Digital Words: Moving Forward with Measuring the Readability of Online Texts

Elissa M. Redmiles, Lisa Maszkiewicz, Emily Hwang, Dhruv Kuchhal, Everest Liu, Miraida Morales, Denis Peskov, Sudha Rao, Rock Stevens, Kristina Gligorić, Sean Kross, Michelle L. Mazurek, Hal Daumé III
Contact: eredmiles, mmazurek@cs.umd.edu

ABSTRACT

The readability of a digital text can influence people's information acquisition (Wikipedia articles), online security (how-to articles), and even health (WebMD). Readability metrics can also alter search rankings and are used to evaluate AI system performance. However, prior work on measuring readability has significant gaps, especially for HCI applications. Prior work has (a) focused on grade-school texts, (b) ignored domain-specific, jargon-heavy texts (e.g., health advice), and (c) failed to compare metrics, especially in the context of scaling to use with online corpora.

This paper addresses these shortcomings by comparing well-known readability measures and a novel domain-specific approach across four different corpora: crowd-worker generated stories, Wikipedia articles, security and privacy advice, and health information. We evaluate the convergent, discriminant, and content validity of each measure and detail tradeoffs in domain-specificity and participant burden. These results provide a foundation for more accurate readability measurements in HCI.

CCS CONCEPTS

• **Human-centered computing** → *HCI theory, concepts and models*; *Empirical studies in HCI*;

KEYWORDS

readability, comprehension, digital literacy, natural language processing

1 INTRODUCTION

The digital world is full of texts: guides to setting up your printer, privacy policies, wikipedia articles, news articles, WebMD resources, and many more. HCI researchers across various domains measure the readability of such texts in order to ensure equity and accessibility to digital information for non-native-language readers [65], evaluate new design options for enhancing text comprehension [3, 52], push for policy changes to improve the readability of terms of service and privacy policies [4, 29, 36, 37], evaluate whether AI systems can comprehend texts similarly to humans [30], and rank search results [58]. Accurate measurement of the readability of online texts is thus crucial for informing research and ensuring digital equity.

Prior work has used a variety of methods for evaluating the reading level, or readability, of a given text: human-expert-written comprehension question tests presented to human readers, automated generation of readability tests deployed to human readers, and purely computed measures requiring no human input [7, 19, 24, 61]. These measures inherently vary in cost and scalability: computed measures are easy to scale and cheap, while writing and administering multiple comprehension questions for a large corpus of documents may be impossible due to time and cost constraints.

Further, these readability measures were developed for grade-school texts and primarily assessed with grade-school readers. They have rarely been re-validated for use with texts encountered online, which are often domain specific and targeted toward adult readers. Such texts may differ from grade-school texts in a number of ways: they may have different structures, including formats such as bullet points, and may include more abstract words and less cohesive paragraph structures [24]. Such differences may affect the accuracy of computed measures and automatically generated readability tests, which are increasingly used to scale readability measurements in the digital world [5, 16, 20].

To our knowledge, no prior work has assessed the validity of these measures for applications in HCI, nor compared these measures with regard to scalability and adaptability to digital texts. Here, we take a first step toward providing guidance to HCI researchers who seek to effectively empirically measure readability. We evaluate the most commonly used methods for measuring the readability of texts [1] available online by comparing readability scores generated from:

1. Human-written comprehension questions.
2. Automatically generated readability tests
   (a) Traditional Cloze tests [61]: created by removing every *n*th word in a given document and requiring the reader to fill in the blank with the correct word.
   (b) A novel multiple-choice *Smart Cloze* test that we developed specifically for domain-specific texts.
3. Subjective measures [52, 55]: a single Likert item asking users: "How easy is this to read?".

[1] Additional work has explored text *quality*, a larger construct, as discussed in the next section.

(4) Formulaic, computed measures such the Flesch Reading Ease Score (FRES) [19], computed based on sentence and word length.

We compare these methods across a total of 100 documents drawn from four online corpora (Table 1): two domain-specific corpora (health information documents and computer security advice articles) and two general corpora (simple, crowdworker generated stories and Wikipedia articles).

We evaluate each measure in terms of content validity: the degree to which different measures relate to theoretically-grounded linguistic components (e.g., text cohesion, syntactic complexity); convergent validity: the degree to which these measures correspond to each other; redundancy: the degree to which one measure is obviated by another; and score precision: the shape of the distribution of scores and how well it distinguishes among documents. We also detail trade-offs between methods in terms of applicability to domain-specific texts and participant burden (measured by time spent).

