Comparison of Methods for Evaluating Complexity of Simplified Texts among Deaf and Hard-of-Hearing Adults at Different Literacy Levels

Oliver Alonzo
oa7652@rit.edu
Golisano College of Computing and Information Sciences
Rochester Institute of Technology (RIT)
Rochester, NY

Jessica Trussell
jwtnmp@rit.edu
National Technical Institute for the Deaf
Rochester Institute of Technology (RIT)
Rochester, NY

Becca Dingman
bad6955@rit.edu
School of Information
Rochester Institute of Technology (RIT)
Rochester, NY

Matt Huenerfauth
matt.huenerfauth@rit.edu
School of Information
Rochester Institute of Technology (RIT)
Rochester, NY

ABSTRACT

Research has explored using Automatic Text Simplification for reading assistance, with prior work identifying benefits and interests from Deaf and Hard-of-Hearing (DHH) adults. While the evaluation of these technologies remains a crucial aspect of research in the area, researchers lack guidance in terms of how to evaluate text complexity with DHH readers. Thus, in this work we conduct methodological research to evaluate metrics identified from prior work (including reading speed, comprehension questions, and subjective judgements of understandability and readability) in terms of their effectiveness for evaluating texts modified to be at various complexity levels with DHH adults at different literacy levels. Subjective metrics and low-linguistic-complexity comprehension questions distinguished certain text complexity levels with participants with lower literacy. Among participants with higher literacy, only subjective judgements of text readability distinguished certain text complexity levels. For all metrics, participants with higher literacy scored higher or provided more positive subjective judgements overall.

CCS CONCEPTS

• **Human-centered computing** → **Accessibility design and evaluation methods**; **Empirical studies in accessibility**.

KEYWORDS

Automatic Text Simplification, Methodological Research, Deaf and Hard-of-hearing, Accessibility

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445038

ACM Reference Format:
Oliver Alonzo, Jessica Trussell, Becca Dingman, and Matt Huenerfauth. 2021. Comparison of Methods for Evaluating Complexity of Simplified Texts among Deaf and Hard-of-Hearing Adults at Different Literacy Levels. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3411764.3445038

1 INTRODUCTION

Automatic Text Simplification (ATS) consists of a variety of computing techniques that aim to rewrite text to make it simpler to read or understand [36]. This is typically done by rewriting phrases or sentences in a text, which is known as syntactic simplification; by replacing individual words, which is called lexical simplification; or hybrid combinations of both [36]. As an emerging field, most of the work in this area is still experimental and most research papers focus on the evaluation of new methods of simplification [36, 37].

While ATS can have computing-based applications such as providing simplifications for a machine-translation pipeline, a growing body of research has focused on its use as an assistive technology for different user groups including people with dyslexia [28], people with aphasia [8], and language learners [4]. Considering that prior literacy research suggests that Deaf and Hard-of-Hearing (DHH) readers[1] have diverse literacy skills (e.g. studies have found fourth-grade reading levels among DHH high-school graduates [39]), researchers have investigated the use of ATS as reading assistance for DHH readers. Studies have found benefits from providing DHH adults with both syntactic and lexical approaches to simplification [3, 18], and also interest from a specific sub-group of DHH readers (computing professionals) in such technologies [2].

Evaluations of ATS output typically focus on three aspects of quality: grammaticality, meaning preservation and complexity [19, 33]. While the former two are typically evaluated with expert or native readers, the former is more commonly the focus of evaluations with target readers (i.e. those who may benefit from ATS). However, while the importance of the question of what are the

[1]Many DHH people prefer identity-first language. They do not perceive being DHH as a disability, but as a linguistic and cultural difference.

best manual methods of ATS evaluation has been raised in prior work [36, 37], there has been little to no methodological research on evaluating any of the different aspects of quality of the output of ATS systems with DHH users.

Metrics that are often used for evaluating complexity of ATS output with target readers, which include reading speed, comprehension questions, or subjective judgements of text, have not been validated with DHH readers. While this paper focuses on the evaluation of text complexity of simplified texts, we note that prior on evaluating errors in other linguistic technologies with DHH readers has found conflicting results when employing similar metrics. Prior work on the evaluation of errors in captioning technologies, for example, has found subjective judgements to be more effective than comprehension questions overall, but also that the literacy levels of participants affect the effectiveness of the metrics [5]. On the other hand, work on the evaluation of errors in American Sign Language (ASL) animations suggests that subjective judgements are no replacement for comprehension questions [12]. However, it remains unknown which type of metrics may be more effective when evaluating text complexity with DHH readers.

Thus, as a first step to reliably evaluate ATS systems with DHH adults, methodological research is necessary to determine what types of metrics are effective in measuring the complexity level of texts when evaluating them among DHH readers. So, in this paper we present an experimental study in which participants read news articles at different levels of complexity and answered a number of metrics for each article, which allowed us to determine which metrics were effective at distinguishing the texts complexity levels. Our results suggest that comprehension questions only work under a number of conditions (when they're used with participants with lower literacy, the questions themselves are written in language simple enough for participants to understand and the differences in complexity level of the texts are of several grade-levels). While all the subjective metrics tested were effective at distinguishing the complexity levels of the articles with lower literacy readers, only one of them was effective with higher literacy readers. The contributions of this work include:

(1) Methodological guidance based on empirical evidence in terms of metrics that are more effective for evaluating text complexity with DHH readers at different literacy levels, as well as the importance of reporting participants' literacy levels when employing any of the metrics included in our study in order to allow for comparison across studies.
(2) Considering that different user groups may face different challenges when it comes to text complexity [25], our methodological framework can be used by other researchers studying the use of ATS for other user groups to investigate the effectiveness of these metrics for evaluating text complexity with those user groups.

2 BACKGROUND AND RELATED WORK

In this section, we provide background on ATS and research focusing on its use as a reading assistance tool. Then, we discuss prior work on the evaluation of ATS, focusing on why each evaluation was conducted, with whom it was conducted, and which methods were used. This analysis of prior work reveals the lack of methodological guidance and lack of consensus as to how text complexity should be evaluated with users. Finally, we consider prior work among DHH adults, with a focus on evaluation of linguistic technology among these users. While some of this work was not specifically related to ATS, our analysis of this prior research provides guidance for how to conduct methodological research on linguistic technology evaluation with DHH users.

2.1 Automatic Text Simplification and Reading Assistance

ATS consists of computing techniques to rewrite text to improve its readability and/or understandability [36]. There are different approaches to ATS, including: at the syntactic level which focuses the simplification of phrases, sentences, or longer chunks of text; at the lexical level which focuses on replacing complex words with simpler synonyms; or hybrid combinations of both [36]. As an emerging field, there are still few widely used ATS systems, and evaluation of these technologies is still an important area of research with no consensus as to what are the best ways to evaluate them yet [37]. While ATS can have uses such as the initial steps of a machine translation pipeline, a growing body of research has investigated its use as an assistive technology, as a reading assistance tool.

Prior work on ATS as an assistive technology has focused on its use for a number of user groups including people with dyslexia, e.g. [28]; with aphasia, e.g. [8]; second-language learners, e.g. [4]; children, e.g. [6]; low-literacy readers, e.g. [42]; and more closely related to our research, DHH adults, e.g. [3, 15, 18]. Part of the reason different researchers focus on specific user groups is that prior work has identified that the sources of reading difficulty for different user groups may vary [25], and thus the findings from one user group may not generalize to another as different user groups may benefit from different approaches to ATS.

