Methods for Evaluating the Fluency of Automatically Simplified Texts with Deaf and Hard-of-Hearing Adults at Various Literacy Levels

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

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

Matthew Watkins
mxw7981@rit.edu
School of Information
Rochester Institute of Technology
Rochester, NY, USA

Sooyeon Lee
slics@rit.edu
School of Information
Rochester Institute of Technology
Rochester, NY, USA

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

ABSTRACT

Research has revealed benefits and interest among Deaf and Hard-of-Hearing (DHH) adults in reading-assistance tools powered by Automatic Text Simplification (ATS), a technology whose development benefits from evaluations by specific user groups. While prior work has provided guidance for evaluating text complexity among DHH adults, researchers lack guidance for evaluating the fluency of automatically simplified texts, which may contain errors from the simplification process. Thus, we conduct methodological research on the effectiveness of metrics (including reading speed; comprehension questions; and subjective judgements of understandability, readability, grammaticality, and system performance) for evaluating texts controlled to be at different levels of fluency, when measured among DHH participants at different literacy levels. Reading speed and grammaticality judgements effectively distinguished fluency levels among participants across literacy levels. Readability and understandability judgements, however, only worked among participants with higher literacy. Our findings provide methodological guidance for designing ATS evaluations with DHH participants.

CCS CONCEPTS

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

KEYWORDS

automatic text simplification, accessibility, deaf and hard-of-hearing, methodological research

CHI '22, April 30–May 6, 2022, New Orleans, LA

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/3491102.3517566

ACM Reference Format:
Oliver Alonzo, Jessica Trussell, Matthew Watkins, Sooyeon Lee, and Matt Huenerfauth. 2022. Methods for Evaluating the Fluency of Automatically Simplified Texts with Deaf and Hard-of-Hearing Adults at Various Literacy Levels. In *CHI '22: The ACM CHI Conference on Human Factors in Computing Systems, April 30–May 6, 2022, New Orleans, LA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3491102.3517566

1 INTRODUCTION

Automatic text simplification (ATS) consists of computing techniques to rewrite text with the goal of reducing its linguistic complexity while maintaining grammatical correctness (or fluency) and preserving the meaning [33, 34]. Research has evaluated the use of ATS to provide reading assistance to various user groups, including low-literacy readers [37], readers with dyslexia [26, 27], or second-language learners [4]. As prior research has identified great diversity in literacy skill among Deaf and Hard-of-Hearing (DHH) adults in the U.S., e.g. [20, 24, 35], the use of ATS to provide reading assistance to DHH adults has also been explored, e.g., [2, 13, 16].

ATS research is rapidly progressing in the natural language processing (NLP) community, focusing on improving the underlying technologies. Thus, human evaluation of the output of ATS systems is important, e.g., [3, 26, 31, 33, 39], and several characteristics are typically evaluated [34]. First, there is the **complexity** of the output, i.e., whether the texts are indeed simpler. However, errors may be introduced in the process of automatically simplifying texts, which are often grammatical or semantic errors [32]. Thus, two other characteristics of the output texts are also important to evaluate: their **fluency** (or grammatical correctness) and **faithfulness** (preservation of meaning of the original text) [34].

Prior work has identified which metrics are effective for evaluating the *complexity* of simplified texts among DHH readers, finding that popular metrics (e.g., comprehension questions) were less effective than subjective judgements, despite the meta-cognitive skills such judgements may require [3]. These findings suggest that methodological research is needed to validate researchers' intuitions about which metrics to use to evaluate characteristics of ATS output. However, that study had only focused on the *complexity* of

ATS output, even though imperfect ATS technologies can damage a text's *fluency*. Without further methodological research, NLP researchers currently lack guidance as to how to fully evaluate both the *complexity* and the *fluency* of ATS output among DHH readers.

Evaluations of *fluency* and *faithfulness* do not typically involve target reader groups. Instead, they are mostly conducted with high-literacy readers, who are often referred to in the literature as "native" or "expert" readers. Because prior research had specifically revealed that the literacy level of a reader influences their judgements about a text's *fluency* [11], we were motivated to investigate how to include target reader groups in evaluations of the *fluency* of simplified texts, as their impressions may differ from those of expert readers. To the best of our knowledge, however, no prior work has identified guidance on whether the *fluency* of simplified texts can be evaluated among a target reader group such as DHH adults and if so, which metrics would be more effective, and whether those metrics would be affected by the participants' literacy levels.

We thus conduct a methodological study to investigate whether there are metrics, for use among DHH readers, that can distinguish between automatically simplified texts that are known to differ in their level of *fluency*. A total of 29 participants read texts carefully engineered to be at various levels of *fluency* using the output of ATS systems combined with human-produced simplifications. We recorded particiants' responses to various metrics, including those from prior work, such as reading speed, comprehension questions written at different levels of linguistic complexity [3], and subjective judgements. Among our DHH participants across a range of reading literacy, we found that their reading speed and subjective judgements of a text's grammaticality were effective at distinguishing between levels of *fluency* in automatically simplified texts.

The contributions of this work include:

(1) Empirical evidence that the *fluency* of simplified texts can be evaluated among DHH adults at various literacy levels.
(2) Methodological guidance as to which metrics are capable of distinguishing different levels of *fluency* of simplified texts, for use in studies with DHH adults at various literacy levels.
(3) A framework for researchers to conduct methodological research on how to evaluate the *fluency* of simplified texts among other target users of reading-assistance tools.

2 BACKGROUND AND RELATED WORK

ATS may be applied: (a) at the syntactic level by modifying the structure of phrases, e.g., [9], (b) at the lexical level by replacing complex words with simpler synonyms or paraphrases, e.g., [19], or (c) at both levels, e.g., [39].

Accessibility researchers have evaluated the use of ATS-based reading assistance tools among multiple groups, including people with dyslexia [26, 27], people with aphasia [10], children [7] and people who are DHH [2, 16]. Numerous user studies have been conducted with these user groups focusing on various aspects of the technology. For instance, some research investigated design aspects of the user interface of ATS tools, e.g., [2, 6], while others have focused on measuring benefits from tools that provide the various forms of ATS outlined above, e.g., [2, 16, 26, 31]. Others yet have evaluated the linguistic needs of various user groups, revealing that specific linguistic properties impact readability differently, depending on the user group [23]. Overall, these findings reveal the need for research that evaluates ATS systems with specific reader groups, as the linguistic needs and preferences of each may vary.