We find that FRES, traditional Cloze tests, and single-item subjective evaluations have relatively high convergent validity, with correlation values around 0.5. Further, these measures also exhibit high content validity, correlating strongly with conceptual linguistic components of readability such as text syntactic complexity and word concreteness [24]. These measures also exhibit a lack of redundancy: each measure correlates with a *different* set of linguistic components, and these linguistic components explain only 30-75% of the variance in measure scores, suggesting that the measures are assessing something beyond computable linguistic components. On the other hand, perhaps surprisingly, human-written comprehension questions have low convergent validity — correlation of 0.25 or less with the other measures — suggesting that comprehension questions may measure a different component of readability or something else entirely.

Finally, we contribute two open-science tools. Our open-source *Smart Cloze* tool, which we developed to automatically generate multiple-choice Cloze-style readability tests for domain-specific texts, exhibited some benefits compared to existing measures but also some drawbacks; in the Discussion section we detail when its use may be most appropriate and effective. To enable follow-up research on related topics, we also release our *Digital Readability* evaluation dataset of 100 documents, including 300 comprehension questions written by human experts.

2 BACKGROUND: READABILITY MEASUREMENT

Readability, broadly defined, is a concept indicating how easy or difficult to read a certain text is for someone [66]. Because reading is a complex phenomenon involving both social [22] and cognitive factors [44], there has been a long history of work attempting to estimate readability.

Classical approaches involved answering **human-written comprehension questions** in the form of short answers or essays, oral readability tests, and eventually, multiple choice tests. These tests were always designed or administered by "experts," typically psychologists or schoolteachers [15, 54]. Texts for which most readers correctly answer the comprehension questions were rated easy, while those that stumped many readers were considered hard. It is important to note a limitation of these methods: these approaches inherently blend the reader's ability (to write an essay, to listen to and answer oral questions) with the difficulty of the text itself [54].

Due in part to this shortcoming, but more to the need to scale readability assessment, alternatives began to be developed. In 1923 a new approach was born: using multiple regression formulae to predict readability [64]. Arguably the most popular such formula[2] is the **Flesch Reading Ease Score** (FRES) [18, 19] – a regression model for predicting readability based on the number of sentences, syllables, and words, in a text:

$$206.835 - 1.015\left(\frac{words}{sentences}\right) - 84.6\left(\frac{syllables}{words}\right)$$

The formula was designed to predict "the average grade level of a child who could answer correctly three-quarters of the test questions asked about a given passage" drawn from the McCall-Crabbs' Standard test lessons in reading, a book containing passages and corresponding comprehension questions [59].The FRES formula assumes that longer sentences and words — which often co-occur with complex syntax — indicate greater reading difficulty [12, 17]. Additionally, since shorter words tend to be more common than longer ones in English [57], longer words are assumed less likely to be familiar to the reader.

In 1953, the **C**loze procedure was proposed as a blend of traditional reading comprehension assessment — with human input — and the purely computational method exemplified in Flesch's Reading Ease Formula [61]. The Cloze procedure involves creating readability tests by removing every *n*th word — typically, every 5th [62] — in a given document, and requiring the entity answering the test to fill in the blank with the correct word. Answers are correct if they are an exact match for the original word in that position. After being validated as scalable method of comprehension assessment through comparison with expert-written comprehension questions for grade-school texts [6, 25, 43, 50], the Cloze procedure has been used to assess the quality of other types of documents including translations [27], comprehension of business texts [60], and the quality of text-simplification tools [31].

[2] The most popular by number of citations, and anecdotally, by wide-spread application.

Far more recently, in the 2000s, researchers explored two 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 [10, 21, 23, 33, 34], and multiple-choice Cloze tests in which test-takers select from a set of distractors rather than filling in an open blank, which avoid potential scoring issues with typos and equally-correct synonyms [8, 21, 23, 26, 41, 42, 46].

In parallel, researchers developed supplementary *computed* measures that could be paired with FRES to provide more detail on linguistic features that had long been theorized to be relevant to readability: narrativity, syntactic simplicity, word concreteness, referential cohesion, and deep cohesion [38]. Texts high in narrativity are story-like and usually easier to comprehend. Syntactic complexity describes the difficulty of the sentence structures: "She reads the newspaper in the morning" is a simple sentence, while "Although she was pressed for time and would be late, she took her time reading the newspaper this morning before leaving" is a complex one. Word concreteness describes whether the words in the document are easy to visualize and comprehend: for example, "ball" is highly concrete while "difference" is not. Referential cohesion represents the degree of overlap between content words in sentences in a document [39]. These connections help readers make connections between concepts and maintain a mental "scaffold" of the document. Finally, deep cohesion represents the ease of detecting the connection between causal and logical concepts within a text. Texts that lack connectives between causal and logical text components require the reader to infer these causal and logical relationships [39].

