning*, 2:97–134.

Jack Mostow and Hyeju Jang. 2012. Generating diagnostic multiple choice comprehension cloze questions. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*. ACL.

Annamaneni Narendra, Manish Agarwal, et al. 2013. Automatic cloze-questions generation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*.

John W Oller, J Donald Bowen, Ton That Dien, and Victor W Mason. 1972. Cloze tests in english, thai, and vietnamese: Native and non-native performance. *Language Learning*, 22(1):1–15.

Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. 2014. Reputation as a sufficient condition for data quality on amazon mechanical turk. *Behavior research methods*, 46(4):1023–1031.

Juan Pino, Michael Heilman, and Maxine Eskenazi. 2008. A selection strategy to improve cloze question quality. In *Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains. 9th International Conference on Intelligent Tutoring Systems, Montreal, Canada*.

Emily Pitler, Annie Louis, and Ani Nenkova. 2010. Automatic evaluation of linguistic quality in multi-document summarization. In *Proceedings of the 48th annual meeting of the ACL*, pages 544–554. ACL.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the conference on empirical methods in natural language processing*, pages 186–195. ACL.

Kevin M Quinn, Burt L Monroe, Michael Colaresi, Michael H Crespin, and Dragomir R Radev. 2010. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1):209–228.

Earl F Rankin and Joseph W Culhane. 1969. Comparable cloze and multiple-choice comprehension test scores. *Journal of Reading*, 13(3):193–198.

Elissa M Redmiles, Sean Kross, and Michelle L Mazurek. 2019. How well do my results generalize? comparing security and privacy survey results from mturk and web panels to the us. In *IEEE Symposium on Security and Privacy*.

Luz Rello, Martin Pielot, and Mari-Carmen Marcos. 2016. Make it big!: The effect of font size and line spacing on online readability. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM.

Matthew Richardson, Christopher JC Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Loukia Sarroub and P David Pearson. 1998. Two steps forward, three steps back: The stormy history of reading comprehension assessment. *The Clearing House*, 72(2):97–105.

Jeff Sauro and Joseph S Dumas. 2009. Comparison of three one-question, post-task usability questionnaires. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1599–1608. ACM.

Cyrus Shaoul. 2010. The westbury lab wikipedia corpus. *Edmonton, AB: University of Alberta*.

Jennifer Slegg. 2018. Google's use of readability, reading level & vocabulary metrics in search algorithms.

Saku Sugawara, Yusuke Kido, Hikaru Yokono, and Akiko Aizawa. 2017. Evaluation metrics for machine reading comprehension: Prerequisite skills and readability. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 806–817.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624. International World Wide Web Conferences Steering Committee.

Wilson L Taylor. 1953. Cloze procedure: A new tool for measuring readability. *Journalism Bulletin*, 30(4):415–433.

Wilson L Taylor. 1956. Recent developments in the use of "cloze procedure". *Journalism Quarterly*, 33(1):42–99.

Chaffai Tekfi. 1987. Readability formulas: An overview. *Journal of documentation*, 43(3):261–273.

Tuck Meng Tham. 1987. *Linguistic variables as predictors of Chinese text readability*. Ph.D. thesis.

Sowmya Vajjala and Detmar Meurers. 2013. On the applicability of readability models to web texts. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 59–68.