When it comes to DHH people, as mentioned in the introduction, one of the key characteristics of this user group is their diversity in literacy skill. While many readers can possess age-appropriate reading skills, prior work has identified sixth-grade reading levels among university students [1, 26]. Other studies estimate that more than 30% of DHH students in the United States are "functionally illiterate" [39] (functional literacy roughly corresponds to between fourth- and eighth-grade reading levels in the United States [22]). ATS has the potential to benefit DHH readers given that prior literacy research has identified both syntactic and lexical aspects of text as key sources of difficulty for lower literacy DHH readers. In fact, prior research on ATS has found benefits from both syntactic [18] and lexical [3] approaches to ATS. Furthermore, recent research has found interest in having such tools from a specific sub-group of DHH adults: those in the computing field [2] who may need to learn new skills on their own for their career [9, 34].

2.2 Evaluating Automatic Text Simplification

Much ATS research has focused on its evaluation, e.g. [18, 28, 33], whether that is to evaluate new machine learning techniques or to evaluate its benefits as a reading assistance tool for a particular user group. When evaluating the former, most research focuses on evaluating the text from a linguistic perspective, to determine the quality of the output, e.g. [19, 21, 33], typically evaluating the

three different aspects outlined in the introduction: grammaticality, meaning preservation and simplicity. When evaluating ATS with a potential target user group, which is the focus of this paper, most research has focused on only one of those dimensions: the simplicity (or complexity) of the text. Researchers often distinguish between two dimensions of text complexity, namely understandability and readability. Understandability, which focuses more on the user, refers to how easy a text is to understand for a particular reader. Readability, in turn focuses more on the text, referring to how easy the text is to read intrinsically [36].

There has been wide diversity in how these evaluations have been conducted, including both manual and automatic evaluation methods. Manual evaluation of ATS can rely on "expert" native readers of the language, e.g. [33, 44], at times recruited through crowdsourcing, e.g. [19, 21]. However, most accessibility research with ATS relies on evaluating it with the target users themselves, e.g. [3, 18, 28, 33]. Automatic evaluation of ATS, in turn, uses software to analyze texts based on automatic metrics. These metrics can be based on machine translation metrics (e.g. BLEU) or on comparisons against human references of complexity (e.g. SARI) [46]. While the latter directly relies on human judgements, all automatic metrics ultimately rely on human judgements for validation. Thus, manual evaluation of ATS is still crucial even for the development of automatic metrics for automatic evaluation.

Most research that relies on manual evaluation typically provides participants with texts at different levels of complexity; these stimuli texts often consist of an original version and simplifications obtained as the output of ATS, e.g. [3, 21, 28, 33]. However, there is a lot of diversity in how complexity is measured. Objective metrics rely on behavioral metrics or scores from a task, and subjective metrics typically rely on direct judgements from participants.

In text complexity evaluation with target users, objective metrics of text complexity have varied. Some studies measure reading speed, under the premise that better comprehension would lead to higher reading speed, e.g. [20, 28, 31]. Other studies include comprehension questions, where the expectation is that better comprehension will lead to better scores, e.g. [18, 28, 33]. Other behavioral metrics have included eye-tracking, based on the premise that text complexity affects fixations duration [16, 28, 37]. However, while some studies have found significant differences from objective metrics such as comprehension questions, e.g. [18], and reading speed, e.g. [48], there have also been challenges: For example, in one study, expert readers judged texts as simpler, but evaluations with target readers using comprehension questions showed no significant differences [33]. Other studies revealed significant differences in subjective judgements from target readers, but not on comprehension questions [3, 28]. Inconsistencies with objective metrics have also been observed in the context of evaluating text complexity among DHH readers. For instance, in prior work on measuring text complexity in the health domain with both DHH and hearing readers, significant differences in comprehension questions scores were only revealed for the DHH group, who had significantly lower functional health literacy than the hearing group [18]. On the other hand, another study on evaluating ATS tools with DHH readers did not observe significant differences in comprehension questions scores, while there were significant differences in understandability ratings [3]. Finally, prior work has also suggested that comprehension questions can be confounded by a number of different factors, such as the participants' understanding of the questions themselves [13].

In research with expert readers, it is common for researchers to collect subjective judgements of grammaticality, meaning preservation and simplicity scores, e.g. [19, 21, 33]. In research with target readers, which typically focuses more on evaluating text complexity, subjective judgements typically focused on the two dimensions of complexity described above: understandability and readability. These are typically asked on some form of Likert scale. Understandability is typically measured with a user-focused statement such as "I was able to understand this text well," whereas readability uses text-focused statements such as "This text was easy to read," e.g. [3, 29, 30, 48]. While the two can seem similar, researchers have argued that they can be measured independently by using examples such as a well-written scientific paper which may be highly readable, but difficult to understand without the proper training [36]. Other work has focused on asking participants for an estimate of one of the objective metrics, under the assumption that better estimates (e.g. a participant thinking they did better on a quiz or read faster) may indicate better comprehension [20]. Studies that employ subjective responses with target readers tend to reveal significant differences between text complexity levels more readily, e.g. [20, 28], including in the context of evaluation among DHH readers, e.g. [3]. However, when using subjective metrics in accessibility, there are concerns of positive bias as well as accessibility barriers preventing users from appropriately judging their own performance [40].

All of this prior work relies on being able to effectively measure differences in complexity using the metrics described above (i.e. that if two texts are indeed at different complexity levels, the metrics would reveal significant differences). However, to the best of our knowledge, no prior work has examined whether any of these metrics are indeed effective at distinguishing different levels of text complexity with DHH readers across literacy levels. Prior work on evaluating other linguistic technologies (as summarized in the next section) with DHH users has identified that certain metrics work better with users at particular literacy levels [5]. This work has also revealed differences in terms of how participants with different literacy levels judge those technologies [5]. However, it remains unknown whether similar methodological findings would be obtained in the context of evaluating the complexity of ATS output.

2.3 Evaluating Linguistic Technologies with DHH Adults

There has been prior methodological research (presented at CHI'18) that focused on the evaluation of the quality of imperfect captioning tools with DHH adults. That study found that subjective impressions were the most effective with this particular user group [5]. In contrast, prior work on the evaluation of ASL animations, has suggested that subjective impressions of understandability should not replace comprehension questions [12, 14]. However, researchers in the captioning study also conducted analysis of sub-groups based on their literacy level, finding that the effectiveness of different

metrics even within that user group (DHH adults) may vary by their literacy level, with fewer metrics being effective for participants with lower literacy [5]. Considering these issues of literacy levels, we thus closely follow the methodology of [5], by using literacy level as a factor and, in our case, text complexity levels as another factor. While there is overlap, our study also differs in terms of the specific metrics used since we focus on those identified from prior work on evaluating ATS.

3 HYPOTHESES

With this context, we set out to investigate which metrics are effective for evaluating the complexity of the output from ATS with DHH readers, and how respondents' literacy levels may influence that effectiveness. Thus, as a methodological study, we employ carefully selected texts with human-made simplifications that are known to be at different complexity levels, as judged by a DHH literacy expert in our team. Then, we evaluate the following hypotheses for each metric included in our study:

(1) H1: When English texts of different complexity levels are evaluated by DHH individuals with *(a) lower English literacy skill* and *(b) higher English literacy skill*, the response scores for this metric will reveal statistically significant differences among the complexity levels. We refer to this characteristic as the **discriminative ability of the metric**, and this is a desirable characteristic. If there is a significant difference, then the metric is effective for use in evaluating text complexity.
(2) H2: In an overall analysis of response scores for all texts, when we compare response scores between DHH individuals in the higher-literacy and lower-literacy groups, we will observe a significant difference. We refer to this as the **literacy bias of the metric**. While this is not necessarily a problem with the metric that would prevent its use in evaluating text complexity, researchers who use this metric must ensure that they report the literacy skill level of their participants in a study, in order for results across studies to be comparable.

