Maintaining understandability when representing data usages based on results from Readability and Understandability research

Maximilian Fest
Seminar Inverse Transparency (WS19/20)
Advisor: Valentin Zieglmeier
ge34gub@mytum.de

***Abstract*—Results from readability and understandability research are reviewed, with a focus on their usefulness to potential Inverse Transparency solutions. Much of this research concentrates on the measurement of readability, with results from linguistics providing the framework from which measures are evaluated. A range of measures are described, including traditional readability formulas, modern computational approaches, and user-centric approaches. Traditional formulas attempt to predict text readability based on a small number of simple characteristics, imposing limitations on their reliability. More recent computational approaches analyse texts based on hundreds of features, and are therefore better placed to deal with the complexity of readability assessment. User-centric measures heed common critique of these approaches, by emphasising the interaction between the reader and the text in their assessment. A short survey suggests the difficulties of ensuring or enforcing understandability in an Inverse Transparency solution.**

I. Introduction

Any data privacy solution built on the concept of Inverse Transparency relies on the use of data usage reports, from which an employee may gauge how and to what end their personal data was used. Of course such a solution is ineffective if these usage reports are not understandable to the employee.

Fortunately, a large body of research has accumulated over the past century, concerned with what makes one text easier to understand than another. Much of this has focused entirely on features of the text itself, measuring 'readability' to encompass the more abstract notion of understandability. Although the terms understandability and readability have been used roughly synonymously, some have pointed out critical differences between the two. Whereas readability measures some property of the text, how understandable it is depends also on the person engaging with it (his education, background, etc…). In the following, we assume the **understandability** of a text to mean the ease with which a person can understand it, and consider **readability** a useful proxy for this quality.

Readability research can be divided roughly into the following four sections, by which this literature review is organised:

1) Traditional Readability Formulas
2) Computational linguistics
3) User-centric approaches
4) Linguistics and presentational factors

The first three sections of research were motivated in large part by the need for an objective measure of readability. Traditional Readability formulas first addressed this need in the mid-20th century, providing easily-calculable text-based scores that saw wide application. With improved computational resources, these formulas were superseded by approaches in computational linguistics, which analysed a text based on more complex features ignored by the traditional formulas. User-centric approaches highlight the importance of taking characteristics of the target reader into account when evaluating the understandability of a text. These approaches may produce scores, like the text-based measures, or simply provide feedback. Finally, linguists provide the frame of reference by which these methods can be compared and evaluated, and approaches discussed in the latter two sections were inspired by results in linguistics.

Such measures may then be used to ensure that texts are written at a level that is easily understandable to the intended reader. Their most common use is in assigning books to school grade levels for which they would be most appropriate. However, the value of readable writing permeates all areas where a writer must effectively communicate with an audience. That includes lawmaking. For example, in 2010 the U.S. Securities and Exchange Commission (SEC) introduced their plain writing initiative, motivating research into what measures could be used to objectively enforce a minimum level of readability.

To conclude the review, a questionnaire will be evaluated on the Amazon Mechanical Turk (AMT), intended to evaluate AMT as a tool for collecting readability judgements, and to suggest how a diverse population of AMT workers responds to passages resembling data usage reports.

To further focus this review, some assumptions are made about the data usage reports provided in an Inverse Transparency solution:

1) Usage reports will be short, i.e. no longer than a few sentences
2) They must be automatically selected or generated following a data usage

Given the number of possible data usages, it is not unlikely that usage reports must, to some extent, be generated. Research in each of the sections above will be evaluated according to these assumptions.

A distinction must also be drawn between the readability of prose and of software. The latter measures readability of program code, based on concepts unique to programming like comments or loop structure, and therefore has no relevance to this literature review.

II. Traditional Readability Formulas

A. Overview

The notion that the readability of a text may be predicted based on a few simple characteristics dates back at least as far as 1921, to the publication of Kitson's 'The Mind of the Buyer' [1]. In it, he describes how and why different newspapers and magazines attract different readerships, positing that this depends in large part on their readability, and that readability may be predicted by measuring average sentence and average word length of texts within these publications. Readability research continued in much the same vein throughout the 20th century, and by the 1980's, there were over 200 readability formulas and over 1000 studies published about them [2].

A readability formula aims to predict the readability of a text based on a very limited number of basic syntactic and semantic features. One popular such formula, the Flesch-Kincaid score relates a texts readability to its average number of syllables **s** per word and its average words per sentence **w** [3]:

$$\mathbf{G} = 0.39 * \mathbf{w} + 11.8 * \mathbf{s} - 15.59$$

It is intended to return the academic grade level **G** of students for whom the text is most appropriate. **s** and **w** respectively serve as proxies for sentence (syntactic) complexity and word (semantic) difficulty.

Since the mid-20th century, these formulas have been applied extensively, in diverse contexts. To many researchers, this suggests that they have stood the test of time, and their validity as objective measures of readability is often simply assumed (for example [4]). Even in the 21st century, such simple measures are being developed [5], and comprehensive reviews point to consistently strong performance in predicting the readability of texts [2].

A useful property of these formulas is that they assign a numeric score to a text. This score can be used to define a minimum readability level for a specific class of texts, for instance to prevent obfuscation in informed consent forms [6] or in annual reports [7]. It also serves as a useful point on which to compare different texts.

Formulas that rely on word-frequency lists to estimate vocabulary difficulty form an important subclass of readability measures. The new Dale-Chall readability formula is the most frequently cited of these [8]. It classifies words into 'easy' and 'hard', based on containment in a list of 3000 'easy' words. Such an approach can, to some extent, take into account the background of the reader, by adjusting these vocabulary lists to the target audience.

The Lexile Measure is a more recent readability formula that has seen widespread application. It was developed in 1989 in a commercial venture owned by Metametrics Inc., who have now ranked 300,000 books according to their Lexile score [1]. Like the Dale-Chall measure, it uses a word frequency measure, calculated from a 5 million word corpus, as well as the log average sentence length. Despite the slight increase in complexity, it did not perform significantly better than 9 other readability formulas tested in the defining study [9]. What sets apart the Lexile measure, is that it seeks to provide a score for both the reader and the text, by testing the reader's comprehension of passages with a known lexile score. This way, someone with a lexile score of 750 may be matched to books with the same score, and be expected to answer roughly 75% of multiple-choice comprehension questions about this book correctly [9]. This is a unique approach, and perhaps why the Lexile measure has garnered more attention than comparable formulas.

B. Limitations

Especially when computational resources were limited, these measures were appealing for their simplicity. Calculating readability scores with them requires only measuring basic properties of the text, without needing to involve human test subjects. However, many have criticised the traditional readability formulas, in large part because they oversimplify a complex issue. Many characteristics of text that are known to influence its understandability are ignored by readability formulas [10]. In many cases, this is because it was believed that simpler index variables were sufficiently correlated to more complex ones to omit the latter, with the benefit that the formulas are easier to apply [11].

The readability formulas were not developed according to any model of reading comprehension. Rather, they observed simple statistical correlations between surface-features of a text and perceived readability, implying weak construct validity [12].