Cohmetrix is one of the commonly used tools for measuring linguistic indices that correspond to these theoretical constructs [24]. Cohmetrix uses NLP techniques such as part-of-speech tagging and latent semantic analysis to measure 108 features associated with reading ease and text cohesion. The principal components of these features align with the five aforementional conceptual components of readability, providing supplementary information about the readability of a text beyond single-number measures like FRES. Along with FRES, these five indices have been shown to well-represent the variance in K-12 texts when evaluated in over 70 different corpora.

Even more recently, in HCI some researchers have used **subjective assessments** of reading ease (typically variants of "How easy is this document to read?") to assess document readability as a component of usability [52]. This approach was modeled on single-item assessments of usability, such as "how easy was this task to complete", which were shown to correlate strongly with other usability measures) [55, 63].

In the past, these various readability measures have been assessed through comparison with text readability ratings given by "reading level experts" (typically K-12 teachers) or through comparison to one other measure (e.g., Flesch Reading Ease compared to comprehension questions). Such assessments are typically very small-scale (10 documents or fewer) and performed exclusively with analog grade school texts. Our work fills these gaps, which limit the relevance of prior evaluations for the HCI community and online texts: (1) we evaluate and compare the most common readability measures with regard to various measures of construct validity, as well as participant time and attention, (2) we evaluate the measures on a larger set of texts (100 documents) that were collected from online environments and in two cases, generated by online volunteers or workers.

Finally, there has also been recent work that goes beyond readability to assess the overall *quality* of text, including factors such as how interesting a topic is to the reader or how grammatically correct the text is (which may be correlated with readability, but is a separate construct) [35, 47]. In this work, we focus strictly on readability — in part because it has been used so frequently for HCI applications — and exclude other quality measures from our comparative evaluation.

3 CORPUS

Here, we describe the digital texts used in our evaluation. We draw our final evaluation corpus from four source corpora:

- Simple stories created by crowdworkers, from [53]
- Wikipedia articles, from [56]
- Health information documents, from [40]
- Security and privacy advice documents, created by the authors

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. We sample 25 documents from each of the four source corpora to form our final evaluation corpus.

Source Corpora

Story corpus. We drew our crowd-worker-created stories from the MCTest [53] dataset which consists of 500 simple stories created by Amazon Mechanical Turk crowdworkers and validated manually for quality. As prior work in Cloze-test generation focuses heavily on simple, general grade-school texts we included these digital variants of simple stories as a baseline.

Wikipedia corpus. We drew our Wikipedia articles from a corpus of 20,000 Wikipedia articles scraped from Wikipedia and cleaned for quality by Shaoul [56]. 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 [40]. 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[3] 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.

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. [4] We manually reviewed all selected documents to ensure that they were on-topic and appropriately clean. Table 1 summarizes the evaluation corpus.

4 IMPLEMENTING READABILITY MEASURES

Here we describe how we scored documents within the evaluation corpus. We apply the most commonly used readability measures summarized in the Background section: multiple-choice comprehension questions, a readability formula (specifically, FRES), Cloze tests (the traditional Cloze procedure, as well as a domain-specific procedure that we developed), and subjective ease measurement using a single item ("How easy is this to read?").

Corpus	Source	Original	Evaluation
Health	[40]	500	25
Security	Author Created	1,878	25
Wikipedia	[56]	20,000	25
Stories	[53]	500	25
Total	–	22,878	100

Table 1: Summary of corpora used in evaluation experiments.

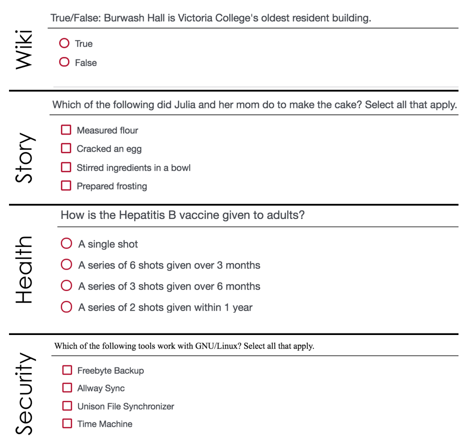

Figure 1: Example comprehension questions.