2.1 Interests and Benefits Among DHH Adults

Research using standardized testing has revealed U.S. fourth-grade reading levels (typically corresponding to students who are 9 or 10 years old) among subsets of DHH high-school graduates (who are typically around 18 years old) [35]. Other research has described 30% of deaf-high school graduates in the U.S. as "functionally illiterate" [20]. However, these are subsets of the samples in these studies and do not reflect the entire population. What these suggest is that while many DHH readers have age-appropriate reading skills, significant subsets of DHH adults may face challenges when reading and thus, there is great diversity in literacy skill among this user group.

Prior work with DHH readers has revealed benefits from syntactic approaches when simplifying medical texts [16], and perceived benefits from lexical simplification for science-related texts [2]. Other research has investigated the interests and benefits of ATS among a particular subset of DHH adults: those with experience in the computing and information technology field [1].

2.2 Evaluating Complexity of Simplified Texts

As discussed in Section 1, ATS systems are typically evaluated in terms of the resulting text's *complexity*, its *faithfulness* in preserving the meaning of the original text, and its *fluency* (grammatical correctness) [34]. Human evaluations of *complexity*, which are done with both expert/native readers and target users of ATS, are typically conducted using various metrics, including comprehension questions (e.g., [16, 26]), reading speed (e.g., [26, 29]), and judgements of understandability (e.g., [2, 31]) or readability (e.g., [2, 31]).

Despite the use of various metrics for measuring text *complexity* in studies with target users of ATS technologies, little methodological work had established the validity of these instruments with such users. The most closely related work to the current paper consisted of a methodological study, with DHH adult readers across a range of reading literacy levels investigating whether various metrics could measure differences between texts known to be at different *complexity* levels [3]. Researchers in that study determined the effectiveness of metrics for distinguishing between complex or simplified texts, finding that some metrics only worked among DHH readers of particular reading literacy levels. For instance, comprehension questions that had been specifically written with low-linguistic complexity were able to measure differences in *complexity* between texts among DHH readers with lower literacy, and several subjective metrics were able to distinguish some differences in text *complexity*. When evaluating texts with participants with higher literacy, only subjective judgements of readability were able to measure any differences. These findings suggest objective metrics such as comprehension questions need to be carefully crafted to ensure they can be used to measure such differences [3]. A key limitation of this prior study was that it had only considered how to evaluate text *complexity*, even though errors may be introduced when employing ATS. Further, that prior study had only evaluated texts produced by human editors, rather than examining texts which had actually been processed by ATS technology.

2.3 Evaluating Fluency of ATS

The aforementioned methodological study [4] had only provided guidance on how DHH readers could evaluate *complexity*. However, researchers often wish to measure a simplified text's *fluency* and its *faithfulness* in preserving the meaning of the original text, e.g., [31, 39]. Researchers traditionally evaluate these two aspects exclusively with expert/native readers by using scalar instruments to compare a simplified text to the original, e.g., [31].

While there may be benefits from asking expert readers to examine an original and simplified text to determine whether the meaning has been faithfully preserved, the exclusion of target users from evaluations of text *fluency* is more difficult to justify. Prior work has revealed how specific target groups differ in their linguistic needs and preferences when reading texts, e.g., [23], and other studies have revealed that a reader's level of literacy influences their subjective judgements about a text's *fluency*, e.g. [11]. These suggest that it may be more valid if the *fluency* of ATS output could be evaluated with target reader groups. However, there has been a lack of prior methodological work on evaluation of text *fluency* among lower-literacy readers, specifically among DHH adults.

2.4 Evaluating other Linguistic Technologies among DHH Adults

There has been prior methodological research on evaluating other linguistic technologies, which may produce imperfect output, among DHH adults at various literacy levels. For instance, considering that approximately half a million people use American Sign Language (ASL) as their primary means of communication [21] prior methodological research has been conducted on the evaluation of ASL animations in studies with DHH participants, revealing benefits from collecting responses to both subjective judgements and objective comprehension questions about the content [12].

Berke et. al investigated how to evaluate imperfect video captions from automatic speech-recognition (ASR) with DHH adults at various literacy levels [5]. They found it was easier to measure differences in caption quality with higher-literacy participants using subjective metrics, while for lower-literacy participants, subjective metrics requiring meta-cognitive insight into text quality were not effective. Only some objective metrics were effective at evaluating caption quality among lower-literacy participants. In contrast, the aforementioned study with DHH adults evaluating the *complexity* of simplified text [3] had found that objective metrics like comprehension questions only worked among lower-literacy participants. Given these mixed results, we have included both subjective-judgement questions and objective comprehension questions in the set of metrics evaluated in our current study.

3 HYPOTHESES

We investigate which metrics (e.g., participants' reading speed, responses to comprehension questions and subjective judgements) are effective among DHH adults for evaluating the *fluency* of simplified texts. We also investigate whether participants' responses to metrics varies depending upon their reading-literacy level.

For each metric, we evaluate the following hypotheses:

(1) **H1:** When evaluating English texts at different *fluency* levels, due to grammatical errors introduced by ATS, participants' responses for this metric will reveal statistically significant differences among texts of different *fluency* levels. This characteristic, which is desirable, has been referred to as the **discriminative ability of the metric** in prior work [3, 5]. If there is a significant difference, then the metric is effective for use in evaluating the *fluency* of the texts. For each metric, we investigate this hypothesis among two sub-groups of DHH readers: **(H1a)** those with lower English literacy skill and **(H1b)** those with higher English literacy skill.
(2) **H2:** When comparing the response scores of DHH individuals in a higher-literacy and lower-literacy group for all texts, a significant difference will be observed. This has been referred to in prior work as the **literacy bias of the metric** [3, 5] and it is is not necessarily a problem with the metric nor would it prevent its use for evaluating the *fluency* of texts. Instead, when using this metric in a study, researchers need to consider and report the literacy skill level of their participants so that results across studies can be comparable.