4 METHOD

4.1 Reading Stimuli

We needed texts at different levels of complexity as our source for ground truth for our study. One approach to obtain these could have been using an ATS system for acquiring the different levels of complexity (by simplifying an already-complex text). However, because ATS systems are still rather experimental, that approach would likely introduce errors [35], which could confound our results. Our study focuses on evaluating one aspect of the quality of ATS output text: complexity, not on its level of error. Thus, to ensure having texts at different levels of complexity – and with good quality – we employed a Wizard-of-Oz approach by using human-made simplifications for our study. We obtained our texts from Newsela[2], a website that provides news articles at different levels of linguistic complexity as a resource for school teachers, because its texts have been used by prior work in NLP as a source of simplifications for training ATS systems [45].

4.1.1 Text selection. We first selected 10 articles from the Science section of Newsela that were relatively similar in word-length (between 550 and 750), had Flesch-Kincaid grade levels close to 12th grade and also had the same number of simplified versions available of similar complexity. For each article, we selected three versions: the original version, the one with medium-level complexity as well as the simplest version available. A researcher in the team who is an expert in Deaf literacy then selected a subset of 6 articles to use in our study based on 1) whether the simplifications provided by Newsela would actually be simpler for DHH readers; 2) the background knowledge one could expect an average Deaf reader to have (e.g. one of the articles that were part of the initial set of 10 related to the use of sounds); and 3) prioritizing the ones who had the least number of grammatically complex comprehension questions (described in next subsection). Specifically, articles were omitted if the accompanying questions had more than one dependent clause or phrase because that makes them more difficult for DHH readers to understand [11].

The final set of 6 articles used in our study had an average length of 669 words (SD = 85.15) and Flesch-Kincaid grade level of 12.41 (SD = 0.86) in their original versions (henceforth referred to as our high-complexity condition), which means a high-school graduate in the U.S. should be able to understand them. In their medium-complexity version, their average length was of 714 words (SD = 76.78), with an average Flesch-Kincaid grade level of 8.9 (SD = 0.76), meaning a student in their first year of high-school in the U.S. should be able to understand them. Finally, in their low-complexity version, their average length was 412 words (SD = 128.51), and the average Flesch-Kincaid grade level was 4.3 (SD = 0.31), which should be understandable for a 4th grade student in the U.S. Links to all of these articles are included in Appendix A.

4.2 Metrics

We selected a set of metrics identified from prior work, including both objective and subjective metrics to evaluate in our study. These metrics included reading speed, comprehension questions (at different levels of linguistic complexity), participants' predictions of how well they did on the comprehension questions, as well as subjective judgements of the understandability and readability of the texts.

4.2.1 Reading Speed. We included reading speed as a potential objective metric of readability, since it had been used in prior work, as discussed in the Background and Related Work section, under the assumption that more readable text leads to higher reading speed. The reading speed was measured in words per minute (wpm).

4.2.2 Comprehension Questions. Our comprehension questions consisted of main-factual questions written as multiple-choice, which varied in their linguistic complexity. This is because, first, when using comprehension questions, most prior work has written them as multiple-choice e.g. [3, 10, 13, 18, 28]. Second, considering that there are also different types of questions in terms of what they ask of the users [10], we decided to use main-factual questions (i.e. questions that ask about relatively important aspects of a text)

[2]https://newsela.com

Table 1: One of the comprehension questions for an article, in its two versions.

High Linguistic Complexity	**Which of the following statements was true about the animal's physical composition?** a) It had no mechanism for chewing b) Its head was relatively small in relationship to its body c) It was able to blow fire out of its mouth
Low Linguistic Complexity	**What is true about the animal?** a) It had no way to chew. b) Its head was small. c) It could blow fire.

because prior work has identified them as easier than other types of questions [10], which allowed us to control for question-type difficulty. Finally, prior work has suggested that the varying linguistic complexity of questions (i.e. some questions can be harder to read than others) can affect whether a user actually understands the questions themselves [13]. So, we decided to test whether varying the degree of linguistic complexity of the comprehension questions themselves would have an effect on their effectiveness.

Now, we could have authored different comprehension questions for each article version (i.e. their high, medium and low complexity conditions) each article. However, we wanted to determine whether asking *the same questions* would effectively distinguish the varying complexity levels of the different article versions. This approach thus required that we authored the questions in a way that would not favor a particular article version by asking about a fact that could have been trimmed during the simplification process and thus not be present in one of the simplified versions. To address this, we identified facts that would be present in all three versions of each article. Then, we created 10 pairs of multiple-choice main-factual comprehension questions for each article. As illustrated in Table 1, each pair contained the same question (i.e. they were asking about the same fact and the multiple-choice options were the same fact), but both the question items and the options were written at a high and a low linguistic complexity level. When looking at the low-linguistic-complexity question in each pair across all of the texts, their average Flesch-Kincaid grade level was 3.75 (SD = 0.67) which means we could expect a U.S. 3rd- to 4th-grader to understand them. In turn, the average Flesch-Kincaid grade level for the high-linguistic-complexity questions was 9 (SD = 0.42), which should be understandable for a student in their first year of high school in the U.S. Then, we selected a subset of 6 pairs of questions by discarding those with higher grammatical complexity in their high-linguistic-complexity versions, or questions that asked about facts that were not directly tied to the main idea of the article.

The final set of questions consisted of 6 pairs of questions for each article. Now, while we could have shown both versions in each pair to participants, the version that would come second would have been likely to seem repetitive to participants and to elicit the same answer as the version of the same question that appeared first. So, instead we chose to show each participant 3 questions in their low-linguistic complexity version and 3 in their high-linguistic complexity version for each text, and rotated their selection for each article across participants. The 6 questions were then arranged randomly as a single quiz. Then, we scored participants' scores on the low-linguistic-complexity questions separately from the high-linguistic complexity questions to determine if either set was more effective than the other at identifying the complexity of the articles.

4.2.3 Score Prediction. We also included a question that asked participants to give a subjective estimate of how well they would do on a task. Specifically, we asked participants to estimate on a 0-100 scale the grade they expected to get on the quiz they had just completed. This question is similar to one used in prior research among DHH individuals, in which they were asked to predict their success at other academic tasks [41].

4.2.4 Understandability and Readability. We also included two subjective questions that have been widely used in the literature on the evaluation of text simplification. The first question, which focuses on a text's understandability and is thus more focused on the reader, reads as "I was able to understand this text well" and uses a 5-point Likert-type scale of agreement going from "Strongly disagree" to "Strongly agree." The second question, which focuses on the text's readability and is thus more focused on the text, reads as "This text was easy to read," and also uses the same 5-point Likert-type scale.

4.3 English Reading Literacy

In order to group participants by their literacy level, we needed a reliable metric of their English reading literacy skill. Thus, we administered the sentence comprehension sub-test of the Wide Range Achievement Test in its 4th edition (WRAT4) [43], which has been used in prior work as a measurement of literacy level with DHH readers because it is brief, can be administered without audio stimuli and has been previously validated with DHH people [17, 27].

4.4 Data Collection

4.4.1 Procedure. This IRB-approved study was conducted remotely due to social-distancing restrictions during the COVID-19 pandemic. Thus, participants were provided with a consent form via e-mail ahead of the study. Then, participants met via Zoom with a researcher on the team who is hard-of-hearing and fluent in ASL for a 70-minute appointment. Participants were directed to a website created using jsPsych [7], which contained all of the stimuli for the study. The first screen on the website was an introduction, which contained detailed instructions of the study and indicated to participants that they would be reading texts that had all been simplified using different simplification tools. We told participants that all texts had been simplified to avoid a placebo effect. No other indications of whether the texts had any transformations were provided to participants.

