Comparing Evaluation Techniques for Text Readability Software for Adults with Intellectual Disabilities

Matt Huenerfauth
City University of New York (CUNY)
Queens College & Graduate Center
65-30 Kissena Blvd, Flushing, NY
+1-718-997-3264
matt@cs.qc.cuny.edu

Lijun Feng
City University of New York (CUNY)
Graduate Center
365 Fifth Avenue, New York, NY
+1-212-817-8190
lijun7.feng@gmail.com

Noémie Elhadad
Columbia University
Dept. of Biomedical Informatics
622 West 168th Street, New York, NY
+1-212-305-5780
noemie@dbmi.columbia.edu

ABSTRACT

In this paper, we compare alternative techniques for evaluating a software system for simplifying the readability of texts for adults with mild intellectual disabilities (ID). We introduce our research on the development of software to automatically simplify news articles, display them, and read them aloud for adults with ID. Using a Wizard-of-Oz prototype, we conducted experiments with a group of adults with ID to test alternative formats of questions to measure comprehension of the information in the news articles. We have found that some forms of questions work well at measuring the difficulty level of a text: multiple-choice questions with three answer choices, each illustrated with clip-art or a photo. Some types of questions do a poor job: yes/no questions and Likert-scale questions in which participants report their perception of the text's difficulty level. Our findings inform the design of future evaluation studies of computational linguistic software for adults with ID; this study may also be of interest to researchers conducting usability studies or other surveys with adults with ID.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *language generation, machine translation*; K.4.2 [**Computers and Society**]: Social Issues – *assistive technologies for persons with disabilities*.

General Terms

Design, Experimentation, Human Factors, Measurement.

Keywords

Assistive Technology, Intellectual Disabilities, Natural Language Processing, Text Readability Assessment, Text Comprehension.

1. USERS & MOTIVATING APPLICATION

People with cognitive disabilities are diverse and may include people with intellectual disabilities (ID), dementia, autism, stroke, brain injury, or other impairments. Our goal is to create tools that benefit people with ID, specifically those classified in the "mild level" of mental retardation, IQ 55-70. About 5% of the U.S. population has intelligence test scores of 70 or lower [31].

People with ID face challenges understanding written and spoken information. Most read below their own mental age-level [19], and their difficulties differ from other low-literacy adults. World knowledge and working memory are key resources for language comprehension, and these individuals often struggle to infer and remember information from text [12]. An important distinction between people with ID and other adults is their speed of semantic encoding [25, 15]. Due to slower encoding speed, units are lost from the working memory before they are processed. These individuals also have trouble building cohesive representations of discourse [15]: as less information is integrated into the mental representation of the current discourse, less is comprehended.

Adults with ID are limited in their choice of reading material. Most texts that they can readily understand are targeted at the level of readability of children. However, the topics of these texts often fail to match their interests since they are meant for younger readers. Because of the mismatch between their literacy and their interests, users may not read for enjoyment and thus miss valuable reading practice time. To make reading material of interest more accessible, our long-term research goal is to build a web-based automatic system that can simplify selected text for people with ID and read it aloud using text-to-speech technology. Specifically, our goal is to give these users access to local news information.

Despite our interest in a read-aloud application, this paper uses the term "readability" when discussing the difficulty level or complexity of a text. We do this to remain in line with the use of this term by researchers working on related problems in computational linguistics (section 2.1). Most researchers studying automatic tools for measuring or improving texts have focused on readers without disabilities, who do *read* the text for themselves. While we also use the term "readability," it may be more appropriate to say that we are interested in the complexity of a text as displayed and read-aloud to adults with mild ID.

1.1 Related Work

While the creation of software for automatically simplifying and presenting articles to adults with ID is novel, other researchers have studied the syntactic *simplification* of texts for people with aphasia [1] or *generation* of texts for people with low-literacy [32]. (In language generation, the input to the system is a symbolic encoding of the information to be expressed; in simplification, the input is an already-written text that must be modified.) Researchers have also studied the design of various technologies for people with cognitive disabilities: portable reading systems [5], reminder systems [23], navigation [11],

ASSETS'09, October 25–28, 2009, Pittsburgh, Pennsylvania, USA.
Copyright 2009 ACM 978-1-60558-558-1/09/10...$10.00.

handheld devices [7], and web browsers [4, 28], to name a few. While our initial focus is on the underlying linguistic technology, we anticipate this related research to provide an important foundation for the HCI design aspects of our overall system.