4 METHOD

4.1 Reading Stimuli

In typical evaluations of ATS output, participants may respond to a set of *trusted* question-instruments to assess texts with *unknown* levels of *fluency*. However, as this is a methodological study, the study design differs (informally, it may feel "backwards"): We need a set of texts known to be at specific levels of *fluency*, and ask participants to respond to question-instruments to determine whether those question-instruments are able to discriminate between texts of different *fluency* and whether there are any literacy biases in the response to those items. Thus, it is essential that our text stimuli are carefully engineered such that they have specific levels of *fluency*.

As our goal is to inform researchers evaluating ATS systems, we wanted our stimuli texts to exhibit realistic levels of quality from automatic systems. However, it would have been difficult to simply process texts with ATS while carefully controlling for both *fluency* and *complexity* levels. Thus, we employed a semi-automated process by mixing sentences from the output of two state-of-the-art ATS systems and manual simplifications, such that the resulting texts consisted of a mixture of sentences from: the original (complex) text, the output of one of the ATS systems, and a simplified text that had been produced by a human author. Details of our procedure for generating text stimuli are described below.

4.2 Stimuli Generation Procedure

4.2.1 Articles. We selected 6 articles from Newsela[1], an educational website that provides human simplified versions of news articles. Our 6 articles, which were about science-related topics, had been identified as appropriate to use with DHH readers and used in prior methodological work for evaluating the *complexity* of simplified texts among DHH readers [3]. Before simplification, these articles had an average Flesch-Kincaid grade level of 12.4 (SD = 0.86). We will refer to these original versions of these articles as "Original."

4.2.2 Simplifications. We processed the original articles through:

[1] https://newsela.com

(1) A state-of-the-art hybrid ATS system that incorporates both rule-based and data-driven models [18]. A parameter controlling the number of words copied from the original input when paraphrasing was set at 70%, obtaining output with an average Flesch-Kincaid grade level of 8.8 (SD = 0.77). We refer to these versions of the articles as "Hybrid."
(2) A state-of-the-art Transformer-based model trained on datasets presented in [14], which provided output with an average Flesch-Kincaid grade level of 6.7 (SD = 0.9). We refer to these versions of the articles as "Transformer."

While one of our goals was to include texts that exhibited realistic levels of *fluency* from ATS systems, we also needed to control the *fluency* and *complexity* levels of the texts. So, to obtain sentences with high-*fluency* and low-*complexity* (for use as text stimuli), we selected a human-authored simplification for each article from the Newsela dataset, with an average Flesch-Kincaid grade level of 8.9 (SD = 0.76). These were the closest to the average literacy grade-levels observed in prior work with DHH adults (e.g. in [3, 5]). We will refer to these versions of the articles as "Newsela."

Thus, for each of our six articles, we had four versions: (1) an Original version, (2) output from the Hybrid ATS, (3) output from the Transformer ATS, and (4) a human-authored Newsela simplification. Our goal was to assemble text stimuli using sentences from these four sources. Notably, sentences from (2) and (3) are direct output from ATS, without human editing. To guide this engineering of stimuli texts, we needed to know the *fluency* and *complexity* of each sentence in each source text so that sentences could be selected from each source to achieve a final text of a specific *fluency* and *complexity* level.

4.2.3 Annotations. We aligned each sentence from each original article to sentence(s) produced by ATS or human simplifications. In many cases, the ATS systems and human authors had split sentences, and thus multiple simplified sentences would align to a single original sentence. Then, two high-literacy native English speakers rated each original sentence and its possible replacements on 5-point Likert scales for grammaticality and complexity. To collect judgements of grammaticality, we used an item from prior work [31] that asked annotators to indicate their agreement with the statement: "This sentence is grammatically correct."

For judgements of complexity, we employed a scale in which disagreement indicated harder to read texts, using an item from prior work: "This sentence is easy to read." [3, 27, 28, 38]. For clarity, we reverse-scored annotators' responses for this particular item, using 5 for Strongly Disagree and 1 for Strongly Agree, so that we may more intuitively refer to the resulting score as *complexity*.

Our annotators' average judgements were 1.6 for *complexity* and 4.5 for *fluency* for all sentences in Original texts; 2.8 for *complexity* and 3 for *fluency* for all sentences in Hybrid texts; 2.5 for *complexity* and 3.2 for *fluency* for all sentences in Transformer texts; and finally, 1.1 for *complexity* and 4.8 for *fluency* for sentences in Newsela texts.

Finally, a DHH literacy expert[2] judged a subset of 24 randomly-selected sentences from the six articles. We computed inter-rater reliability on that subset to determine the agreement in the ratings of the expert and our annotators. We computed Krippendorf's alpha for these ordinal data, obtaining a moderate alpha value of 0.554.

4.2.4 Generation. For each sentence, we had four possible sources to use. Thus, the set of all possible texts that could have been generated would be combinatorially numerous. For instance, given that for each sentence, we had four options (the original, the two obtained from ATS, and one human simplification), an article with 26 sentences could generate 4^{26} (281,474,976,710,656) combinations. Thus, instead of trying to generate all possible options, we created a Python script to execute a top-N greedy algorithm, as follows:

(1) First, it considers the annotators' judgements for each original sentence and its three possible replacements.
(2) For each sentence, it selects the best of the four options, such that the *fluency* and *complexity* of the overall article remains closest to the desired level. For instance, if attempting to create a stimulus with low *fluency* but medium *complexity*, it favors local choices to achieve this result.
(3) It outputs the top-N articles closest to the desired levels of *fluency* and *complexity* (after identifying an article, it backtracks until N articles are obtained).

We identified articles with average levels of *complexity* of 2.0, but average levels of *fluency* of: 3.5 for our *low fluency* condition, 4 for our *medium fluency* condition, and 4.5 for our *high fluency* condition. Table 1 shows an excerpt illustrating each *fluency* condition.