Each participant proceeded through all 6 articles, reading 2 at each complexity level (high, medium and low). The order of articles and text complexity conditions was rotated using a Graeco-Latin square design. After reading each article, participants answered the quiz described above, which contained 6 comprehension questions in total, 3 with low linguistic complexity and 3 with high linguistic complexity. Which 3 questions were selected at each linguistic complexity level was rotated across participants, and the order in which they were displayed in the quiz was randomized. After the quiz, participants were asked to predict their grade, followed by the

subjective metrics of understandability and readability. After the third article, participants were encouraged to take a quick break to prevent fatigue.

After reading all 6 articles, participants filled out the sentence comprehension sub-test of the WRAT, followed by a demographics questionnaire. Participants were then compensated with $40 for their participation.

4.4.2 Participants. Participants were recruited through social-media and email advertising based on the criteria of identifying as Deaf or Hard-of-Hearing (DHH) and being over 18 years old. We recruited a total of 59 participants for our study. Participants' self-identified genders included female (N = 31), male (N = 25) and non-binary (N = 1). Participants' average age was 27.33 (SD = 10.17, range = 18 - 63). A total of 19 participants identified as hard-of-hearing, with 28 identifying as culturally Deaf [3], 9 as deaf and one as Deaf/Blind. Participants' average WRAT score was 87.82 (SD = 15.58, range = 63 - 126), which is lower than the national average in the U.S. of 100 [43]. To investigate hypothesis H1, which focused on participants with lower (H1a) and higher literacy skills (H1b), we split participants into two groups based on their median WRAT score, which was 86. Our two groups, labeled as **WRAT-L** and **WRAT-H**, respectively, were as follows:

- **WRAT-L**: participants with WRAT scores of 86 or lower (mean = 75, SD = 6.95, range = 63 - 85).
- **WRAT-H**: participants with WRAT scores higher than 86 (mean = 100.6, SD = 10.34, range = 87 - 126).

Three participants did not have time to complete the WRAT sentence comprehension form, and thus their responses were not included in the analysis. Upon careful analysis of the reading speed of participants, following the methodology of [20], we excluded the data from two participants whose reading speed was higher than the median reading speed plus 3 times the Interquartile Range (IQR), which means they may have just been skimming through the text [4]. This left us with 54 participants and, after splitting them into two groups based on their median WRAT score as detailed above, each group consisted of 27 participants.

4.5 Data Analysis

After conducting Shapiro-Wilk tests of normality for all of the results, none of them followed a normal distribution, and thus we conducted non-parametric tests for our difference testing. For each of the literacy groups, we conducted Kruskal-Wallis tests to determine if there were statistical differences between the text complexity conditions (H1).

If there were significant differences, then we conducted post-hoc pairwise comparisons using Mann-Whitney U-tests with Bonferroni corrections. To compare between the two literacy groups (H2), given that this was a non-parametric between-groups comparison we conducted Mann-Whitney U-tests as well.

[3]We employ the custom of capitalizing the word Deaf for those who identify as members of the Deaf community [24].
[4]One participant only met this criteria for one of the articles, so we only excluded the data for that specific article for that participant.

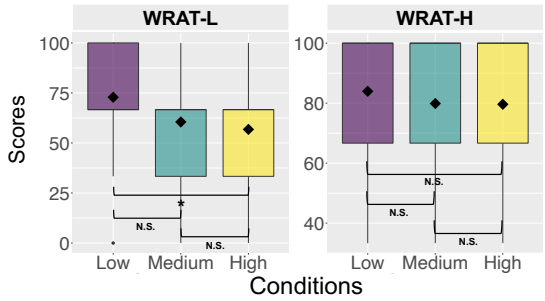

Figure 1: Low-linguistic complexity comprehension questions success for H1, with a max. value of 100% ($* = p < 0.05$).

5 RESULTS

As summarized in Table 2, we observed significant differences between text complexity levels when using low-linguistic-complexity comprehension questions as well as all of the subjective metrics (i.e. the score prediction, and the understandability and readability judgements) with the WRAT-L group. With the WRAT-H group, in turn, we only observed significant differences between text complexity levels when using readability judgements. There were significant differences between the two groups (WRAT-L and WRAT-H) with all of the metrics.

In the rest of this section, we present the detailed results of the statistical analysis first for H1, and its sub-hypotheses, followed by the results for H2 and all of its sub-hypotheses. Complementary to these detailed results, figures 1 through 10 illustrate significant results using whisker plots for continuous data (i.e. the reading speed, comprehension question scores for both the high-linguistic and low-linguistic complexity questions, and their score predictions), and stacked bar charts for the Likert-type data (i.e. understandability and readability judgements) which are the recommended way of plotting Likert-type scales [32]. Figures 1 through 4 show the results for each text complexity, grouped by literacy group (for hypotheses H1a and H1b), while figures 5 through 10 show the results for hypotheses H2, which looked at the results for each group as a whole.

5.1 Discriminative Ability (H1)

5.1.1 Reading Speed. For each group, the statistical analysis revealed no statistically significant differences for either group (p-value of .37 for WRAT-L and .12 for WRAT-H).

5.1.2 High-linguistic-complexity Comprehension Questions. The results from the statistical tests were not significant for either group (p-value of .78 for WRAT-L and .76 for WRAT-H).

5.1.3 Low-linguistic-complexity Comprehension Questions. The difference between the different text complexity levels was statistically significant for the WRAT-L group (χ^2 = 8.1383, p = .017). Pairwise comparisons revealed statistical differences between the low and high complexity conditions. For the WRAT-H group, however, the results were not statistically significant (p = .75). These results are illustrated in Figure 1.

Table 2: A summary of the results for each metric across each hypotheses

Metric	Discriminative Ability among Lower Literacy DHH Respondents (H1a)	Discriminative Ability among Higher Literacy DHH Respondents (H1b)	Literacy Bias (H2)
Reading speed	**H1a was not supported**. This metric was not discriminative between any text complexity levels.	**H1b was not supported**. This metric was not discriminative between any text complexity levels.	**H2 was supported**. Higher literacy readers had significantly higher reading speed than lower literacy readers.
High-linguistic-complexity comprehension questions	**H1a was not supported**. This metric was not discriminative between any text complexity levels.	**H1b was not supported**. This metric was not discriminative between any text complexity levels.	**H2 was supported**. Higher literacy readers had significantly higher scores than lower literacy readers.
Low-linguistic-complexity comprehension questions	**H1a was partially supported**. Worked well to distinguish between lowest and highest text complexity only.	**H1b was not supported**. This metric was not discriminative between any text complexity levels.	**H2 was supported**. Higher literacy readers had significantly higher scores than lower literacy readers.
Score prediction	**H1a was partially supported**. Worked well to distinguish between lowest and highest text complexity only.	**H1b was not supported**. This metric was not discriminative between any text complexity levels.	**H2 was supported**. Higher literacy readers predicted significantly higher scores than lower literacy readers.
Understandability "I was able to understand this text well"	**H1a was partially supported**. Worked well to distinguish low-complexity texts from both the medium and high-complexity texts.	**H1b was not supported**. This metric was not discriminative between any text complexity levels.	**H2 was supported**. Higher literacy readers had significantly higher judgements than lower literacy readers.
Readability (Best Metric) "This text was easy to read."	**H1a was partially supported**. Worked well to distinguish low-complexity texts from both the medium and high-complexity texts.	**H1b was partially supported**. Worked well to distinguish between lowest and highest text complexity only.	**H2 was supported**. Higher literacy readers had significantly higher judgements than lower literacy readers.