The readability researcher GR Klare was one of the first to discuss the limitations of these formulas and to investigate their usefulness [13]. In a 1976 meta-study, Klare analysed 36 studies that attempted to improve understandability by improving readability as measured by such formulas. He found that only half succeeded, and that they had to improve readability by an average of 6.5 grade levels to improve reader comprehension [11]. In a later study, Charrow and Charrow [14] tested the readability of jury instructions before and after revising them to improve comprehension, and found that readability even decreased in some cases. Proponents of the formulas justified this by arguing that the scores produced by them can be likened to the readings on a thermometer: you can use them to tell the temperature of a room, but you cannot manipulate them to change the temperature [15].

Furthermore, most of these studies validated their formulas only in terms of earlier readability formulas. The earliest were, in turn, validated using the McCall-Crabbs lessons, although these were never intended to measure understandability [16]. This indicates at best a weak statistical basis for the validity

[1] https://lexile.com/for-researchers/

of these formulas as predictors for understandability [16]. Their validity can further be called into question because they depend heavily on the context they were devised in. A text that would have been deemed readable in the mid-20th century could very well no longer be, and vice versa [17].

Bailin and Grafstein point out that, given these inconsistencies, it can be counter-productive to attach scores that appear objective to texts, because this apparent objectivity might give it credibility over expert opinion or even intuition, when it is not clear that it should [10]. Moreover, she argues that such a score makes little sense in the first place, because the understandability of a text is a product of the interaction between the text and a person, and is therefore not a measurable property of the text.

In a recent study, Loughran and McDonald [18] evaluate the Fog Index as a predictor of the readability of financial disclosures. Like the Flesch formula, the Fog Index is calculated based on average sentence length and the proportion of words with more than two syllables. They conclude that "measures like the Fog Index are poorly specified in the realm of business writing", and find that the file-size of these financial disclosures is a better predictor of their readability. To justify this, they make the common argument that readability formulas fail to take into account the readers' background. Interestingly, this study was motivated by the SEC's consideration of using the Fog Index to help identify poorly written financial documents.

In spite of the weaknesses of traditional readability formulas, useful conclusions can be drawn from this research. First, the understandability of a text can very likely not be determined from a few basic characteristics. Consequently, texts can not be improved by improving their readability scores, as calculated by traditional readability formulas. Second, there is a component of understandability that is personal. This means that the understandability of a text changes over time and varies from person to person. Third, a measure that has predictive power does not necessarily provide useful information about how to improve a text. Failing to address these criteria, the traditional readability measures are misplaced in scenarios in which accurate assessment of readability is necessary.

III. Computational readability assessment

To address the construct weakness of classical formulas, researchers devised methods to analyse texts based on cognitive theories of reading and reading comprehension. Rather than evaluating text difficulty based on surface features like sentence difficulty, they hypothesised that understandability of a text was more closely related to higher level discourse features like coherence [19]. The measurement of readability became an interdisciplinary endeavour, with important contributions from computer science, cognitive psychology, linguistics, artificial intelligence, and more. As a result, relevant studies were to be found under a variety of headings: automated readability assessment, computational linguistics, natural language processing (NLP), discourse processing, and more. For more detailed reviews of such approaches, see [20] and [21].

Because of the complexity inherent in representing these cognitive mechanisms, measures inspired by them usually require some form of automation. Typically, rich, annotated versions of texts are used because they offer more linguistic features from which readability can be predicted, for example with sophisticated machine learning methods. Other approaches develop language models that represent certain classes of texts (for example: 8th grade science textbooks), to predict the likelihood of a given word or phrase within this model. It is then argued that less likely phrases are ones the reader will be less familiar with, and thus will be harder to understand [22].

A. Training corpora

What these approaches have in common, is that they require large training corpora to learn from. Therefore none of the measures described in the following section can be seen as standalone measures of readability. Rather, reliable readability measurements of texts in the training corpus are given, and based on these measurements a program learns how to predict the readability of unknown texts. Many such labelled corpora exist, but often they rely on measurements using traditional readability formulas. Other text corpora like the Penn Discourse Treebank contain texts expert-annotated with higher-level discourse features, but no judgement is made of their readability. Particularly when training models for a specific target audience, it may be necessary to compile a corresponding training corpus. Studies concerning readability in languages other than English have pointed to a lack of training corpora containing texts written in those languages [23], [24].

In their 2017 study, Crossley et. al. describe a number of NLP tools that they used for part-of-speech tagging and automatic feature extraction [25]. Simply put, these tools automate the process of annotating words in a text as belonging to a specific *part of speech*, accounting for their definition and their context within the text. Examples of such are the Tool for the Automatic Analysis of Cohesion (TAACO) [26] and the Tool for the Automatic Analysis of Lexical Sophistication (TAALES) [27]. Such tools can significantly reduce the effort involved in gathering annotated training corpora.

One difficulty often encountered by researchers in gathering readability judgements from a population, is that humans have difficulty discerning nuanced classes of readability levels. For example, given some text, it is difficult to assign a grade-level from 1 to 12. This makes it difficult to gather reliable training data for machine learning measures. Researchers have circumvented this problem by asking test subjects simply to chose the more readable between two texts, from which they build a comparator that allows them to sort texts in a corpus by their readability [25], [28]. This approach forfeits the usefulness of simply assigning a score by which relative differences between the texts may be estimated. However, for an Inverse Transparency solution an ordering of possible usage reports by reading ease is arguably sufficient. It is worth noting, that Pitler et. al. [29] found that readability predictors

behave differently depending on whether they are ranking texts, or predicting their readability, though no guesses are made as to why this is the case.

B. Cognitively inspired measures

In response to linguists' complaints that traditional readability measures relied only on surface-level features, researchers began integrating higher-level semantic and discourse features. This includes characteristics of the text like its average parse tree height or the average number of verb phrases per sentence, as well as higher-level properties like discourse relations [29]. Intuitively, this resembles the traditional readability formulas, in that the importance of each fiture is 'fit' to given readability measurements. However, advanced machine learning frameworks can more effectively assess the interplay between the many variables that may affect a texts readability, and can express sophisticated 'decision spaces' that capture this [20]. Two important variables in this are feature selection and choice of learning framework. However, in a review of more recent approaches to readability measurement, Collins-Thompson [20] reports that it is typically found that the choice of features can lead to more significant performance gains than the choice of learning framework. This suggests that the choice of learning framework shouldn't necessarily be made because of accuracy, but on demands of the problem itself. For instance, if confidence estimates are required, probabilistic learning frameworks may be beneficial [20]. The most commonly used framework seems to be the Support Vector Machine (SVM), for regression or classification tasks [29], [30], [31]. Other methods employed include k-Nearest-Neighbor [32], and multiple linear or logistic regression [33].

1) Features for readability assessment: Even in the 20th century, the focus of many readability studies was to find the most constrained feature set that best explained readability. To this end, feature ablation studies were performed to gauge the impact of a single feature on the effectiveness of a readability predictor. In certain cases, removing a feature or group of features led to improved assessment by the predictor [29], [30], [34]. While classical features like sentence length or vocabulary difficulty were commonly found to be strong predictors of readability, they invariably found that including non-classical features led to improved accuracy [29], [30], [31]. Franccois et. al. [30] note that certain features for readability assessment may behave in a non-linear manner, and can therefore not be effectively captured by linear regression techniques. Some of these non-classical features and techniques to measure them are discussed briefly in the following.