4.2.5 Validation. To ensure that our text stimuli met our conditions, we conducted an expert review with our team's DHH literacy expert. The expert considered the three versions of each article labeled with generic labels (A, B, C), and ranked them by *fluency* from the perspective of how an average DHH reader would interpret the texts. The expert ranked all of the articles in the same order as ranked in the generation step, providing further confidence that these articles were indeed at the three different levels of *fluency*.

4.3 Metrics

To identify which metrics could distinguish texts engineered to be at different levels of *fluency*, we compared several metrics:

4.3.1 Reading Speed. This metric has been used in prior work measuring readability when comparing the impact of user-interface elements, e.g. [17, 29], and evaluating ATS systems, e.g. [26]. The idea is that more readable texts leads to higher reading speed. Prior methodological work on evaluating the *complexity* of texts among DHH adults did not identify reading speed as an effective metric [3], but it is unclear if it can distinguish texts at different *fluency* levels. We measured reading speed in words per minute (wpm), i.e. the number of words in a text over the minutes taken to read it.

4.3.2 Comprehension Questions. We obtained comprehension questions from prior methodological work on how to evaluate the *complexity* of ATS output among DHH adults [3]. Researchers in that study had written two versions of each multiple choice question, at two different levels of linguistic complexity; each question asked about the same fact, but the wording of the question item and answer choices varied in their linguistic complexity. Following the approach of [3], we created a quiz for each participant that selected a low- *or* high-linguistic complexity question for each fact, so that

[2]This co-author is a university professor of Deaf and Hard-of-Hearing (DHH) literacy and education who has published over 10 journal articles in DHH literacy venues, and worked as a teacher of DHH students for several years before pursuing a PhD.

Fluency Level	Excerpt
Low	"Much of this has been swallowed up by agriculture, there is still much land," said farmers who don't like agriculture. "We are committed to continuing to look for this small, elusive lizard, elusive and lizards."
Medium	"Much of this has been swallowed up by agriculture, there is still much land," said farmers who don't like agriculture. "We are committed to continuing to look for this small, elusive and cryptic lizard."
High	He added that much of this land has been taken up by agriculture, but there is still a lot more land to survey. "We are committed to continuing to look for this small, elusive and cryptic lizard."

Table 1: Excerpt from each condition, illustrating different levels of fluency.

the quiz contained 3 of each level. This controls the difficulty level of responding to a question about a specific fact, while enabling the study to examine the efficacy of comprehension questions at different levels of linguistic-complexity in their wording.

4.3.3 Score Prediction. This subjective-response item asked participants to predict how well they had done on the comprehension questions, by providing a numerical value from 0% to 100%. This had been used in prior work on whether reading comprehension can predict academic achievement [36]. In prior methodological work on evaluating the *complexity* of texts among DHH adults, this item had not been able to distinguish among texts' *complexity* levels [3]. However, it is unclear whether this item might distinguish between different levels of *fluency* among DHH readers.

4.3.4 Likert Subjective Judgements. Four subjective items used a 5-point scale: "Strongly Disagree" to "Strongly Agree."

- **Readability.** Judgements of readability ("This text was easy to read") had been widely used in prior work, e.g., [2, 26, 31], including methodological work on how to evaluate text *complexity* with DHH adults [3].
- **Understandability.** Judgements of understandability, which read "I was able to understand this text well." had also been used in prior methodological work on evaluating text *complexity* with DHH readers [3].
- **Grammaticality.** Judgements of grammaticality had been used in prior work on evaluating the *fluency* of ATS output with high-literacy readers, e.g., [31, 39], but they had not been validated with DHH readers. We used an item from prior work [31]: "This sentence is grammatically correct."
- **System Performance.** Prior work on evaluating ASR-based video captions had asked DHH adults whether the ASR had done a "good job" [5]. Thus, we included an adapted version: "The automatic text simplification system did a good job simplifying this news story."

4.4 Data Collection Procedure

Participants completed an informed consent form for our IRB-approved study and met over Zoom with a research assistant fluent in ASL. Participants read articles sequentially on a website built using the jsPsych library [8]. Articles were counterbalanced using a Graeco-Latin-Square schedule, which rotated the order of the articles and the order of the *fluency* conditions. Each participant read the six articles, with two of each condition.

After each article, participants responded to 6 comprehension questions as a single quiz containing 3 low- and high-linguistic complexity questions as described above. Participants then predicted their scores on the comprehension quiz (0% to 100%) and responded to the other subjective Likert-scale items (understandability, readability, grammaticality, and system performance). Participants were given the option to rest after the third article to avoid fatigue.

At the end of the study, participants filled out the sentence comprehension sub-test of the Wide-Range Achievement Test (WRAT), which has been validated to measure literacy levels for DHH readers [15, 25]. Finally, participants completed a demographic questionnaire and were compensated with $40 USD for their participation.

4.5 Participants

A total of 29 DHH participants were recruited through social media. Participants' average age was 25.6 (range 18 to 35, SD = 5.4). Ten participants identified as culturally Deaf[3], 6 identified as deaf, 12 as hard-of-hearing, and one as Deaf-blind (who indicated that they were able to adjust the font size of the text on their web browser to read it comfortably during the study.) Participants self-identified as female (N = 21), male (N = 7), and one preferred not to say.

Participants' average WRAT scores were 94.6 (range 67 to 128, SD = 16.8, higher score indicated higher literacy). To compare the effectiveness of the metrics for different literacy groups among our participant pool, and to investigate potential literacy bias of our metrics, we split our participants into two groups based on their median WRAT score (93) following the approach of [3]:

- **WRAT-H**: 13 participants with scores higher than 93.
- **WRAT-L**: 16 participants with scores of 93 or lower.

5 RESULTS

Table 2 summarizes the results for each metric in terms of: **discriminative ability** (its effectiveness in measuring differences among the three *fluency* levels of text) and **literacy bias** (whether scores were overall higher/lower for a particular literacy group). These items correspond to the two Hypotheses presented in section 3.

In regard to discriminative ability, both reading speed and judgements of grammaticality were effective at measuring differences between some *fluency* levels with WRAT-H and WRAT-L readers. However, judgements of understandability and readability were only effective among the WRAT-H group. No other metrics were effective at measuring differences among the *fluency* levels of texts.