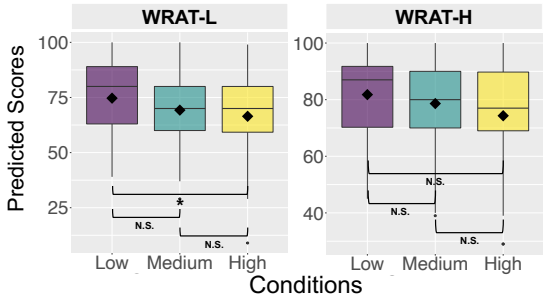

Figure 2: Score predictions for the comprehension questions for H1, with a max. value of 100% ($* = p < 0.05$).

5.1.4 Score prediction. As illustrated in Figure 2, there were significant differences between the text complexity conditions for the WRAT-L group (χ^2 = 8.1135, p = .017), with pairwise tests revealing differences between the low and high text complexity conditions. On the other hand, the results for the WRAT-H were not significant (p = 0.08)

5.1.5 Understandability. The results revealed significant differences between the text complexity conditions for the WRAT-L group (χ^2 = 13.0794, p = 0.001), with pairwise tests showing significant differences between the low and the medium complexity conditions, as well as between the low and the high complexity conditions. However, the results for the WRAT-H group revealed no significant differences (p = .35). Figure 3 illustrates these results.

5.1.6 Readability. Finally, as illustrated in Figure 4, the tests revealed significant differences between the text complexity levels for both literacy groups. For the WRAT-L group (χ^2 = 24.2346, p

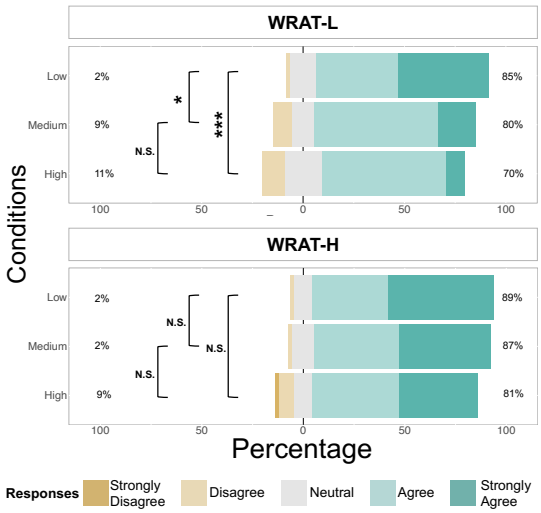

Figure 3: Understandability judgements for H1 using a Likert-type agreement scale ($* = p < 0.05, *** = p < 0.001$).

< 0.0001), pairwise comparisons revealed significant differences between the low and the medium complexity conditions, as well as between the low and the high complexity conditions.

For the WRAT-H group (χ^2 = 7.0867, p = .028), pairwise comparisons only revealed significant differences between the low and high text complexity conditions.

5.2 Literacy Bias (H2)

H2 was supported for all of the metrics, with the WRAT-H group obtaining higher scores or providing higher predictions of their

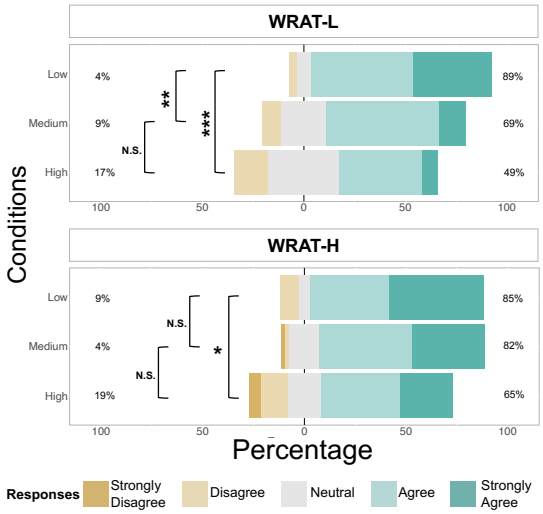

Figure 4: Readability judgements for H1 using a Likert-type agreement scale ($** = p < 0.01, *** = p < 0.001$).

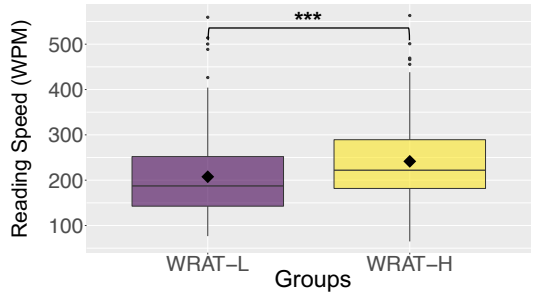

Figure 5: Reading speed for H2, calculated in words per minute (wpm) ($*** = p < 0.001$).

scores, as well as providing higher ratings for understandability and readability. For each metric, the results of the Mann-Whitney U-tests were:

- Reading speed: z-score = -3.7662, p < .001 (Figure 5).
- High-linguistic-complexity comprehension questions: z-score = -6.76736, p < .00001 (Figure 6).
- Low-linguistic-complexity: z-score = -5.21405, p < .00001 (Figure 7).
- Score prediction: z-score = -4.4865, p < .00001 (Figure 8).
- Understandability: z-score = -3.4653, p = .001 (Figure 9).
- Readability: z-score = -2.5901, p < .01 (Figure 10).

6 DISCUSSION

In this section, we discuss the results and their implications first for the hypothesis related to the discriminative ability of the metrics (H1), followed by the literacy bias of the metrics (H2).

6.1 Discriminative Ability (H1)

Overall, the metric that was most effective with both groups (WRAT-L and WRAT-H) was the subjective readability question "This text

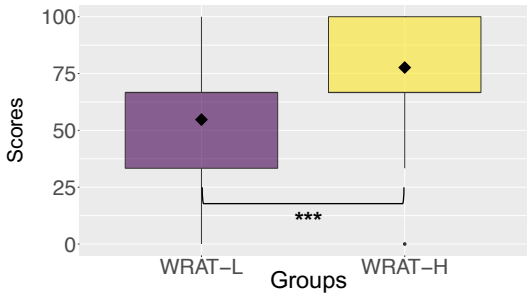

Figure 6: High-linguistic complexity comprehension questions success for H2, with a max. value of 100% ($*** = p < 0.001$).

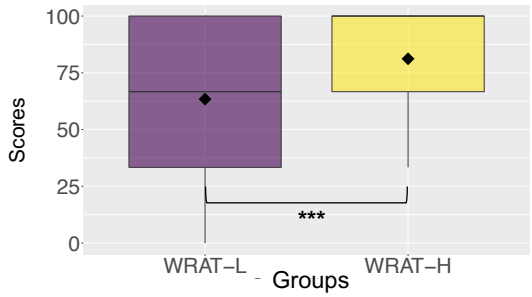

Figure 7: Low-linguistic complexity comprehension questions success for H2, with a max. value of 100% ($*** = p < 0.001$).

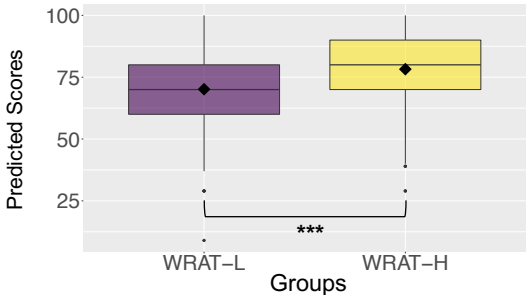

Figure 8: Score predictions for the comprehension questions for H2, with a max. value of 100% ($*** = p < 0.001$).