A concept in the forefront of discussions about higher level determinants of readability is cohesion, approximated by features like the number of propositions or some measure of sentence overlap. This captured the notion that a text is not merely a collection of sentences in an arbitrary order, but that the readability of a sentence may depend on context provided in surrounding text.

Similarly, Latent Semantic Analysis (LSA) is a technique that allows for the analysis of semantic relatedness between texts, or between different segments of a text [35]. By representing text content (maybe individual words) as vectors in semantic space, the model is imbued with a 'knowledge' of the contexts in which this word normally appears. For example, the words 'music' and 'guitar' are likely to appear frequently in the same context, thus they would have a strong semantic relationship [20]. This method has also been used to match students with texts most appropriate to their reading level, by estimating the conceptual overlap between their personal topic knowledge and text content [36].

Statistical Language Modeling (SLM) provided a powerful alternative to the word frequency lists of the traditional readability measures. A language model is trained on a corpus of representative texts, and then predicts the likelihood of certain words or phrases occurring in this text as a proxy for reader familiarity. Si and Callan [22] use a unigram language model, and find this to be a stronger predictor of readability than sentence length. Later studies [29], [31] compute similar language models and use them as features in analysis with SVMs. Ostendorf et. al. [31] point out notable accuracy improvements with a more complex trigram language model, and found that combing SLMs with other features using an SVM produced the best results.

2) Coh-Metrix: There are currently not many openly available tools for measuring the readability of a text, and even fewer that were validated against other measures of readability. The most substantial of the tools that remain is the Coh-Metrix [37], which compiles measures of over 200 features related to cohesion, language and readability to analyse a text, and to return specific measures requested by the user. Some studies (e.g. [32]) used the Coh-Metrix to identify specific features with which different classes of texts could be distinguished. McNamara et. al. [38] found Coh-Metrix was successful in its primary purpose of distinguishing between highly and less cohesive texts. While the Coh-Metrix does not explicitly measure readability, it measures many of its determinants, and can thus be used for this purpose.

3) Future directions: The value of these computational approaches lies not only in their superior accuracy, but also in their potential for further improvement. The task of personalised text retrieval or text simplification is particularly relevant to making texts readable for all persons, rather than the average person. In the context of an Inverse Transparency solution, data usage reports might be generated based on prior knowledge of the employees reading proficiency profile.

Moreover, most studies focus on evaluating longer pieces of text, like books or documents. Less work has been done on local readability measures, for example measuring the readability of a single sentence. Dell' Orletta et. al. [24] found this to be a more difficult problem than general readability, and that certain features that are useful in assessing document readability are less useful in assessing sentence readability. Pilan et. al. [23] achieved 71% accuracy in the classification of second-language sentences. Because usage reports may be short, such research is particularly relevant to Inverse Transparency solutions.

Though little reference to this is made in the readability literature, the production of texts at a certain reading-level would be useful for Inverse Transparency solutions. Text-production is a standard problem in Natural Language Processing, thus it is possible that the NLP models used in these computational approaches can somehow be extended to serve this purpose. Related work in text simplification is indicative of this potential [33].

C. Evaluation

Computational approaches to readability assessment are flexible, in that they can be trained on text corpora specific to a certain subject area, for instance a body of data usage reports ranked in understandability by employees. Moreover, they can adapt over time, as their environment changes and they are provided more information. For example, after reviewing a data usage report, employees may be asked to judge the understandability of the report, allowing the system to update its estimate of both the employees reading proficiency and the understandability of the report.

Researchers consistently found that including more, and especially higher-level characteristics of the text increased prediction accuracy. This complements linguistic theories of text comprehension, suggesting a stronger construct validity of these approaches.

Computational approaches can be expected to grow more powerful with improved methods in artificial intelligence as a whole and more effective application of these methods to the problem of readability assessment. Perhaps the biggest challenge in developing such a solution is the assembly of gold-standard training corpora, but this is becoming easier with NLP tools like TAACO and crowdsourcing platforms like the Amazon Mechanical Turk.

IV. User-centric measures

The shortcomings of text-based readability measures implied the need for measures that somehow took into account characteristics of the reader that might impact their understanding of a given text. This led to more user-centric approaches towards estimating understandability, that will be discussed in this section and can be divided into quantitative and qualitative approaches. Because they are necessarily more difficult to employ, these methods are usually used to accurately measure understandability, from which text-based readability predictors may be validated or created, or to provide in-depth feedback about what made a text difficult to understand. It is worth discussing them for this reason.

A. Quantitative approaches

Perhaps the most well-known way of estimating how understandable readers find a given text is to have them take a standard reading comprehension test, in which they answer **multiple-choice questions** about the text. In general, results from such multiple-choice tests are what other measures are validated against, though this has been called into question. These tests are difficult to administer, because they are difficult to construct objectively and may take long for test participants to complete. Kobayashi et. al. [39] found that text organization and test format had a significant impact on test results. Moreover, the way a test subject reads a passage under test-conditions differs substantially from the way he would under non-testing conditions [40].

Many of the traditional readability formulas of the 20th century were validated against cloze scores on the same passage. The **cloze test** is administered by deleting words from a given passage, and letting the reader try to infer these words from the context [41]. An estimate of how well the reader understands the passage is derived from the percentage of words the reader can guess correctly [42]. This test is still commonly used in assessments of language skill, especially foreign language, like the Aptis test or the PTE Academic test. Many variants exist, the simplest of which deletes every n-th word without taking into account the significance of the deleted word within its context. However, it is not clear what exactly the cloze test measures, and how closely this corresponds to comprehension [43]. It may, for instance, simply measure textual redundancy.

Multiple variants of the cloze test were developed using rational decision procedures rather than 'random' deletion of every n-th word. Bachman [44] argued that such cloze-tests may be able to measure higher-level syntactic and discourse features of the text. In a later study, he argues that these tests can be designed specifically to test a range of abilities, and develops criteria for this [45]. More recently, Greene [46] found that cloze tests are useful for assessing understanding of discourse at a university level, if deletions are concentrated on the connections between sentences. A similarity can be observed between the assumptions underlying statistical language modelling, and the word-deletion and inference approach used in the cloze procedure, in that textual difficulty is approximated by some concept of reader familiarity with the text.

B. Qualitative approaches

In a critique of traditional readability measures, Redish [17] argues that "the methodology which allows you to deal with the complexity of a real document for real users is **usability testing**". In this, representative users are allowed to work with a document while trained observers take notes. This allows for useful feedback about the entire document, not limited to a small set of specific features, therefore it becomes especially relevant when assessing non-textual determinants of understandability, like effective use of diagrams. The **think-aloud protocol** is a comparable approach, in which subjects are asked to voice their thoughts as they work through a document. These monologues are transcribed, and provided as feedback to authors [47]. However, as DuBay [2] points out, on its own these approaches are incapable of matching the reading levels of test subjects to those of the target audience.