[3]Deaf with capital "D" refers to people who identify as members of Deaf culture [22].

Metric	Discriminative Ability among Lower Literacy DHH Respondents (H1a)	Discriminative Ability among Higher Literacy DHH Respondents (H1b)	Literacy Bias (H2)
Reading speed (Best Metric)	**H1a was partially supported**. Worked well to distinguish the lowest fluency texts from both the medium and highest fluency texts.	**H1b was partially supported**. Worked well to distinguish the lowest fluency texts from both the medium and highest fluency texts.	**H2 was not supported**. There were no measurable differences between lower and higher literacy readers.
High-linguistic-complexity comprehension questions	**H1a was not supported**. This metric was not discriminative between any text fluency levels.	**H1b was not supported**. This metric was not discriminative between any text fluency levels.	**H2 was supported**. Higher literacy readers had significantly higher scores than lower literacy readers.
Low-linguistic-complexity comprehension questions	**H1a was not supported**. This metric was not discriminative between any text fluency levels.	**H1b was not supported**. This metric was not discriminative between any text fluency levels.	**H2 was supported**. Higher literacy readers had significantly higher scores than lower literacy readers.
Score prediction	**H1a was not supported**. This metric was not discriminative between any text fluency levels.	**H1b was not supported**. This metric was not discriminative between any text fluency levels.	**H2 was not supported**. There were no measurable differences between lower and higher literacy readers.
Understandability "I was able to understand this text well"	**H1a was not supported**. This metric was not discriminative between any text fluency levels.	**H1b was partially supported**. Worked well to distinguish between the lowest and highest text fluency only.	**H2 was supported**. Higher literacy readers had significantly higher judgements than lower literacy readers.
Readability "This text was easy to read."	**H1a was not supported**. This metric was not discriminative between any text fluency levels.	**H1b was partially supported**. Worked well to distinguish the high-fluency texts from both the medium and low-fluency texts.	**H2 was not supported**. There were no measurable differences between lower and higher literacy readers.
Grammaticality "This text was grammatically correct."	**H1a was partially supported**. Worked well to distinguish between the lowest and highest text fluency only.	**H1b was partially supported**. Worked well to distinguish between the lowest and highest text fluency only.	**H2 was not supported**. There were no measurable differences between lower and higher literacy readers.
System performance "The tool did a good job simplifying the news story."	**H1a was not supported**. This metric was not discriminative between any text fluency levels.	**H1b was not supported**. This metric was not discriminative between any text fluency levels.	**H2 was not supported**. There were no measurable differences between lower and higher literacy readers.

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

Our analysis revealed literacy bias for understandability judgements and comprehension questions (at both levels of linguistic complexity of questions). WRAT-H readers gave higher understandability judgements, and achieved higher scores on comprehension questions. To determine whether other metrics had statistically equivalent response scores, when comparing WRAT-H and WRAT-L readers, we conducted Two One-Sided Tests (TOST), which revealed statistically equivalent responses for: reading speed (within a margin of 74 wpm), score predictions (within a margin of 15 percentage points), and judgements of readability, grammaticality, and system performance (within a margin of 0.75 for all three).

Table 2 may sufficiently summarize the results for many readers, but the following subsections provide detailed results. We first present the results for Hypothesis 1 (the metrics' discriminative ability). Figures 1 through 4 accompany significant results, and include whisker-plots for continuous data and stacked divergent bar charts[4] for Likert-type data, separated by literacy group. We then present the results for Hypothesis 2 (the metrics' literacy bias), accompanied by Figures 5 through 7, which include the same type of plots as above, but compare the groups WRAT-L and WRAT-H.

5.1 H1: Discriminative Ability

5.1.1 Reading Speed (wpm). The results from the Kruskal-Wallis test revealed significant differences in reading speed between the conditions for both groups (WRAT-L: χ^2 = 18.707, df = 2, p-value < 0.001; WRAT-H: χ^2 = 27.41, df = 2, p-value < 0.001). Pairwise comparisons revealed statistically significant differences for both the WRAT-L and the WRAT-H group between the low and medium *fluency* conditions (p-value < 0.001 for both groups), and between the low and high *fluency* conditions (p-value < 0.001 for both groups).

5.1.2 Low-Complexity Comprehension Questions. There were no statistically significant differences revealed by the analysis between the conditions for either group (p-value = 0.895 for WRAT-L, and p-value = 0.858 for WRAT-H).

5.1.3 High-Complexity Comprehension Questions. No statistically significant differences were revealed for the high-complexity comprehension questions (p-value = 0.425 for WRAT-L, and p-value = 0.176 for WRAT-H).

5.1.4 Score Prediction. No statistically significant differences were revealed for the score predictions with either group (p-value = 0.326 for WRAT-L, and p-value = 0.422 for WRAT-H).

5.1.5 Understandability. The judgements of understandability only revealed significant differences between the *fluency* conditions for the WRAT-H group (χ^2 = 7.751, df = 2, p-value = 0.02). Pairwise comparisons revealed differences between the low and high *fluency* conditions (p-value = 0.016).

5.1.6 Readability. As illustrated in Figure 3, judgements of readability also revealed statistically significant differences between

[4]Stacked divergent bar charts are recommended to display Likert-type data [30]. Bars indicate each responses' percentage and are centered on the "neutral" response item.

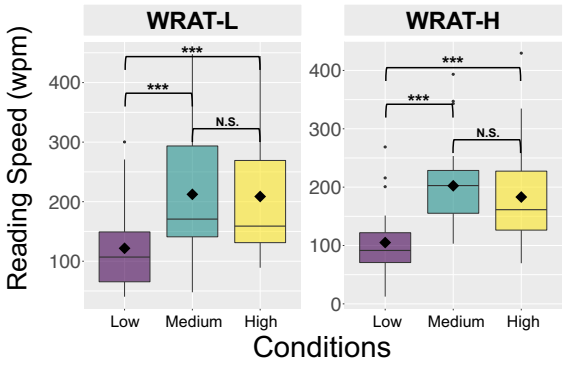

Figure 1: Reading speeds for H1, measured in words per minute (*** = p < 0.001).