Comprehension Question Generation

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 [2, 13]. 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 creator of the SAT, Discovery Science, and other similar organizations; the specialist spent more than 10 hours editing and refining the questions. We will open source this dataset of texts and comprehension questions.

Computed Measure: FRES

We selected the FRES as our computed measure, as it is the most-used by number of citations, and anecdotally, by

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

[4] bin_1 was up to one standard deviation below the mean, bin_2 was up to $\frac{1}{4}$ standard deviations below, bin_3 was $\pm\frac{1}{4}$ standard deviations of mean, bin_4 was between $\frac{1}{4}$ and one standard deviation above the mean, and bin_5 was more than one standard deviation above the mean.

wide-spread application. We computed the FRES for each document using the Python `textstat` package [5].

Traditional Cloze Test.

We created an open source platform to automatically construct Cloze tests for our corpus and collect answers to them. To create these tests, we remove every nth word of a given document and replace that word with a text box in which the respondent can type the answer. Prior work on Cloze suggests that the frequency of blanks does not significantly affect results [61]; as such we select set n=5, up to a maximum of 35 target words, as was done in the traditional Cloze procedures [62].

Smart Cloze Test.

Prior work to improve Cloze tests has also offered a multiple-choice variant, in which distractors (incorrect answer choices) are randomly drawn from a general dictionary containing other words with the 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 with words 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[6]) 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. [7] More precisely, a distractor w_d will be accepted if:

$$\left[\tfrac{1}{100}P(w_d|w_{c+1}) \geq \tfrac{1}{100}P(w_c|w_{c+1})\right] \land \left[\tfrac{1}{100}P(w_{c-1}|w_d) >= \tfrac{1}{100}P(w_{c-1}|w_c)\right]$$

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. We release this method as part of our open source Cloze platform.

Subjective Ease Measurement

Drawing on prior work in HCI [52, 55, 63], we constructed the single-item question "How easy is this document to read?" with 5-point Likert-item response choices ranging from "Very Easy" to "Very Hard."

5 EVALUATION APPROACH

Next, we describe how we evaluated these measures: specifically, how we collected and analyzed our evaluation data, as well as the limitations of our approach.

Data Collection

To collect evaluation data, we needed humans to answer the comprehension questions, Cloze tests, and ease question for our documents. We recruited Amazon Mechanical Turk workers (MTurkers) to complete these tasks. Each worker saw four documents, one randomly selected from each of the four corpora, and only one randomly selected readability measure. For example, a worker randomly assigned to comprehension questions answered four comprehension questions, one from each corpus, and no other questions. We recruited U.S. MTurkers with a 95% approval rating or above to avoid the need for explicit attention check questions [45].MTurkers were compensated with $1.50 for completing the task. We recruited at least 5 distinct MTurkers for each type of measure and each document (n=841).

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

[6] https://spacy.io

[7] 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.

Analysis

We compare the five measures described in the prior section by examining their validity, their applicability to domain-specific documents, their precision of measurement, degree of redundancy, and the burden they impose on participants. Scores for readability measures are scaled to be out of 100: for human-generated scores, as a percentage of correct answers. Documents are assigned a mean score from each procedure that required human test takers, and a single score from the linguistic measures.

Specifically, we explore construct validity [11]: 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.
- Convergent validity: the degree to which related measures (e.g., multiple measures of the same construct) are correlated.

We also explore three additional 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.
- 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 the five linguistic components (narrativity, syntactic simplicity, word concreteness, referential cohesion, and deep cohesion) theorized to be related to readability (see Background section for details) can *explain the variance* in the measure scores. We measure these components using the Cohmetrix tool [24]. 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 [9]. Factors are considered significant at $p < 0.05$. We also report R^2 as the measure of variance explained by the model. 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. If significant, this factor can tell us the degree to which scores on a measure are correlated to domain.

To assess **convergent validity**, we compute the Pearson correlation between the scores for each readability method in our evaluation dataset. We report the rho value (strength of the correlation) for correlations significant at $alpha < 0.05$; Holm-Bonferonni [1] 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 [49]. 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) [14].

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.

Limitations

Our work is subject to three primary limitations. First, we recruit MTurkers to complete our measures, in part because they are commonly used in HCI studies [32, 32]; however, 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 [28, 51]. Future work could evaluate readability measures for HCI tasks on these populations. Second, 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 [45].