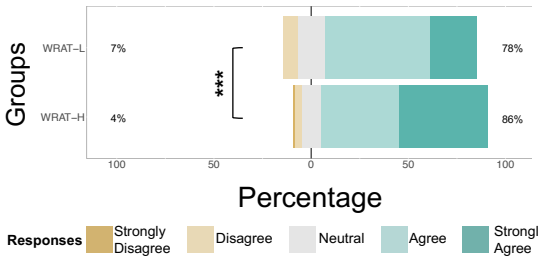

Figure 9: Understandability judgements for H2 using a Likert-type agreement scale ($*** = p < 0.001$).

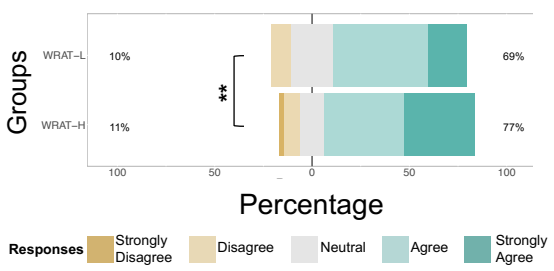

Figure 10: Readability judgements for H2 using a Likert-type agreement scale ($= p < 0.01$**).**

was easy to read." This was also one of only two metrics that was able to distinguish between the low and medium complexity conditions with WRAT-L, as well as the only one that was effective with the WRAT-H group.

When focusing on the objective metrics included (i.e. reading speed and comprehension questions), only comprehension questions were effective under three different conditions. First, comprehension questions were effective only when focusing on readers with lower literacy levels; we only observed significant differences across the complexity conditions when focusing on the WRAT-L group. The wording of the questions themselves has to be written in a way that is easy enough for those readers to understand them; we only observed significant differences when focusing on the low-linguistic-complexity comprehension questions. Finally, the texts also need to be far apart in complexity; we only observed significant differences between the low and high complexity conditions. Thus, we conclude that while some of these conditions may be of particular interest for researchers who want to employ comprehension questions, the results of comprehension questions as an objective metric of text complexity should be interpreted carefully unless those questions have been either carefully validated or piloted prior to using them as a metric with the particular literacy group of interest. However, as discussed in the next paragraph, subjective questions were highly effective for evaluating text complexity; so, while comprehension questions may have value in ensuring that participants pay attention to readings because of the pressure of being tested afterwards, subjective questions may be sufficient to evaluate the complexity of texts with DHH readers when those texts are known to be of good quality (as they were in our study given that they were human-made simplifications).

The subjective metrics (i.e. score prediction, understandability and readability judgements) were all effective with the WRAT-L group. While prior work has highlighted concerns around the meta-cognitive literacy knowledge skills of DHH readers (i.e. their ability to make judgements about their literacy skills), especially those with lower literacy levels [23, 38, 47], our results suggest that when focusing on the complexity of texts, readers with lower literacy are able to effectively judge this difference in relative complexity. However, only subjective judgements of readability were effective with the WRAT-H group, which highlights a trend in our results: it is more difficult to effectively differentiate complexity levels with higher literacy readers.

Our results are in line with prior work on the evaluation of the quality of captioning tools with DHH readers in terms of subjective metrics being more effective [5]. This is in contrast with prior work on the evaluation of ASL animations with DHH adults, which suggested subjective metrics should not be a replacement for comprehension questions [12]. However, our results are in stark contrast with the results of the captioning evaluation study [5] in terms of which group the metrics were more effective with; while their study revealed it was harder to evaluate imperfect captioning tools with lower literacy readers, our results suggest that it is harder to evaluate text complexity with higher literacy readers. We speculate that in our case, the articles with high complexity (which were around the 12th grade level) may not have been challenging enough for the higher literacy readers. Two of our subjective metrics focused on the readers: predicting how they had scored on the quiz and judging how easy the text was for them to understand. However, when focusing on the text, namely how easy the text was to read, higher literacy readers were able to judge it properly.

These results also helps illuminate inconsistencies described in the background and related work section on the evaluation of text complexity with DHH readers. In a study [18] that had revealed significant differences using comprehension questions among DHH readers, these significant differences were only revealed for the DHH group, who had significantly lower functional health literacy than the hearing control group. This suggests that one of the conditions identified in our results for comprehension questions may have been met: having a group with lower literacy. The text complexity levels in that study were also reportedly around 4 grade-levels apart, which may have met another condition identified in our results: text complexity levels being far apart. However, the researchers in that study did not report the linguistic complexity of their comprehension questions, so it is hard to determine if the condition of having questions that are easy to understand was met. In [3], their results did not reveal significant differences in comprehension questions scores, but there were significant differences in understandability ratings among DHH readers. However, researchers in that study only reported participants' average WRAT score without controlling for it when conducting the statistical analysis; so, it is difficult to determine if the lack of significance in comprehension question scores was due to not having differences in complexity of the text, or to not having a group of DHH readers with lower literacy. Further, they included more than one type of comprehension questions without reporting their linguistic complexity, so it is difficult to determine if their questions met the condition of being simple enough for participants to understand the questions themselves. Finally, the simplified text conditions were only about a grade level less than their original conditions, which suggests that they may not have been far apart enough for comprehension questions to reveal differences. Thus, it may have been the case that the significant differences in subjective judgements indeed aligned with text complexity, but the comprehension questions simply did not meet the conditions to be effective suggested by our results.

Our results suggest that subjective metrics may not only be discriminative of text complexity levels when evaluating them with DHH readers, but they are also more effective than our objective

metrics. Thus, we recommend always including subjective metrics in text-complexity studies with DHH readers. However, if researchers still want to include comprehension questions in their studies, we recommend being careful about the interpretation of a lack of significant differences in comprehension questions, unless they have been carefully crafted with low linguistic complexity and validated after controlling for the literacy level of participants. It is important to note that these recommendations are in the context of non-erroneous texts. However, given that, as with any automatic system, errors are inevitable in the output of ATS systems [35], further research is necessary to determine how the introduction of both semantic and grammatical errors would impact our results since some metrics may be able to capture the different types of errors better than others. Further, while we acknowledge we cannot guarantee that with more statistical power statistically significant differences would not emerge for some of the objective metrics in our study, we note that most prior work on evaluation of ATS systems with DHH readers has used smaller sample sizes without controlling for literacy levels. Thus, if we did not observe significant differences with our sample size and controlling for literacy levels, it is unlikely they would emerge in a study with fewer participants.

6.2 Literacy Bias (H2)

Unlike prior work on the evaluation of captioning tools, in which some metrics did not exhibit a literacy bias, our results suggest that all of the metrics we used had a literacy bias. More specifically, participants in the WRAT-H group scored and predicted higher scores, and also judged texts to be easier than participants in the WRAT-L group. This does not mean that these are bad metrics. Instead, this suggests that researchers should always carefully report and/or control for the literacy levels of the participants in their study when evaluating the complexity of texts with DHH readers. Readers of published research studies should then consider this literacy bias when comparing across studies, since texts of equal complexity may elicit different scores depending on the literacy level of the participants they are evaluated with.

It is difficult to generalize this result to other user groups (i.e. whether these metrics would have a literacy bias with other user groups) given that prior work has established that the sources of difficulty with texts may be different for different user groups [25]. Nevertheless, these results are still useful to researchers working with other user groups to know that these metrics may be biased for participants' reading skill. Similar analysis using measurements of the relevant skill a specific user group struggles with could reveal similar trends.