Readability assessment is a task that is relatively easy for a human to do, but difficult for machine intelligence. Thus, it is a good fit for **crowdsourcing** services like the Amazon Mechanical Turk (AMT). 'Requesters' post Human Intelligence Tasks (HITs) for evaluation by a diverse population of 'workers',

who are paid small amounts of money upon completion of the HIT. De Clerq et. al. [48] used crowdsourcing to obtain readability judgements, and found that the non-expert judgements produced by the workers were of comparable quality to those of experts. As outlined in section III, they asked participants to chose the more readable between two texts in order to produce a ranking. Complementing this, Chen et. al. [49] developed a statistical model to reduce the number of 'worker' judgements necessary to achieve a given level of ranking accuracy. AMT also allows for filtering of target demographics through 'qualifications', that a worker must fulfil before they are allowed to complete the HIT. While these are likely not sufficient for effectively screening workers that fit a certain reader profile, a portion of the HIT could be dedicated to this, for example by including a cloze test or a questionnaire. Consequently, Kittur et. al. [50] found that the success of crowdsourcing tasks can be very sensitive to task and interface design. Overall, crowdsourcing presents a time- and cost-effective alternative for estimating understandability.

Expert-judgement is a final method of gaining readability judgements and comprehensive feedback as to what could be improved. However, studies have described a "knowledge effect in writing", observing that readers with more extensive knowledge of a topic poorly judged the understandability of this text to a layperson [51]. This suggests that expert judgements may be only weakly representative of judgements by the target audience, which are more important. Therefore, expert judgements should be used in conjunction with other methods described here.

C. Case study: PEMAT

The **PEMAT** (Patient Education Materials Assessment Tool) provides an interesting case study for the development of an Inverse transparency solution. The PEMAT was designed based on consumer testing and on the feedback of a panel of health literacy experts, to assess the understandability and actionability of patient education materials [52]. Scores are calculated from user survey responses. The test consists of 26 prompts about the material, partitioned into understandability and actionability prompts, for each of which the user must choose to agree (1) or disagree (0) [2]. These prompts touch on properties like layout, word choice, use of visual aids, understandability, and more. In the defining study, Shoemaker et. al. were unable to find a relationship between PEMAT understandability scores and consumer comprehension scores. However, they found a significant positive correlation between PEMAT results for actionability and comprehension scores for all media types tested. Beyond the results published in this study, little can be found in the literature about the validity of this measure.

Only one of the criteria in the PEMAT is domain-specific to healthcare materials, related to word choice. This could easily be adjusted to create assessment tools for other domains. These tests can also be administered more quickly and objectively than usability testing, and produce a numeric score.

D. Evaluation

User-centric measures are more difficult to apply than simple text-based measures. However, user-centric measures are broadly considered to be more valid indicators of understandability, because they measure an interaction between a user and a text. As a result, they have been used commonly to validate text-based measures and to test language proficiency, and they may become relevant in building text corpora labelled with readability judgements. Both quantitative and qualitative user-centric measures provide the kind of feedback that text-based measures are unable to, and the two classes of readability measures should be seen as complementing each other, neither particularly useful on its own.

V. Linguistics and presentational factors

A. Linguistics

Of course there exist many guidelines and heuristics for writing more understandable texts, and such information may be found readily on the internet [3]. However, as pointed out by Klare [11], there is little agreement among these heuristics, making it difficult to determine exactly what steps should be taken in writing understandable text. Results in readability research, both on the computational and on the purely linguistic front, complement this, in that they suggest that the understandability of a text may depend significantly on the person reading it. It follows that there can be no one-size-fits-all approach to writing simple text.

This is reflected in the number of (in some cases competing) approaches to linguistics. Cognitive-linguistic theories analyse the interaction between cognition and language, built on the idea that linguistic constructs reflect our perception of the world. In comparison, psycholinguistic approaches are not interested in the structure of language, but rather on the mental processes taking place during language production and comprehension, as well as during language acquisition. Corpus linguistics assume a statistical approach to the study of language, making few assumptions about its structure. All of these approaches play an important advisory role for readability research, and many studies integrate results from more than one of them. For example, language models are commonly used in conjunction with a feature-based text analysis [31], applying findings from both cognitive and corpus linguistics. Based on psycholinguistic research, Kidwell et. al. [54], build a model that predicts a words most likely acquisition age as a proxy for word familiarity. It does this by analysing authentic texts on the Web.

However, beyond the empirical research already described in this literature review, little can be found that is directly concerned with the understandability of prose. In a paper on the value of readability formulas for the development

[2]https://www.ahrq.gov/ncepcr/tools/self-mgmt/pemat.html

[3]https://centerforplainlanguage.org/learning-training/five-steps-plain-language/

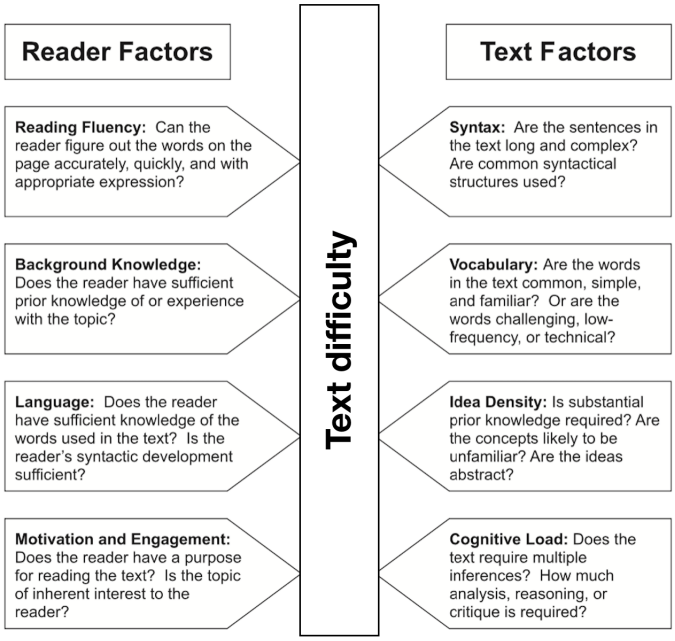

Fig. 1. Factors contributing to text difficulty [53]

and adaptation of short test items, Oakland [53] proposes a grouping of determinants of understandability, emphasising the importance of considering 'Reader Factors', as depicted in Figure 1.

The Karlsruhe comprehensibility concept (KCC) is one attempt at a framework of the understandability of prose, building on the earlier Hamburg comprehensibility concepts [55]. In the KCC Göpferich defines 6 dimensions of comprehensibility that are roughly in line with feature groups shown in Figure 1, drawing on research from cognitive science and educational psychology. These are (with simplified success requirements):

1) **Concision**: the text contains the minimum amount of information required to fulfill its communicative function.
2) **Correctness** (consistency): the text does not contain mistakes, whether conceptual, grammatical, logical…
3) **Motivation**: the text arouses motivation in its reader.
4) **Structure**: the text is cohesive, that is the reader is forced only to make an appropriate number of inferences.
5) **Simplicity**: the lexis (word choice) and syntax are simple for the target audience of the text.
6) **Perceptibility**: the text can be perceived easily, through effective layout, typography, colors, …

Göpferich defines text quality as the extent to which the text fulfils its communicative function. She points out that an instructional text may be easily comprehensible, but not very usable, and thus have low text quality, and that her KCC may be used not only to evaluate comprehensibility, but also the broader notion of text quality. I was unable to find a similar analysis or an application of such a model for shorter texts or sentences, though the KCC could reasonably be expected to be applicable to such texts too.

While both Oakland and Göpferich point out the factors that determine understandability, neither makes judgements as to the relative importance of each factor. This is the subject of study of computational approaches described in section 2, and may vary widely depending on the target group and the type of text. For a comprehensive comparison of such features in automated readability assessment, see [56].