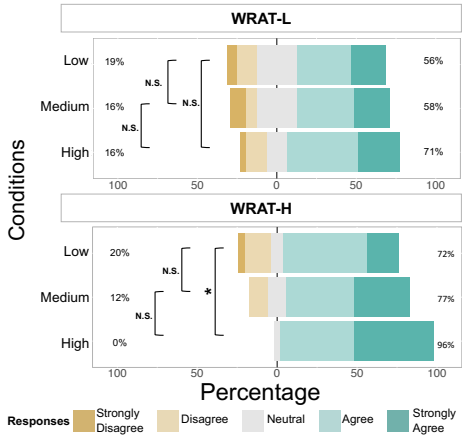

Figure 2: Understandability judgements for H1 using a Likert-type agreement scale (* = p < 0.05).

the *fluency* conditions for the WRAT-H group (χ^2 = 7.932, df = 2, p-value = 0.019), with pairwise comparisons revealing differences between the low and high *fluency* conditions (p-value = 0.045) and the medium and high *fluency* conditions (p-value = 0.041).

5.1.7 Grammaticality. Judgements of whether the text were grammatically correct revealed significant differences for both groups (WRAT-L: χ^2 = 11.482, df = 2, p-value = 0.003; WRAT-H: χ^2 = 13.355, df = 2, p-value = 0.001). For both groups, pairwise comparisons revealed differences between the low and high *fluency* conditions (p-value = 0.004 for WRAT-L, and p-value < 0.001 for WRAT-H). Figure 4 summarizes these results.

5.1.8 System Performance. No statistically significant differences were revealed for either group for system performance judgements (p-value = 0.191 for WRAT-L, and p-value = 0.058 for WRAT-H).

5.2 H2: Literacy bias

H2 was only supported for some metrics. In the cases in which it was supported, participants in the WRAT-H scored higher overall (in the case of comprehension questions) or provided higher judgements

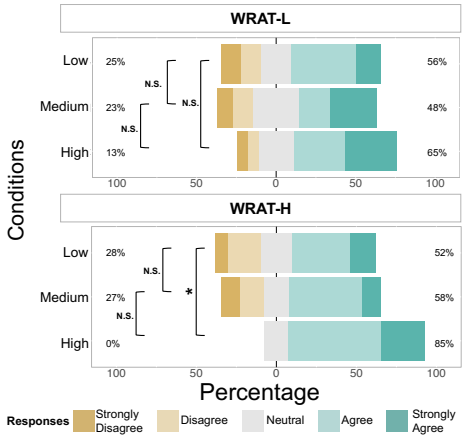

Figure 3: Readability judgements for H1 using a Likert-type agreement scale (* = p < 0.05).

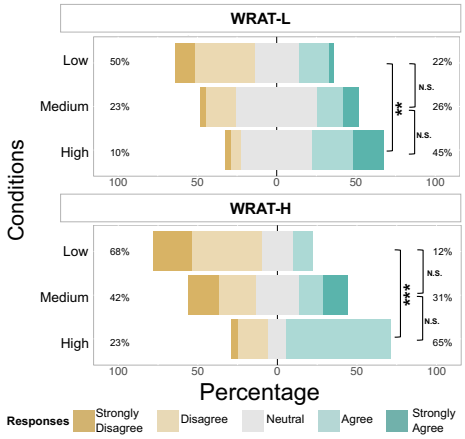

Figure 4: Grammaticality judgements for H1 using a Likert-type agreement scale (** = p < 0.01, *** = p < 0.001).

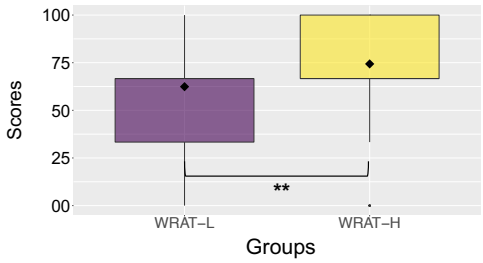

Figure 5: Low-linguistic complexity comprehension questions scores for H2, with a max. value of 100% (** = p < 0.01).

(of understandability). The metrics for which H2 was not supported all passed TOST equivalence tests, at the alpha=0.05 level.

- **Reading Speed (words per minute):** Z-score = -0.506, p-value = 0.61. TOST equivalence testing revealed no significant difference (with a margin of 74 words per minute).

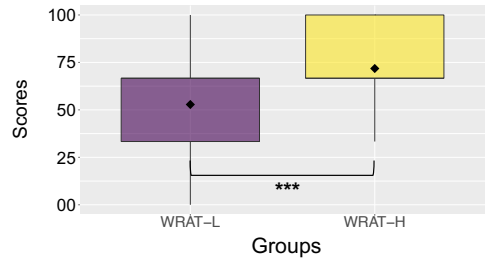

Figure 6: High-linguistic complexity comprehension questions scores for H2, with a max. value of 100% (*=p<0.001).**

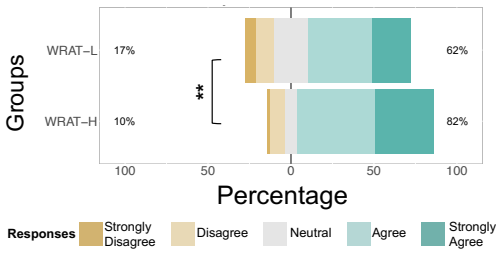

Figure 7: Understandability judgements for H2 using a Likert-type agreement scale (** = p < 0.01).