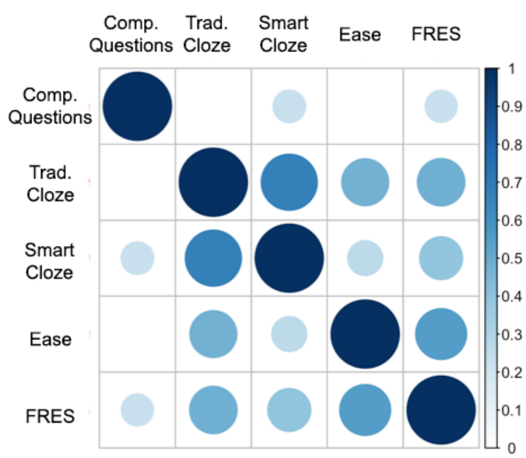

Figure 2: Correlation matrix showing the convergent validity of the measures. That is, the correlation between readability measurement methods. Non-significant correlations ($p > 0.05$) are not shown.

6 RESULTS

In this section, we report our results for content validity (including domain sensitivity of measurements), convergent validity, redundancy, score precision, and participant burden. We summarize our results in Figure 2 and Table 2.

Content Validity

First, we seek to understand the relationship between the different measures and the five linguistic components discussed in the Background section. In addition to providing a measure of **content validity**, examining these relationships allows us to evaluate whether the human-input methods provide any advantage over linguistic methods.

Comprehension questions are significantly related to the narrativity ($p = 0.003$) and syntactic complexity ($p = 0.035$) of the document. [8] They are not related to the other three factors or to the source corpus.

Traditional Cloze scores are significantly related to the narrativity ($p < 0.001$) and referential cohesion ($p = 0.035$) of the document. They are not related to the other factors or to document domain. Smart Cloze scores are also significantly related to narrativity ($p = 0.040$) and referential cohesion (0.008), but in addition they are significantly related to syntactic complexity ($p = 0.005$) and to document domain. 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 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 any other features) 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 each other.

Finally, FRES scores are significantly related to narrativity ($p < 0.001$), concreteness ($p < 0.001$), and syntactic complexity ($p < 0.001$). 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 [9] ($R^2 = 0.504$), 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 ($R^2 = 0.132$).

Convergent Validity

Next, to examine **convergent validity**, we examine the correlation between scores from different measures (Figure 2). We find that the comprehension question scores have the least correlation with scores from the other methods: no significant correlation with scores generated by 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 may get at a different aspect

[8]All regression coefficients reported in this section go in the anticipated direction. For example, more narrative documents correlate with higher readability scores on a given measure, while syntactically more complex documents had lower scores.

[9]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 [48].

	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 2: 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.

of readability than the other measures. Specifically, we hypothesize — in line with theoretical work on reading comprehension [54] — that comprehension questions assess a combination of the readability of the text and the reader's cognitive abilities. This is further supported by the fact that performance on comprehension questions does not vary by document type – the cognitive load required for completing a comprehension question may be equal, even for arguably simpler texts such as stories.

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 ($\rho = 0.56$)

Redundancy

We also evaluate **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.