B. Presentational Factors

The 6th dimension proposed in the KCC covers text-content-external factors, like legibility or use of diagrams, that may contribute to understandability. Because most of the research discussed in this literature review focused on characteristics of the content, some relevant findings will be described briefly in the following.

As one might expect, typography and line spacing have an influence on reading speed. Studies have suggested that larger fonts and average line spacing lead to optimal readability [57]. Moreover, font and background colors should be chosen to maximise contrast between the two [58]. Hill [59] found that font, word style and concrete color combination had little to no impact on readability.

Tinker [60] found that text that uses mixed upper- and lower-case letters was easier to read. Moreover, proportionally spaced text can be read more quickly than non-proportionally spaced texts (in which every letter occupies the same amount of space) [61].

Loman et. al. [62] found that including signalling techniques, like underling or bolding important text segments, could modify students reading strategies to effectively increase the understandability of expository prose.

Vahabi [63] suggests that illustrations can aid understanding. In an Inverse Transparency solution, this might promote the use of diagrams.

However, these are all not particularly significant findings, with studies often in contexts of reading disabilities or language acquisition, where relevance to the specific needs of an Inverse Transparency solution is not clear. As a result, the choice of presentational factors may benefit greatly from user feedback, for example by way of usability-testing.

VI. Survey

A. Design

To complement this literature review, I posted a short survey on the Amazon Mechanical Turk (AMT). This survey had two purposes:

1) To evaluate AMT as a crowdsourcing tool (see section IV)
2) To suggest how people respond to texts that narrowly resemble the usage reports produced in an Inverse Transparency solution.

To refrain from making too many assumptions about usage reports in an inverse transparency solution, texts were chosen according to two characteristics: they were 2-3 sentences

long, and concerned with technical subjects. To this end, passages were taken from the sections on 'Informational texts: Mathematics, Science and technical subjects' from the Common Core State Standards Appendix B, a collection of short passages organised by textual complexity. These ratings were informed by expert judgements, feedback from the public and other existing standards [4]. Passages in this Text corpus are generally longer, so I took excerpts that I judged to be representative of the difficulty of the full passage. The passages were chosen such that survey participants evaluated 3 texts from each of the Grade levels 6-8, 9-10, 11-CCR (college- and career-ready). Two additional passages were included as validation measures: one very simple (grade-level 3-4), and one containing multiple nonsensical words (see below). It was assumed that, if they actually read the passage (rather than selecting answers at random), they would invariably select the expected answer to these validation questions. Survey participants were then asked to evaluate their understanding of this passage, by indicating whether they felt the passage required further clarification or they felt completely confident in their understanding of the passage. They were instructed to select the former option whenever they felt any doubt. This was intended to imitate a situation in which an employee reads a usage report, and must decide whether or not to request further clarification. The test layout, as well as some of the test passages may be found in the Appendix.

B. Results

Participants were paid $1.00 if I approved their submissions, which I did only when they 'correctly' responded to both validation questions. Within an hour, 86 people had submitted survey responses, 33 of which were approved. The average time to complete the survey was 7 minutes, and the results are summarised in Figure 2. The large number of rejected responses can be attributed to poorly constructed validation questions. It appears that even participants who thoughtfully responded to each passage selected the 'wrong' answer, many of whom contacted me afterwards to justify their choices. It is also possible that many of the participants did not properly read the instructions, aiming to complete the survey as quickly as possible.

The goal in constructing the difficult validation passage was to change the original passage only in minor ways (see bold text), such that it is not immediately obvious that this was to be used for validation. Unfortunately, some survey participants were able to infer the original meaning of the passage and then decided simply to ignore the words that did not fit. Some respondents recognised that they didn't entirely understand what these words meant but could guess, and that this lack of understanding of parts of the passage did not impact their understanding of the passage as a whole. This was enough for 54.7% of respondents to feel confident in their understanding of the passage (adapted from [5]):

> It is spring in McAllen, Texas. The morning sun **bathes in blue light**. The streets are lined with **palm falls** and **curious derivations**. McAllen is in Hidalgo County, who have the lowest household **flunders** in the country, but it's an **order town**, and a thriving **inter-trade zone** has kept the **cargo rate** below ten percent. McAllen calls itself the Square Dance Capital of the World. "Lonesome Dove" was set around here.

Other participants pointed out that understanding is an entirely subjective thing, and that rejections based on these validation passages are unwarranted. For the easy validation passage, 72.1% respondents indicated they were confident of their understanding of it, lower than anticipated. However, this was the passage respondents judged the most favorably.

Because of the apparent weakness of the validation prompts, I decided to include all responses that took longer than 2 minutes to complete in the results. Others were removed, as it is (in my eyes) not possible to read all the passages and respond thoughtfully in less time.

Prior to the survey, I expected that having participants choose between two ratings of understandability would simplify their task. Based on e-mail feedback, it may have complicated matters, as many felt it was unclear where exactly to draw the line between fully understandable and requiring further clarification. This may be related to the notion of levels of understanding. For instance, understanding a word might mean that you know situations in which to use it, or it might mean that you are aware of its origin and know its precise definition. I would hypothesize that this confusion would not be as much of a problem in an Inverse Transparency solution, where it is in an employees own interest to make sure he has a complete understanding of the usage report, and the employee is the judge of what constitutes complete understanding.

Although there is a slight dip in the perceived understandability of passages as their grade level increases, passages ranked Grade level 9-10 were less well understood than the supposedly more difficult 11-CCR passages. A moderate correlation (R=-0.78) can be observed between Grade-level groupings and perceived understandability. However, within each grade level there is a large amount of variance in these ratings. The text that was rated by far the least favorably was a passage rated at the 9-10th grade level, and one of the passages at grade level 6-8 was perceived as requiring clarification by more survey participants than two of the 11-CCR passages. Interestingly, the two most difficult passages share one characteristic: they both discuss purely mathematical concepts, whereas the other passages concentrate on scientific and technical subjects. This suggests that the topic may have more of an impact on the perceived difficulty of short excerpts than features analysed in rating the full passage they were taken from. Because the validation questions did not fulfil their function, it was not possible to effectively filter responses that did not adhere to the survey instructions, which may have added noise to the data.