- **Low-Complexity Comprehension Questions:** Z-score = -2.994, p-value = 0.003 (Figure 5)
- **High-Complexity Comprehension Questions:** Z-score = -3.964, p-value < 0.001 (Figure 6)
- **Score Prediction:** Z-score = -1.59, p-value = 0.11. TOST equivalence testing revealed no significant difference (with a margin of 15 on a 0 to 100 scale).
- **Understandability:** Z-score = -2.627, p-value = 0.009 (Figure 7)
- **Readability:** Z-score = -0.21, p-value = 0.8. TOST equivalence testing revealed no significant difference (margin 0.75 on a 1 to 5 scale).
- **Grammaticality:** Z-score = -1.296, p-value = 0.19. TOST equivalence testing revealed no significant difference (with a margin of 0.75 on a 1 to 5 scale).
- **System Performance:** Z-score = -0.381, p-value = 0.7. TOST equivalence testing revealed no significant difference (with a margin of 0.75 on a 1 to 5 scale).

6 DISCUSSION

In the following subsections, we first discuss the results of the discriminative ability of the metrics (H1), followed by the discussion of the results of the literacy bias of the metrics (H2).

6.1 H1: Discriminative Ability

The best metric overall was reading speed; it was able to measure differences between more text *fluency* levels and worked with both WRAT-H and WRAT-L participants. To the best of our knowledge, reading speed had only been used in prior work for measuring reading comprehension, and in prior methodological work on evaluating the *complexity* of simplified texts with DHH readers, this metric had not been an effective way to measure text *complexity* [3]. This lack of use in prior work for evaluating *fluency* may come from the fact that most evaluations of *fluency* have relied upon an expert reader making side-by-side comparisons between original and simplified texts, and measuring reading speed is less suitable when a participant is making such side-by-side comparisons. Our findings suggest that displaying the output text to a DHH participant and measuring their reading speed is also an effective way of measuring the *fluency* of that text.

The other objective metrics in our study were the comprehension questions written at different complexity levels. However, given that these were not able to distinguish between any of the *fluency* conditions with either one of the groups, we do not recommend their use for evaluating the *fluency* of simplified texts. This result is in line with prior methodological work on evaluating the *complexity* of simplified texts [3], in which comprehension questions had only worked under a number of conditions (namely, that the texts were far enough in *complexity*, that the group had lower literacy, and that the questions were written in low complexity). As highlighted in that prior study, comprehension questions may have value in keeping participants engaged with the reading task, even if they are not useful as metrics for evaluation of the text itself. In this case, we found no evidence of them being effective in distinguishing between the *fluency* levels of texts included in our study.

Among the subjective metrics, judgements of grammaticality were the most effective, revealing the difference between the low and the high *fluency* conditions of text. This metric may be used in cases where reading speed may not be available, e.g., if it is not feasible for a researcher to capture reading time due to their study setup. Among the remaining subjective metrics, we found that judgements of readability and understandability were only effective with the WRAT-H participants. Thus, in our study, we found that a greater number of our metrics were effective among higher-literacy participants, as compared to the number of metrics that had been effective among lower-literacy participants. These results are in line with prior methodological work on evaluating the quality of automatic video-captioning tools with DHH readers [5]; in that prior study, more of the caption-quality metrics that researchers had investigated were effective among their higher-literacy readers. We speculate that this may be because these judgements (of understandability and readability) require higher metacognitive awareness when disfluencies are introduced in the text—and are therefore more suitable among higher-literacy readers.

In summary, future researchers who wish to use the methodological findings of our study should utilize reading speed to measure *fluency* of ATS texts among DHH readers. As an additional or alternative measurement, we secondarily recommend the use of a Likert-scale subjective judgement of the grammaticality of the text.

6.2 H2: Literacy Bias

As discussed previously, finding that a metric has a literacy bias does not necessarily mean that it is undesirable to use it within a study. It simply means that if a researcher were to use such a metric, then they should also report the literacy levels of the participants

in their study, e.g., WRAT scores. Without information about the literacy level of the specific participants in a study included in a publication, it would not be possible for readers of that paper to compare the results to other published work, since the literacy characteristics of the participants may have influenced the scores.

Our study revealed that three metrics had a literacy bias: Likert-scale subjective judgements about the understandability of a text, comprehension questions written at a higher complexity level, and comprehension questions written at a lower complexity level. Since none of these three metrics had actually been effective in measuring text *fluency* in our study, our finding of literacy bias for these three metrics may be moot for researchers interested in measuring text *fluency*. However, since comprehension questions and subjective judgements of text *complexity* are used in other evaluation contexts, our findings may be of interest to such researchers. We speculate that the higher literacy of WRAT-H readers made it easier for them to understand texts and to answer comprehension questions.

For the remaining metrics (reading speed, score prediction, readability and grammaticality judgements, and judgements of whether the tool had done a good job), we did not find evidence of a literacy bias, i.e., there were no statistically significant differences between the responses from WRAT-H and WRAT-L readers. TOST confirmed that responses were statistically equivalent, within margins. This finding should be interpreted with caution: Prior work on evaluating the *complexity* of simplified texts had observed literacy biases in reading speed, score prediction, and readability judgements. In the general case, we believe that it is reasonable that literacy biases may exist for these metrics, e.g. that higher-literacy readers read more quickly than lower-literacy readers. Our findings should be interpreted more narrowly: In the context of a study in which participants were asked to read texts that contained dis-fluencies introduced by ATS technology, our findings suggest that literacy bias for these metrics was less substantial. From the perspective of a researcher who is only evaluating *fluency* of texts, it may be less necessary to report the literacy level of DHH participants in the study. However, reporting of participant literacy level is generally recommended, especially when evaluating the *complexity* of the texts, as found in prior work [3].

7 LIMITATIONS AND FUTURE WORK

There were several limitations in our study and several possible avenues for future work. First, we cannot guarantee that some of the metrics found to be ineffective in our study would not work in studies with larger sample sizes and more statistical power. Of course, since evaluations of ATS typically involve *fewer participants* than the number included in our current methodological study, our contribution is still useful to researchers. Namely, metrics that did not reveal significant differences in our study would be unlikely to reveal significant differences in evaluations with an equal or smaller sample size. Future work, however, could benefit from employing participants across broader ranges of literacy levels. Furthermore, our study only included news stories about science-related topics which may have affected the effectiveness of some of our metrics. For instance, while we carefully controlled the stimuli to be at the same level of *complexity*, the level we employed may have prevented some metrics from displaying significant differences among our lower literacy group. Thus, future work could explore whether our findings are generalizable to other text domains and controlled to be at lower levels of *complexity*.