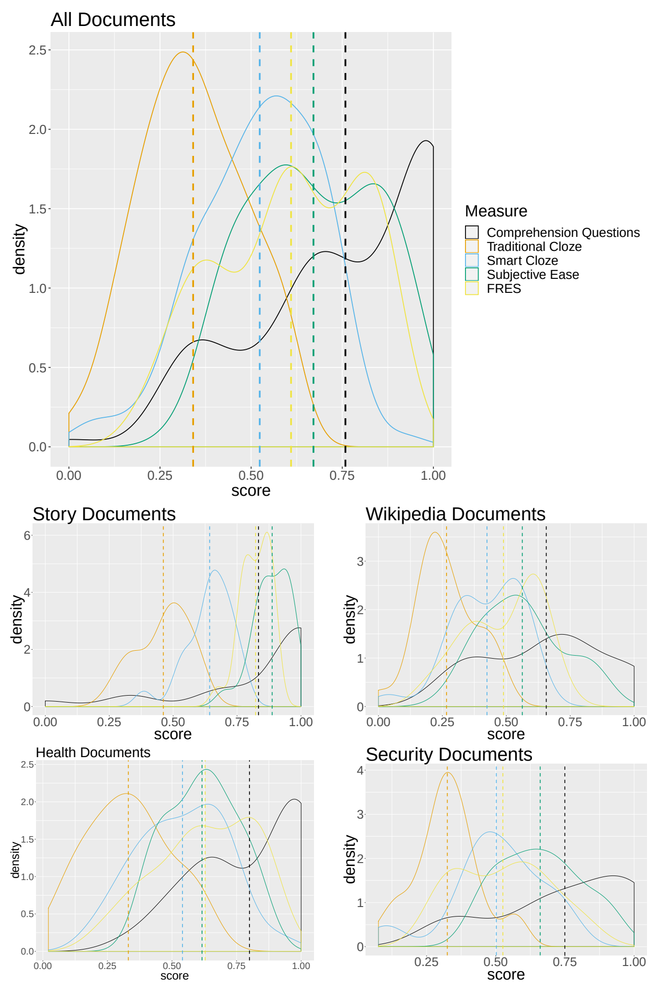

Figure 3: Score distributions by method, across all corpora (top) and by corpus (bottom).

Score Precision

Researchers selecting a readability measurement method may also wish to consider the **score precision**: that is, are

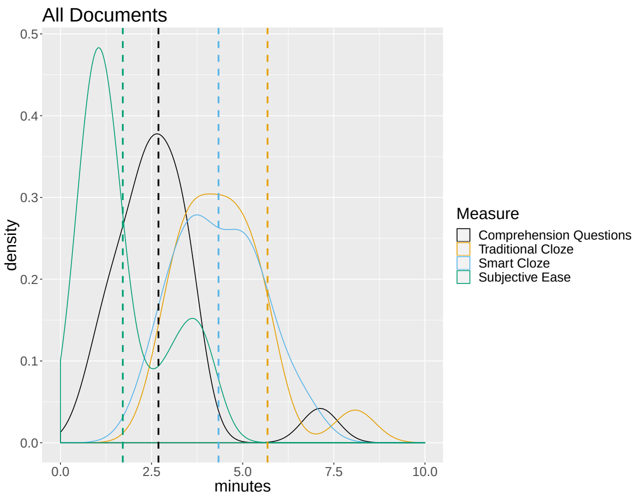

Figure 4: Distribution of completion times from each method across all for corpora.

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 3 shows the score distributions by method across all documents and for each document type.

Across domains, the Cloze tests provide the most normal distributions (average traditional Cloze kurtosis = 2.34, average Smart Cloze kurtosis = 3.08; kurtosis of 3 is normal) of scores. 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 (an example is a uniform 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]).

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 participants. With this in mind, we evaluate **participant burden** by assessing the time required for MTurkers to complete tests across the different measures. Ease perception (one question) is the fastest, with participants spending an average of 1.67 minutes (95% CI: [1.56, 1.78]) per document. Comprehension questions (three questions) were second-fastest, at an average of 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. Traditional Cloze takes slightly but significantly longer than Smart Cloze (up to 35 questions each): an average of 5.05 minutes per document ([4.72, 5.42]). As mentioned above, the non-overlapping confidence intervals indicate significant differences between all four measures. Figure 4 summarizes these results. Translated into research costs, if researchers sought to pay U.S. federal minimum wage ($7.25) then it would cost $0.20 per document for participant ease ratings, $0.35 for participant responses to comprehension questions (plus at least $3 to create three expert-written and reviewed comprehension questions), $0.55 per document for Smart Cloze answers, and $0.61 per document for traditional Cloze. FRES is free, excepting computational power for computing the measure depending on corpus scale.

7 MOVING FORWARD

In sum, in our examination of content validity — the degree to which the measures relate to concepts that have been theorized to be relevant to readability — we find that all of the measures but ease relate to the narrativity of a given document. 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.

Most of the measures are also relatively correlated in their measurements, that is, they exhibit convergent validity. The three traditional methods (traditional Cloze, subjective ease, and FRES) exhibit relatively strong correlation, with ρ near 0.5. Smart Cloze is similar to traditional Cloze overall, but less (although still significantly) correlated with ease than traditional. Further, none of the measures are redundant: a significant portion of the variance in each remains unexplained by the others.

Additionally, the different methods tend toward different levels of score precision: the Cloze methods trend toward normal distributions with low (traditional) and centered (Smart) means. On the other hand, ease 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. Finally, ease perception is the fastest measure for readers to complete, followed by comprehension questions, Smart Cloze, and finally traditional Cloze.