Overall, AMT proved to be a quick and cost-effective way

[4] http://www.corestandards.org/about-the-standards/development-process/

[5] A. Gawande, "The cost conundrum," *The New Yorker*, vol. 1, pp. 36-44, 2009

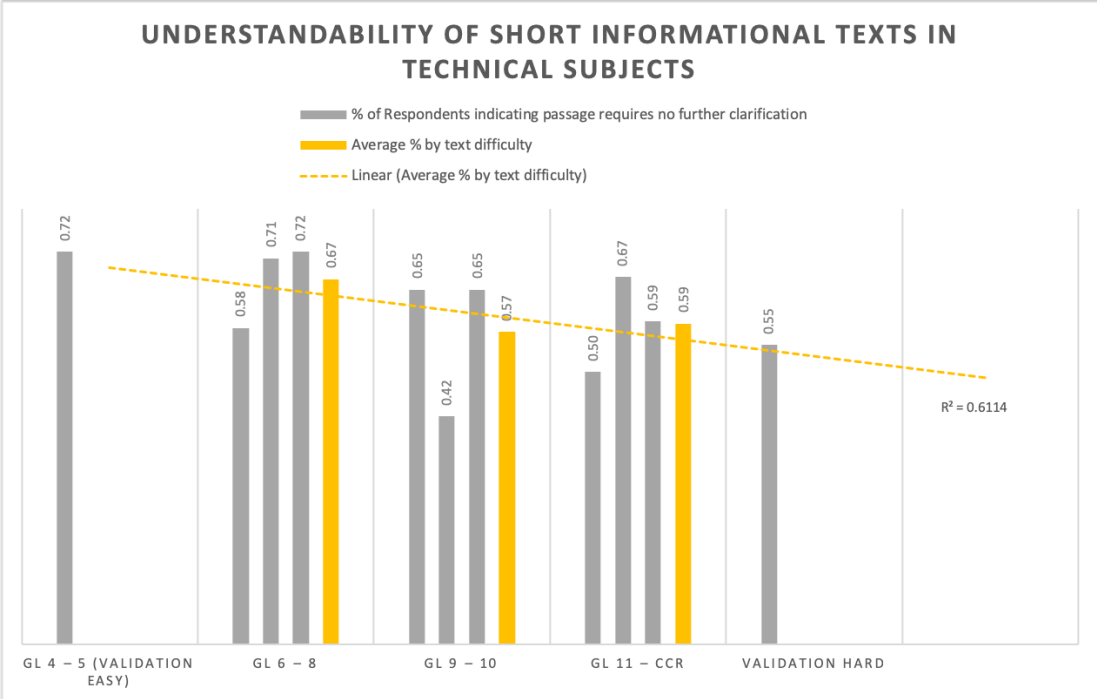

Fig. 2. Understandability of short informational texts in technical subjects

of gathering understandability judgements. However, because of the small amounts paid for the completion of HITs, AMT workers will try to complete these tasks as quickly as they can. This must be taken into account when designing HITs. Survey results have little explanatory power, but serve to indicate some of the difficulties that may be encountered in the development of an Inverse Transparency solution.

VII. Related works

DuBay provides a comprehensive review of traditional readability research in his book *Principles of Readability* [2]. Zamanian and Heydari provide a more concise, and more recent review of the same literature, pointing out both the advantages and disadvantages of traditional readability measures [64].

Schriver gives a general overview of text-quality assessment methods, describing a "continuum from text-focused to reader-focused methods", as well as introducing the problem of text quality assessment from a linguistic perspective [65].

In a 2014 review, Collins-Thompson discusses more recent developments in computational readability assessment, as well as indicating promising directions for future research [20]. Benjamin [21] reviews recent traditional-style formulas, like the lexile measure, as well as more complex computational approaches like the ones outlined in section III. She discusses linguistic theory underlying these computational approaches, and provides general recommendations for the use of readability assessment tools in an educational context.

Graesser et. al. [66] provide an overview of psycholinguistic frameworks for discourse comprehension. Walter Kintsch proposed the Construction-Integration model of discourse comprehension in a frequently cited paper published in 1988 [67].

Although this literature review provides an overview of some of the possible methods for determining the readability of a text, it is far from complete. The topic of what makes one text easier to understand than another is a complex one, that much research has gone into, in a broad array of disciplines. Any of the related works mentioned in this section delve far more thoroughly into certain fractions of this research, and may thus be used to attain a more complete impression of the developments readability research has undergone in the last 50 years.

VIII. Conclusion

The aim of this literature review was to summarise results in readability and understandability research, and to evaluate them for their usefulness in the development of an Inverse Transparency solution. In this, it was not only important how reliable their understandability measurements might be, but also:

- how easily measurements might be collected on a large scale,
- whether measurements provide useful feedback,
- whether measures have evaluative or predictive power, and
- whether measures were sufficiently sensitive to changes in the reader

In this, it was also important how reliably measurements of understandability can be performed on shorter texts. Because of the subjectivity inherent in the notion of understandability, the task of accurately estimating or predicting readability remains a lively topic of research. This review provides a broad

overview of the field, at the cost of likely omitting results that might prove particularly relevant in the future.

It should be clear that although the traditional formulas played an important role in inspiring a growing body of research, they face many limitations that make them an unlikely candidate for readability assessment within an Inverse Transparency solution. Computational approaches initially built on the idea that understandability can be predicted based solely on features of the text, but offer more reliable measurements based on hundreds of features of the text, rather than just 2 or 3. These require training on large, annotated corpora of text, which may not be easy to come by. Natural Language Processing techniques have made text annotation easier in the past decade, and crowdsourcing may provide a cost- and time-efficient way of determining how a specific population judges the readability of a text. User-centric measures may serve to validate the more easily applicable text-based measures, and are likely to play a role in accurately assessing how well a person understands a text even without regard to their personal impression. Historically, results from other linguistic disciplines have informed the approaches described in this literature review, though little of it is concerned directly with the issue of readability, especially its measurement.

REFERENCES

[1] H. D. Kitson, *The mind of the buyer: A psychology of selling*. Macmillan, 1921, vol. 21549.

[2] W. H. DuBay, "The Principles of Readability." *Online Submission*, 2004.

[3] R. Flesch, "A new readability yardstick." *Journal of applied psychology*, vol. 32, no. 3, p. 221, 1948.

[4] R. Lehavy, F. Li, and K. Merkley, "The effect of annual report readability on analyst following and the properties of their earnings forecasts," *The Accounting Review*, vol. 86, no. 3, pp. 1087–1115, 2011.

[5] S. B. Bonsall IV, A. J. Leone, B. P. Miller, and K. Rennekamp, "A plain english measure of financial reporting readability," *Journal of Accounting and Economics*, vol. 63, no. 2-3, pp. 329–357, 2017.

[6] M. K. Paasche-Orlow, H. A. Taylor, and F. L. Brancati, "Readability standards for informed-consent forms as compared with actual readability," *New England journal of medicine*, vol. 348, no. 8, pp. 721–726, 2003.

[7] P. M. Linsley and P. J. Shrives, "Examining risk reporting in UK public companies," *The Journal of Risk Finance*, vol. 6, no. 4, pp. 292–305, 2005.

[8] J. S. Chall and E. Dale, *Readability revisited: The new Dale-Chall readability formula*. Brookline Books, 1995.

[9] D. Smith, A. Stenner, I. Horabin, and M. Smith, "The lexile scale in theory and practice: Final report for NIH grant HD-19448," in *meeting of the International Reading Association, New Orleans*, 1989.

[10] A. Bailin and A. Grafstein, "The linguistic assumptions underlying readability formulae: A critique," *Language & Communication*, vol. 21, no. 3, pp. 285–301, 2001.

[11] G. R. Klare, "A second look at the validity of readability formulas," *Journal of reading behavior*, vol. 8, no. 2, pp. 129–152, 1976.

[12] K. A. Schriver, "Readability formulas in the new millennium: what's the use?" *ACM Journal of Computer Documentation (JCD)*, vol. 24, no. 3, pp. 138–140, 2000.

[13] G. R. Klare *et al.*, *Measurement of readability*. Iowa State University Press, 1963.