7 LIMITATIONS AND FUTURE WORK

There were several limitations to our study. While it was our original intention to include measures obtained through eye-tracking (e.g. fixation duration) in our study, given that they have been used in prior work on evaluating ATS with target readers, we were prevented from using eye-tracking because we had to conduct our study remotely to abide by social-distancing restrictions due to COVID-19. Future work can thus investigate the effectiveness of eye-tracking for distinguishing text complexity levels among DHH readers.

In this study, we decided to focus only on one aspect of the potential output of an ATS system: the level of complexity of the output. But as mentioned in our discussion, the introduction of errors in the output is inevitable [35]. So, future work should focus on how to measure the overall quality of the output of simplification systems with DHH readers, and how the introduction of semantic and grammatical errors could affect the effectiveness of the metrics for evaluating text complexity suggested by our results.

As mentioned in our discussion, the high-complexity conditions for the articles in our study may not have been challenging enough for participants with higher literacy levels. Future work could determine whether some of the metrics analyzed in our study would be effective for evaluating texts with higher complexity among participants with higher literacy. Further, while we grouped participants into two groups using their WRAT scores, we could have recruited a greater number of participants or used a different literacy test to determine whether further subdivision could reveal other patterns. Another limitation of our study was that our sample size did not allow us to guarantee that with a larger sample size, and thus more statistical power, significant differences would not emerge for metrics that we did not observe significant differences in this study.

In this study, our participants identified as DHH, but not necessarily as users of ASL. A future study could recruit a more narrow demographic of participants, with a specific focus on ASL signers, in which ASL-based comprehension questions could also be examined. While we only included texts in the science genre as stimuli, future work could explore whether our findings would still hold with other text genres. Finally, future work could focus on exploring the efficacy of the metrics explored in our study with other user groups and broader literacy levels.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under award No. 1822747.

REFERENCES

[1] J. Albertini and C. Mayer. 2011. Using Miscue Analysis to Assess Comprehension in Deaf College Readers. *Journal of Deaf Studies and Deaf Education* 16, 1 (2011), 35–46. https://doi.org/10.1093/deafed/enq017

[2] Oliver Alonzo, Lisa Elliot, Becca Dingman, and Matt Huenerfauth. 2020. Reading Experiences and Interest in Reading-Assistance Tools Among Deaf and Hard-of-Hearing Computing Professionals. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, Greece) *(ASSETS '20)*. Association for Computing Machinery, New York, NY, USA, 13. https://doi.org/10.1145/3373625.3416992

[3] Oliver Alonzo, Matthew Seita, Abraham Glasser, and Matt Huenerfauth. 2020. Automatic Text Simplification Tools for Deaf and Hard of Hearing Adults: Benefits of Lexical Simplification and Providing Users with Autonomy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376563

[4] Mahmoud Azab, Chris Hokamp, and Rada Mihalcea. 2015. Using Word Semantics To Assist English as a Second Language Learners. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. Association for Computational Linguistics, Denver, Colorado, 116–120. https://doi.org/10.3115/v1/N15-3024

[5] Larwan Berke, Sushant Kafle, and Matt Huenerfauth. 2018. Methods for Evaluation of Imperfect Captioning Tools by Deaf or Hard-of-Hearing Users at Different Reading Literacy Levels. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, Article 91, 12 pages. https://doi.org/10.1145/3173574.3173665

[6] Jan De Belder and Marie-Francine Moens. 2010. Text simplification for children. In *Prroceedings of the SIGIR workshop on accessible search systems*. ACM; New