In this study, we only evaluated one type of comprehension questions, namely, multiple-choice questions. Thus, our findings may not generalize to other types of comprehension questions, e.g., cloze tests or summarizing tasks. Further, while we did not recruit participants with fluency in ASL specifically, future work could also incorporate comprehension questions recorded in ASL as a possible metric for use among participants who are ASL signers. Other methods of evaluation, such as eye-tracking, could be explored during in-person studies, which were not possible at the time this study was conducted due to COVID-19 restrictions.

Our current study specifically focused on how to evaluate the *fluency* of automatically simplified texts, which may be damaged as a result of grammatical errors being introduced. However, as mentioned during the related work, semantic errors may also be introduced in the simplification process, affecting the *faithfulness* of the simplified texts. As *faithfulness* is typically evaluated by asking expert readers to examine the original and simplified texts side-by-side, future work could explore whether this type of evaluation is possible among DHH readers, or whether some of the metrics employed in our study (e.g., comprehension questions) can effectively distinguish varying levels of *faithfulness*.

Finally, it is our hope that future work will include DHH users in actual evaluations of the fluency of the output of ATS technologies. Furthermore, future work can include explorations of DHH readers' preferences among various interface parameters of ATS-based reading assistance tools, their specific linguistic needs and how ATS tools can be adapted to better support those.

8 CONCLUSION

In this study, we conducted methodological research to evaluate several metrics in terms of their effectiveness for evaluating the fluency of automatically simplified texts among DHH adults across a range of English literacy levels, and potential literacy biases when using those metrics. Our findings revealed that reading speed and participants' subjective judgements of grammaticality were effective for distinguishing text fluency levels among DHH participants across a range of literacy levels. Judgements of understandability and readability were only effective among participants with higher literacy. Literacy biases were only observed in comprehension questions scores and judgements of understandability, i.e., participants with higher literacy had more positive scores and judgements overall. Our findings provide methodological guidance for the design of studies with DHH participants evaluating ATS technologies. Namely, we recommend the use of reading speed and judgements of grammaticality when evaluating the fluency of simplified texts among DHH adults at a range of literacy levels.

ACKNOWLEDGMENTS

We thank the research assistants who assisted us with annotations and recruitment for this study, and Wei Xu and Mounica Maddela for providing us with automatic simplifications for our stimuli-building process. This material is based upon work supported by the National Science Foundation under Grant No. 1822747.

REFERENCES

[1] 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 pages. https://doi.org/10.1145/3373625.3416992

[2] 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

[3] 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.* Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445038

[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] Joachim Bingel, Gustavo Paetzold, and Anders Søgaard. 2018. Lexi: A tool for adaptive, personalized text simplification. In *Proceedings of the 27th Int'l Conference on Computational Linguistics.* Association for Computational Linguistics, Santa Fe, New Mexico, USA, 245–258. https://www.aclweb.org/anthology/C18-1021

[7] 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.

[8] 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.

[9] Ashwin Devaraj, Iain J Marshall, Byron C Wallace, and Junyi Jessy Li. 2021. Paragraph-level Simplification of Medical Texts. *arXiv:2104.05767* (2021).

[10] 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

[11] Saoradh Favier and Falk Huettig. 2021. Long-term written language experience affects grammaticality judgements and usage but not priming of spoken sentences. *Quarterly Journal of Experimental Psychology* 74, 8 (2021), 1378–1395. https://doi.org/10.1177/17470218211005228 arXiv:https://doi.org/10.1177/17470218211005228 PMID: 33719762.

[12] 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.

[13] 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

[14] Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF Model for Sentence Alignment in Text Simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* 7943–7960.

[15] Lynda J Katz and Franklin C Brown. 2019. Aptitude and achievement testing. In *Handbook of Psychological Assessment.* Elsevier, 143–168.

[16] 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.

[17] 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/

[18] Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. Controllable Text Simplification with Explicit Paraphrasing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* 3536–3553.

[19] Mounica Maddela and Wei Xu. 2018. A Word-Complexity Lexicon and A Neural Readability Ranking Model for Lexical Simplification. In *Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing.* Assoc. for Computational Linguistics, Brussels, Belgium, 3749–3760. https://doi.org/10.18653/v1/D18-1410

[20] Marc Marschark, John A. Albertini, and Harry G. Lang. 2002. *Educating deaf students: from research to practice.* Oxford University Press.

[21] Ross E. Mitchell. 2005. How Many Deaf People Are There in the United States? Estimates From the Survey of Income and Program Participation. *The Journal of Deaf Studies and Deaf Education* 11, 1 (09 2005), 112–119. https://doi.org/10.1093/deafed/enj004 arXiv:https://academic.oup.com/jdsde/article-pdf/11/1/112/1143760/enj004.pdf

[22] Carol Padden, Tom Humphries, and Carol Padden. 2009. *Inside deaf culture.* Harvard University Press.

[23] 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

[24] 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

[25] 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.

[26] 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

[27] 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

[28] 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, CA, USA) *(CHI '16)*. ACM, New York, NY, USA, 3637–3648. https://doi.org/10.1145/2858036.2858204

[29] 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.

[30] 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.

[31] 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

[32] 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

[33] Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications* 4, 1 (2014), 58–70.

[34] 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

[35] 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

[36] 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 arXiv:https://doi.org/10.1080/08856257.2018.1435010

[37] W. M. Watanabe, A. Candido Jr., M. A. Amâncio, M. De Oliveira, T. A. S. Pardo, R. P. M. Fortes, and S. M. Aluísio. 2010. Adapting Web content for low-literacy readers by using lexical elaboration and named entities labeling. *New Review of Hypermedia and Multimedia* 16, 3 (2010), 303–327. https://doi.org/10.1080/13614568.2010.542620 arXiv:https://doi.org/10.1080/13614568.2010.542620

[38] 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

[39] Xingxing Zhang and Mirella Lapata. 2017. Sentence Simplification with Deep Reinforcement Learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, Copenhagen, Denmark, 584–594. https://doi.org/10.18653/v1/D17-1062