When To Use Which Measure

What do these findings mean for selecting readability measures? First, comprehension questions are least similar to the other measures: they appear to simultaneously measure at least two constructs: readability *and* cognitive ability), as has been theorized in prior work [54], and correlate only with narrativity of texts, not with any other conceptual element theorized to be relevant to readability. Further, comprehension questions are difficult to scale to the needs of HCI research and digital documents, as a single comprehension question costs at least $1 (in expert time) to create. As such, we exclude comprehension questions from further consideration.

Next, it may be tempting to exclusively use linguistic features because they are cheap and easy to obtain. We find, however, that linguistic factors explain only 30-50% of the variance in the measures that require human input; thus, significant information is lost by using only linguistic measures. While a useful approximation, when possible researchers should still consider augmenting these factors with a human-input method.

Which human-input method, then should be selected? Researchers and practitioners may wish to consider whether their application is domain-specific or broad in nature. For domain-specific applications, Smart Cloze may be a good choice for reducing costs and participant burden: scores are higher on average than for traditional Cloze, and tests are 30 seconds faster on average (54 seconds faster for domain-specific documents), suggesting that Smart Cloze tests are easier to take, for corpora focused on a narrow domain or topic. Smart Cloze is, however, 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

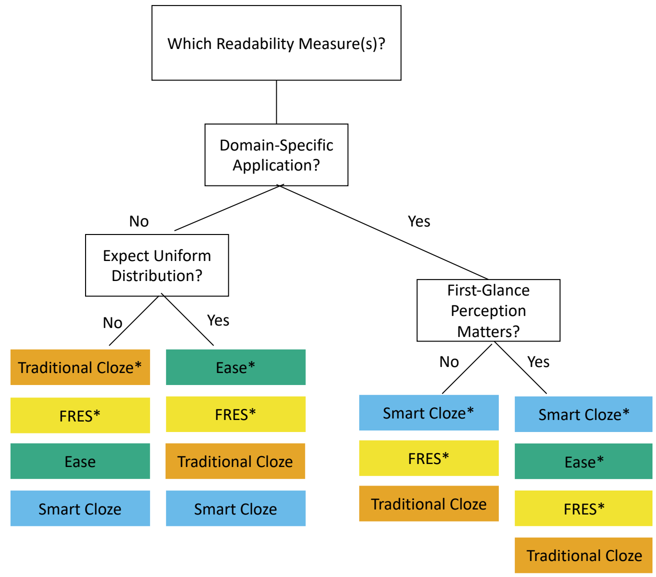

Figure 5: Flow chart for selecting readability measures. The suggested minimum set of measures for a given flow are marked with *.

of readability is less relevant, or in combination with an ease assessment.

For broader corpora of documents, researchers may wish to consider the expected distribution of documents: when a fairly uniform distribution of readability is expected, ease may be the cheapest human-input measure; in contrast, if a more normal distribution is expected, traditional Cloze may be more appropriate. When possible, combining both measures may also provide broader insight.

In conclusion, we find that ease, FRES, traditional Cloze, and Smart Cloze are all relatively valid measures of readability. Each correlates strongly with a different set of linguistic factors, and none is fully explained by another measure. Thus, any combination of computed (e.g., FRES) and human-input (Cloze, ease) measures should be relatively effective. However, considerations of cost, participant burden, and expected readability distribution may suggest particular (combinations of) measures as optimal. We summarize these recommendations in Figure 5.

REFERENCES

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

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

[3] Katrin Angerbauer, Tilman Dingler, Dagmar Kern, and Albrecht Schmidt. 2015. Utilizing the Effects of Priming to Facilitate Text Comprehension. In *Proceedings of the 33rd Annual ACM Conference Extended*

Abstracts on Human Factors in Computing Systems (CHI EA '15). ACM.

[4] Annie I Anton, Julia Brande Earp, Qingfeng He, William Stufflebeam, Davide Bolchini, and Carlos Jensen. 2004. Financial privacy policies and the need for standardization. *IEEE Security & privacy* 2, 2 (2004), 36–45.

[5] 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 (2005), 13–19.

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

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

[8] 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.

[9] 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 (2008), 17.

[10] 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.

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

[12] 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 (1934), 384–412.

[13] Richard R Day and Jeong-suk Park. 2005. Developing Reading Comprehension Questions. *Reading in a foreign language* 17, 1 (2005), 60–73.

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

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

[16] 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 (2002), 2691–2700.

[17] 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*. ACL, 229–237.

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

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

[20] 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 (2006), 352–373.