York, 19–26.
[7] Joshua R De Leeuw. 2015. jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior research methods* 47, 1 (2015), 1–12.
[8] Siobhan Devlin and Gary Unthank. 2006. Helping Aphasic People Process Online Information. In *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility* (Portland, Oregon, USA) *(Assets '06)*. ACM, New York, NY, USA, 225–226. https://doi.org/10.1145/1168987.1169027
[9] Jack Downey. 2010. Careers in Software: Is There Life after Programming?. In *Proceedings of the 2010 Special Interest Group on Management Information System's 48th Annual Conference on Computer Personnel Research on Computer Personnel Research* (Vancouver, BC, Canada) *(SIGMIS-CPR '10)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/1796900.1796912
[10] Mary C Dyson and Mark Haselgrove. 2001. The Influence of Reading Speed and Line Length on the Effectiveness of Reading from Screen. *Int. J. Hum.-Comput. Stud.* 54, 4 (April 2001), 585–612. https://doi.org/10.1006/ijhc.2001.0458
[11] Susan R Easterbrooks and Sharon Baker. 2002. *Language Learning in Children Who Are Deaf and Hard of Hearing: Multiple Pathways*. ERIC.
[12] Matt Huenerfauth. 2008. Evaluation of a Psycholinguistically Motivated Timing Model for Animations of American Sign Language. In *Proceedings of the 10th International ACM SIGACCESS Conference on Computers and Accessibility* (Halifax, Nova Scotia, Canada) *(Assets '08)*. Association for Computing Machinery, New York, NY, USA, 129–136. https://doi.org/10.1145/1414471.1414496
[13] Matt Huenerfauth, Lijun Feng, and Noémie Elhadad. 2009. Comparing Evaluation Techniques for Text Readability Software for Adults with Intellectual Disabilities. In *Proceedings of the 11th International ACM SIGACCESS Conference on Computers and Accessibility* (Pittsburgh, Pennsylvania, USA) *(Assets '09)*. Association for Computing Machinery, New York, NY, USA, 3–10. https://doi.org/10.1145/1639642.1639646
[14] Matt Huenerfauth and Hernisa Kacorri. 2015. Best practices for conducting evaluations of sign language animation. In *30th Annual International Technology and Persons with Disabilities Conference Scientific/Research Proceedings*. California State University, Northridge.
[15] Kentaro Inui, Atsushi Fujita, Tetsuro Takahashi, Ryu Iida, and Tomoya Iwakura. 2003. Text Simplification for Reading Assistance: A Project Note. In *Proceedings of the Second International Workshop on Paraphrasing - Volume 16* (Sapporo, Japan) *(PARAPHRASE '03)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 9–16. https://doi.org/10.3115/1118984.1118986
[16] Marcel Adam Just and Patricia A. Carpenter. 1984. *Using Eye Fixations to Study Reading Comprehension*. Erlbaum, 151–182.
[17] Lynda J Katz and Franklin C Brown. 2019. Aptitude and achievement testing. In *Handbook of Psychological Assessment*. Elsevier, 143–168.
[18] Poorna Kushalnagar, Scott Smith, Melinda Hopper, Claire Ryan, Micah Rinkevich, and Raja Kushalnagar. 2018. Making cancer health text on the Internet easier to read for deaf people who use American Sign Language. *Journal of Cancer Education* 33, 1 (2018), 134–140.
[19] Walter S. Lasecki, Luz Rello, and Jeffrey P. Bigham. 2015. Measuring Text Simplification with the Crowd. In *Proceedings of the 12th Web for All Conference* (Florence, Italy) *(W4A '15)*. Association for Computing Machinery, New York, NY, USA, Article 4, 9 pages. https://doi.org/10.1145/2745555.2746658
[20] Qisheng Li, Meredith Ringel Morris, Adam Fourney, Kevin Larson, and Katherina Reinecke. 2019. The Impact of Web Browser Reader Views on Reading Speed and User Experience. In *CHI 2019*. ACM. https://www.microsoft.com/en-us/research/publication/the-impact-of-web-browser-reader-views-on-reading-speed-and-user-experience/
[21] Angrosh Annayappan Mandya, Tadashi Nomoto, and Advaith Siddharthan. 2014. Lexico-syntactic text simplification and compression with typed dependencies. In *25th International Conference on Computational Linguistics*.
[22] Marc Marschark, John A. Albertini, and Harry G. Lang. 2002. *Educating deaf students: from research to practice*. Oxford University Press.
[23] Carolyn Morrison, Marc Marschark, Thomastine Sarchet, Carol M. Convertino, Georgianna Borgna, and Richard Dirmyer. 2013. Deaf students' metacognitive awareness during language comprehension. *European Journal of Special Needs Education* 28, 1 (2013), 78–90. https://doi.org/10.1080/08856257.2012.749610
[24] Carol Padden, Tom Humphries, and Carol Padden. 2009. *Inside deaf culture*. Harvard University Press.
[25] Gustavo Paetzold and Lucia Specia. 2016. Understanding the Lexical Simplification Needs of Non-Native Speakers of English. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, 717–727. https://www.aclweb.org/anthology/C16-1069
[26] S. J. Parault and H. M. Williams. 2010. Reading Motivation, Reading Amount, and Text Comprehension in Deaf and Hearing Adults. *Journal of Deaf Studies and Deaf Education* 15, 2 (2010), 120–135. https://doi.org/10.1093/deafed/enp031
[27] LeAdelle Phelps and Barbara Jane Branyan. 1990. Academic achievement and nonverbal intelligence in public school hearing-impaired children. *Psychology in the Schools* 27, 3 (1990), 210–217.
[28] Luz Rello, Ricardo Baeza-Yates, Stefan Bott, and Horacio Saggion. 2013. Simplify or Help?: Text Simplification Strategies for People with Dyslexia. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility* (Rio de Janeiro, Brazil) *(W4A '13)*. ACM, New York, NY, USA, Article 15, 10 pages. https://doi.org/10.1145/2461121.2461126
[29] Luz Rello, Roberto Carlini, Ricardo Baeza-Yates, and Jeffrey P. Bigham. 2015. A Plug-in to Aid Online Reading in Spanish. In *Proceedings of the 12th Web for All Conference* (Florence, Italy) *(W4A '15)*. Association for Computing Machinery, New York, NY, USA, Article 7, 4 pages. https://doi.org/10.1145/2745555.2746661
[30] 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* (San Jose, California, USA) *(CHI '16)*. ACM, New York, NY, USA, 3637–3648. https://doi.org/10.1145/2858036.2858204
[31] Luz Rello, Horacio Saggion, Ricardo Baeza-Yates, and Eduardo Graells. 2012. Graphical schemes may improve readability but not understandability for people with dyslexia. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*. 25–32.
[32] Naomi B Robbins, Richard M Heiberger, et al. 2011. Plotting Likert and other rating scales. In *Proceedings of the 2011 Joint Statistical Meeting*. 1058–1066.
[33] Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarevic. 2015. Making It Simplext: Implementation and Evaluation of a Text Simplification System for Spanish. *ACM Trans. Access. Comput.* 6, 4, Article 14 (May 2015), 36 pages. https://doi.org/10.1145/2738046
[34] Tenace Setor, Damien Joseph, and Shirish C. Srivastava. 2015. Professional Obsolescence in IT: The Relationships between the Threat of Professional Obsolescence, Coping and Psychological Strain.. In *Proceedings of the 2015 ACM SIGMIS Conference on Computers and People Research* (Newport Beach, California, USA) *(SIGMIS-CPR '15)*. Association for Computing Machinery, New York, NY, USA, 117–122. https://doi.org/10.1145/2751957.2751962
[35] Matthew Shardlow. 2014. Out in the Open: Finding and Categorising Errors in the Lexical Simplification Pipeline. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), Reykjavik, Iceland, 1583–1590. http://www.lrec-conf.org/proceedings/lrec2014/pdf/479_Paper.pdf
[36] Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications* 4, 1 (2014), 58–70.
[37] Advaith Siddharthan. 2014. A survey of research on text simplification. *ITL - International Journal of Applied Linguistics* 165, 2 (2014), 259–298. https://doi.org/10.1075/itl.165.2.06sid
[38] Barbara K Strassman. 1997. Metacognition and reading in children who are deaf: A review of the research. *Journal of Deaf Studies and Deaf Education* (1997), 140–149.
[39] C. B. Traxler. 2000. The Stanford Achievement Test, 9th Edition: National Norming and Performance Standards for Deaf and Hard-of-Hearing Students. *Journal of Deaf Studies and Deaf Education* 5, 4 (Jan 2000), 337–348. https://doi.org/10.1093/deafed/5.4.337
[40] Shari Trewin, Diogo Marques, and Tiago Guerreiro. 2015. Usage of Subjective Scales in Accessibility Research. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility* (Lisbon, Portugal) *(ASSETS '15)*. ACM, New York, NY, USA, 59–67. https://doi.org/10.1145/2700648.2809867
[41] Dawn Walton, Georgianna Borgna, Marc Marschark, Kathryn Crowe, and Jessica Trussell. 2019. I am not unskilled and unaware: deaf and hearing learners' self-assessments of linguistic and nonlinguistic skills. *European Journal of Special Needs Education* 34, 1 (2019), 20–34. https://doi.org/10.1080/08856257.2018.1435010
[42] Willian Massami Watanabe, Arnaldo Candido Junior, Vinícius Rodriguez Uzêda, Renata Pontin de Mattos Fortes, Thiago Alexandre Salgueiro Pardo, and Sandra Maria Aluísio. 2009. Facilita: Reading Assistance for Low-literacy Readers. In *Proceedings of the 27th ACM International Conference on Design of Communication* (Bloomington, Indiana, USA) *(SIGDOC '09)*. ACM, New York, NY, USA, 29–36. https://doi.org/10.1145/1621995.1622002
[43] Gary S. Wilkinson and Gary J. Robertson. 2006. *Wide Range Achievement Test 4 professional manual*. Psychological Assessment Resources, Inc.
[44] Kristian Woodsend and Mirella Lapata. 2011. Learning to Simplify Sentences with Quasi-Synchronous Grammar and Integer Programming. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Edinburgh, Scotland, UK., 409–420. https://www.aclweb.org/anthology/D11-1038
[45] Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in Current Text Simplification Research: New Data Can Help. *Transactions of the Association for Computational Linguistics* 3 (2015), 283–297. https://doi.org/10.1162/tacl_a_00139
[46] Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing Statistical Machine Translation for Text Simplification. *Transactions of the Association for Computational Linguistics* 4 (2016), 401–415. https://cocoxu.github.io/publications/tacl2016-smt-simplification.pdf

[47] Peixuan Yan. 2018. *Effects of Metacognition on English Reading Outcomes for d/Deaf and Hard of Hearing Students.* Ph.D. Dissertation. The Ohio State University.
[48] 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* (Atlanta, Georgia, USA) *(CHI '10)*. ACM, New York, NY, USA, 2523–2532. https://doi.org/10.1145/1753326.1753709

A READING STIMULI

This appendix contains the links to each of the Newsela articles that were used in this study as reading stimuli in each complexity condition, identified by the codenames we created for each one of them based on their main topic:

- Smartphone typing: high, medium and low complexity.
- Cryodrakon: high, medium and low complexity.
- Bubonic plage: high, medium and low complexity.
- Salmon cannon: high, medium and low complexity.
- Lizard: high, medium and low complexity.
- Garfield: high, medium and low complexity.