[14] R. P. Charrow and V. R. Charrow, "Making legal language understandable: A psycholinguistic study of jury instructions," *Columbia law review*, vol. 79, no. 7, pp. 1306–1374, 1979.

[15] G. R. Klare, "Readable computer documentation," *ACM Journal of Computer Documentation (JCD)*, vol. 24, no. 3, pp. 148–168, 2000.

[16] B. Bruce, A. Rubin, and K. Starr, "Why readability formulas fail," *IEEE Transactions on Professional Communication*, no. 1, pp. 50–52, 1981.

[17] J. Redish, "Readability formulas have even more limitations than Klare discusses," *ACM Journal of Computer Documentation (JCD)*, vol. 24, no. 3, pp. 132–137, 2000.

[18] T. Loughran and B. McDonald, "Measuring readability in financial disclosures," *The Journal of Finance*, vol. 69, no. 4, pp. 1643–1671, 2014.

[19] D. S. McNamara and W. Kintsch, "Learning from texts: Effects of prior knowledge and text coherence," *Discourse processes*, vol. 22, no. 3, pp. 247–288, 1996.

[20] K. Collins-Thompson, "Computational assessment of text readability: A survey of current and future research," *ITL-International Journal of Applied Linguistics*, vol. 165, no. 2, pp. 97–135, 2014.

[21] R. G. Benjamin, "Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty," *Educational Psychology Review*, vol. 24, no. 1, pp. 63–88, 2012.

[22] L. Si and J. Callan, "A statistical model for scientific readability," in *Proceedings of the Conference on Information and Knowledge Management*, vol. 1, 2001, pp. 574–576.

[23] I. Pilán, E. Volodina, and R. Johansson, "Rule-based and machine learning approaches for second language sentence-level readability," in *Proceedings of the 9th workshop on innovative use of NLP for building educational applications*, 2014, pp. 174–184.

[24] F. Dell'Orletta, M. Wieling, G. Venturi, A. Cimino, and S. Montemagni, "Assessing the readability of sentences: which corpora and features?" in *Proceedings of the 9th workshop on innovative use of NLP for building educational applications*, 2014, pp. 163–173.

[25] S. A. Crossley, S. Skalicky, M. Dascalu, D. S. McNamara, and K. Kyle, "Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas," *Discourse Processes*, vol. 54, no. 5-6, pp. 340–359, 2017.

[26] S. A. Crossley, K. Kyle, and D. S. McNamara, "The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion," *Behavior research methods*, vol. 48, no. 4, pp. 1227–1237, 2016.

[27] K. Kyle, S. Crossley, and C. Berger, "The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0," *Behavior research methods*, vol. 50, no. 3, pp. 1030–1046, 2018.

[28] K. Tanaka-Ishii, S. Tezuka, and H. Terada, "Sorting texts by readability," *Computational Linguistics*, vol. 36, no. 2, pp. 203–227, 2010.

[29] E. Pitler and A. Nenkova, "Revisiting readability: A unified framework for predicting text quality," in *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2008, pp. 186–195.

[30] T. François and E. Miltsakaki, "Do NLP and machine learning improve traditional readability formulas?" in *Proceedings of the 1st Workshop on Predicting and Improving Text Readability for target reader populations*. Association for Computational Linguistics, 2012, pp. 49–57.

[31] S. E. Schwarm and M. Ostendorf, "Reading level assessment using support vector machines and statistical language models," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2005, pp. 523–530.

[32] S. A. Crossley, J. Greenfield, and D. S. McNamara, "Assessing text readability using cognitively based indices," *Tesol Quarterly*, vol. 42, no. 3, pp. 475–493, 2008.

[33] S. Aluisio, L. Specia, C. Gasperin, and C. Scarton, "Readability assessment for text simplification," in *Proceedings of the 5th Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, 2010, pp. 1–9.

[34] M. Heilman, K. Collins-Thompson, J. Callan, and M. Eskenazi, "Combining lexical and grammatical features to improve readability measures for first and second language texts," in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, 2007, pp. 460–467.

[35] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse processes*, vol. 25, no. 2-3, pp. 259–284, 1998.

[36] K. Wolfe-Quintero, S. Inagaki, and H.-Y. Kim, *Second language development in writing: Measures of fluency, accuracy, & complexity*. University of Hawaii Press, 1998, no. 17.

[37] A. C. Graesser, D. S. McNamara, M. M. Louwerse, and Z. Cai, "Coh-metrix: Analysis of text on cohesion and language," *Behavior research methods, instruments, & computers*, vol. 36, no. 2, pp. 193–202, 2004.