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

[22] James Paul Gee. 2015. Three paradigms in reading (really literacy) research and digital media. In *Reading at a Crossroads?: Disjunctures and Continuities in Current Conceptions and Practices*. Taylor and Francis Inc.

[23] 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 (2010), 210.

[24] 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 (2004), 193–202.

[25] Michael Heilman. 2011. *Automatic factual question generation from text*. Ph.D. Dissertation. Carnegie Mellon University.

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

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

[28] Panagiotis G Ipeirotis. 2010. Demographics of mechanical turk. (2010).

[29] Carlos Jensen and Colin Potts. 2004. Privacy policies as decision-making tools: an evaluation of online privacy notices. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM.

[30] Douglas Jones, Edward Gibson, Wade Shen, Neil Granoien, Martha Herzog, Douglas Reynolds, and Clifford Weinstein. 2005. Measuring human readability of machine generated text: three case studies in speech recognition and machine translation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE.

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

[32] Aniket Kittur, Ed H Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM.

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

[34] Wen-Pin Lin and Heng Ji. 2010. Automatic Cloze Generation based on Cross-document Information Extraction. In *Asian Conference on Education*.

[35] 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 (2013), 341–352.

[36] Ewa Luger, Stuart Moran, and Tom Rodden. 2013. Consent for all: revealing the hidden complexity of terms and conditions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM.

[37] Aleecia M. McDonald, Robert W. Reeder, Patrick Gage Kelley, and Lorrie Faith Cranor. 2009. *A Comparative Study of Online Privacy Policies and Formats*. Springer Berlin Heidelberg, 37–55.

[38] Danielle S McNamara, Arthur C Graesser, Zhiqiang Cai, and Jonna M Kulikowich. 2011. Coh-Metrix easability components: Aligning text difficulty with theories of text comprehension. In *annual meeting of the American Educational Research Association, New Orleans, LA*.

[39] Danielle S McNamara and Walter Kintsch. 1996. Learning from texts: Effects of prior knowledge and text coherence. *Discourse processes* 22, 3 (1996), 247–288.

[40] Miraida Morales and Nina Wacholder. 2018. Conceptualizing the Role of Reading and Literacy in Health Information Practices. In *International Conference on Information*. Springer.

[41] 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.

[42] 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*.

[43] 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 (1972), 1–15.

[44] Scott G Paris and Steven A Stahl. 2005. *Children's reading comprehension and assessment.* Routledge.

[45] 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 (2014), 1023–1031.

[46] 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.*

[47] 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.* ACL, 544–554.

[48] 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.* ACL, 186–195.

[49] 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 (2010), 209–228.

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

[51] 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 Security and Privacy.*

[52] 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.

[53] 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.*

[54] 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 (1998), 97–105.

[55] 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.* ACM, 1599–1608.

[56] Cyrus Shaoul. 2010. The westbury lab wikipedia corpus. *Edmonton, AB: University of Alberta* (2010).

[57] Bengt Sigurd, Mats Eeg-Olofsson, and Joost Van Weijer. 2004. Word length, sentence length and frequency–Zipf revisited. *Studia Linguistica* 58, 1 (2004), 37–52.

[58] Jennifer Slegg. 2018. Google's Use of Readability, Reading Level & Vocabulary Metrics in Search Algorithms. http://www.thesempost.com/googles-use-readability-vocabulary-search-algorithms/

[59] Kathleen C Stevens. 1980. Readability formulae and McCall-Crabbs standard test lessons in reading. *The Reading Teacher* 33, 4 (1980), 413–415.

[60] Kevin T Stevens, Kathleen C Stevens, and William P Stevens. 1992. Measuring the readability of business writing: The cloze procedure versus readability formulas. *The Journal of Business Communication (1973)* 29, 4 (1992), 367–382.

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

[62] Wilson L Taylor. 1956. Recent developments in the use of "Cloze Procedure". *Journalism Quarterly* 33, 1 (1956), 42–99.

[63] Donna Tedesco and Tom Tullis. 2006. A comparison of methods for eliciting post-task subjective ratings in usability testing. *Usability Professionals Association (UPA)* 2006 (2006), 1–9.

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

[65] Chen-Hsiang Yu and Robert C Miller. 2010. Enhancing web page readability for non-native readers. In *Proceedings of the SIGCHI conference on human factors in computing systems.* ACM.

[66] Mostafa Zamanian and Pooneh Heydari. 2012. Readability of texts:: State of the art. *Theory and Practice in Language Studies* 2, 1 (2012), 43.