[38] D. S. McNamara, M. M. Louwerse, P. M. McCarthy, and A. C. Graesser, "Coh-metrix: Capturing linguistic features of cohesion," *Discourse Processes*, vol. 47, no. 4, pp. 292–330, 2010.
[39] H. Ogawa, H. Kobayashi, N. Matsuda, T. Hirashima, and H. Taki, "Knowledge externalization based on differences of solutions for automatic generation of multiple-choice question," in *Proceedings of the 19th International Conference on Computers in Education*, 2011, pp. 271–278.
[40] A. A. Rupp, T. Ferne, and H. Choi, "How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective," *Language testing*, vol. 23, no. 4, pp. 441–474, 2006.
[41] W. L. Taylor, "Cloze procedure: A new tool for measuring readability," *Journalism Bulletin*, vol. 30, no. 4, pp. 415–433, 1953.
[42] J. R. Bormuth, "Cloze test readability: Criterion reference scores," *Journal of educational measurement*, vol. 5, no. 3, pp. 189–196, 1968.
[43] J. W. Oller Jr, "Cloze tests of second language proficiency and what they measure," *Language learning*, vol. 23, no. 1, pp. 105–118, 1973.
[44] L. F. Bachman, "The trait structure of cloze test scores," *Tesol Quarterly*, vol. 16, no. 1, pp. 61–70, 1982.
[45] ——, "Performance on cloze tests with fixed-ratio and rational deletions," *Tesol Quarterly*, vol. 19, no. 3, pp. 535–556, 1985.
[46] B. Greene, "Testing reading comprehension of theoretical discourse with cloze," *Journal of Research in Reading*, vol. 24, no. 1, pp. 82–98, 2001.
[47] R. Jääskeläinen, "Think-aloud protocol," *Handbook of translation studies*, vol. 1, pp. 371–374, 2010.
[48] O. De Clercq, V. Hoste, B. Desmet, P. Van Oosten, M. De Cock, and L. Macken, "Using the crowd for readability prediction," *Natural Language Engineering*, vol. 20, no. 3, pp. 293–325, 2014.
[49] C. C. Loy, K. Chen, S. Gong, and T. Xiang, "Crowd counting and profiling: Methodology and evaluation," in *Modeling, simulation and visual analysis of crowds*. Springer, 2013, pp. 347–382.
[50] A. Kittur, E. H. Chi, and B. Suh, "Crowdsourcing user studies with Mechanical Turk," in *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2008, pp. 453–456.
[51] J. R. Hayes and D. Bajzek, "Understanding and reducing the knowledge effect: Implications for writers," *Written Communication*, vol. 25, no. 1, pp. 104–118, 2008.
[52] S. J. Shoemaker, M. S. Wolf, and C. Brach, "Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information," *Patient education and counseling*, vol. 96, no. 3, pp. 395–403, 2014.
[53] T. Oakland and H. B. Lane, "Language, reading, and readability formulas: Implications for developing and adapting tests," *International Journal of Testing*, vol. 4, no. 3, pp. 239–252, 2004.
[54] P. Kidwell, G. Lebanon, and K. Collins-Thompson, "Statistical estimation of word acquisition with application to readability prediction," *Journal of the American Statistical Association*, vol. 106, no. 493, pp. 21–30, 2011.
[55] S. Göpferich, "Comprehensibility assessment using the Karlsruhe comprehensibility concept," *The Journal of Specialised Translation*, vol. 11, no. 2009, pp. 31–52, 2009.
[56] L. Feng, M. Jansche, M. Huenerfauth, and N. Elhadad, "A comparison of features for automatic readability assessment," in *Proceedings of the 23rd international conference on computational linguistics*. Association for Computational Linguistics, 2010, pp. 276–284.
[57] L. Rello, M. Pielot, and M.-C. Marcos, "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, 2016, pp. 3637–3648.
[58] R. Hall and P. Hanna, "The effect of web page text-background color combinations on retention and perceived readability, aesthetics and behavioral intention," *Proceedings of the Americas Conference on Information Systems*, p. 277, 2003.
[59] A. Hill and L. Scharff, "Readability of websites with various foreground/background color combinations, font types and word styles," in *Proceedings of 11th National Conference in Undergraduate Research*, vol. 2, 1997, pp. 742–746.
[60] M. A. Tinker, "Legibility of print for children in the upper grades," *Optometry and Vision Science*, vol. 40, no. 10, pp. 614–621, 1963.
[61] M. Helander and B. A. Rupp, "An overview of standards and guidelines for visual display terminals," *Applied ergonomics*, vol. 15, no. 3, pp. 185–195, 1984.
[62] N. L. Loman and R. E. Mayer, "Signaling techniques that increase the understandability of expository prose," *Journal of Educational psychology*, vol. 75, no. 3, p. 402, 1983.
[63] M. Vahabi and L. Ferris, "Improving written patient education materials: a review of the evidence," *Health Education Journal*, vol. 54, no. 1, pp. 99–106, 1995.
[64] M. Zamanian and P. Heydari, "Readability of texts: State of the art." *Theory & Practice in Language Studies*, vol. 2, no. 1, 2012.
[65] K. A. Schriver, "Evaluating text quality: The continuum from text-focused to reader-focused methods," *IEEE Transactions on professional communication*, vol. 32, no. 4, pp. 238–255, 1989.
[66] A. C. Graesser, K. K. Millis, and R. A. Zwaan, "Discourse comprehension," *Annual review of psychology*, vol. 48, no. 1, pp. 163–189, 1997.
[67] W. Kintsch, "The role of knowledge in discourse comprehension: A construction-integration model." *Psychological review*, vol. 95, no. 2, p. 163, 1988.

APPENDIX

Grades 4-5[6] (Easy validation passage):

Life is easy for the Indians here in the Northwest near the great ocean. They feel rich. For them the world is bountiful: the rivers hold salmon and trout; the ocean is full of seals, whales, fish, and shellfish; the woods are swarming with game animals. And there are berries and nuts and wild roots to be gathered. They are not farmers. They don't need to farm.

Grades 6-8[7]: Geology is the scientific study of Earth. Geologists study the planet—its formation, its internal structure, its materials, its chemical and physical processes, and its history. Mountains, valleys, plains, sea floors, minerals, rocks, fossils, and the processes that create and destroy each of these are all the domain of the geologist. Geology is divided into two broad categories of study: physical geology and historical geology.

Grades 9-10[8]: The astrolabe (in Greek, "star reckoner") is a manual computing and observation device with myriad uses in astronomy, time keeping, surveying, navigation, and astrology. The principles behind the most common variety, the planispheric astrolabe, were first laid down in antiquity by the Greeks, who pioneered the notion of projecting three-dimensional images on flat surfaces.

Grades 11-CCR[9]: There is a fundamental property of numbers named after the Greek mathematician Archimedes which states that any number, no matter how huge, can be exceeded by adding together sufficiently many of any smaller number, no matter how tiny. Though obvious in principle, the consequences are sometimes resisted, as they were by the student of mine who maintained that human hair just didn't grow in miles per hour.

[6]Hakim, Joy. A History of US. Oxford: Oxford University Press, 2005. (2005) From Book 1: The First Americans, Prehistory to 1600; Chapter 7: "The Show-Offs"

[7]"Geology." U*X*L Encyclopedia of Science. Edited by Rob Nagel. Farmington Hills, Mich.: Gale Cengage Learning, 2007. (2007)

[8]Nicastro, Nicholas. Circumference: Eratosthenes and the Ancient Quest to Measure the Globe. New York: St. Martin's Press, 2008. (2008), from "The Astrolabe"

[9]Paulos, John Allen. Innumeracy: Mathematical Illiteracy and Its Consequences. New York: Vintage, 1988. (1988) From Chapter 1: "Examples and Principles", Archimedes and Practically Infinite Numbers

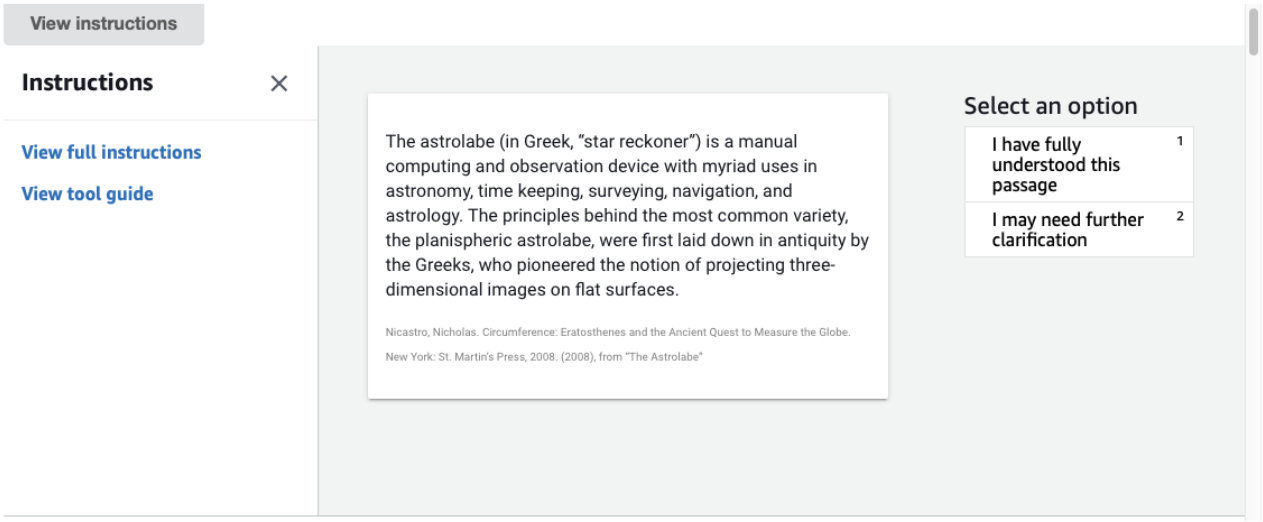

Fig. 3. Test layout; participants responded to 11 such questions