Some researchers discuss issues that arise when conducting usability experiments with these users [4, 28, 22]. Some recruited participants though agencies for adults with ID [4], an approach we have used. Others discuss how task lists, time-limited interviews, and other common methods in usability studies are not suitable for adults with ID [22]. Assistive technology research for these users has included various forms of evaluation: interviews with experts [4], expert evaluations of prototypes [22], surveys on computer use [9], or ethnographic interviews with users [7]. While this research provides a valuable foundation for designing evaluation studies with these users, it does not address the specific issue of how to best measure text comprehension by these users, which is the focus of this paper.

1.2 A Pilot Study with Adults with ID

We conducted a pilot study with 14 adults with mild ID from two day-habilitation programs in New York City. After obtaining IRB approval, we recruited from programs with clinical screening criteria; so, our study did not contain intellectual screening tests. Participants were financially compensated for their time. This interest/satisfaction study used a Wizard-of-Oz prototype of a text-simplification system to examine whether adults with mild ID would respond favorably to a system that automatically simplified news articles and read them aloud. A human editor manually simplified 10 news stories, simulating sets of simplification operations that are reasonably within the future state-of-the-art of computational linguistics (details in section 3.2). TTSReader [29] was used to display a text buffer on the screen, read the text aloud, and highlight each word as it is read. User reaction was positive to the prototype. Ten participants said they would like to use such software at their day program, and two said yes if they had more practice. (Thirteen participants reported using a computer at least once a week at their day program.) Comments included: "It's interesting. Because I can read articles," "It's nice to see the text and hear it," and "I liked the voice and how the words light up."

During the pilot study, participants saw a variety of news stories. Stories about topics relevant to daily life, such as a new subway line in Manhattan, elicited much interest. Participants were also asked what they enjoyed learning or reading. Most reported watching the local news on TV (other interests mentioned: sports, news about celebrities, comic books, weather, and the Bible). For our research, we have therefore decided to focus on the domain and genre of local news stories. Community inclusion is a goal for many of our users, and providing them with stories in language they understand easily may aid them in becoming more aware of and part of their communities. Providing texts that are more interesting and readable is also likely to encourage users to read more. While some initiatives provide online information about government programs in simplified language for people with ID [30], it is not practical for human editors to simplify news sources that are frequently updated and specific to a limited geographic area (like local news). Therefore, automatic methods to process these texts and make them accessible are desirable.

During this pilot study, we also examined which forms of questions were best at measuring text comprehension with adults with mild ID. Some of the texts shown to participants were in their original (complex) form, and some texts were simplified by a human editor. Some participants saw the original form of an article, and others, the simplified form (no participant saw both versions). After each article, participants were asked five-point Likert-scale evaluation questions, e.g. "Was this story easy or hard to understand? Very easy, easy, medium, hard, very hard." Participants were also asked multiple-choice questions and yes/no questions about facts in the story. Some multiple-choice questions had long answer choices (full sentences), and others had shorter answer choices (single verb or noun phrase). As discussed in sections 2.3 and 2.4 below, none of the comprehension question types seemed effective at measuring how well participants understood texts of different levels of complexity.

2. MEASURING TEXT COMPREHENSION

In other venues, we have described the linguistic components of our system [10]. This paper is focused on how we can best *measure comprehension* of the information content of articles presented using our software for adults with ID. Laying the groundwork for how to conduct experimental evaluations of the understandability of the texts will allow us to track the progress of our system over time and compare it to alternative approaches to making textual information accessible for these users. This section will discuss five possible strategies for measuring text comprehension with these users: automatic readability scoring, timed reading, self-perception of readability with Likert-scales, objective evaluation questions, and expert evaluation. Some of these approaches have been examined in our pilot study (section 1.2) and in a new study presented in this paper (section 3).

2.1 Automatic Readability Scoring Systems

One method of evaluating the complexity level of a text is to design a computer program that can analyze various linguistic features of the text and assign a score to it automatically. If possible to build such an automatic readability scoring system (ARSS), then texts could be evaluated without the need to conduct experiments with human subjects. There are several traditional readability formulas that can assign a readability score to a text. The Flesch-Kincaid grade level formula [20] uses average sentence length and average syllables per word to calculate the "grade level" of a text. The New Dale-Chall readability formula consults a static, manually-built list of "easy" words to determine whether a text contains unfamiliar words [2]. These traditional metrics are widely used, especially in educational settings, partly because they are simple to calculate. However, these metrics do not always capture the reading complexity of a text accurately, and researchers have found that adapting texts in an effort to minimize these traditional metrics can actually reduce the true readability of the text [6]. They can easily misrepresent the complexity of technical texts [3] or reveal themselves un-adapted to a set of readers with particular reading difficulties. Given the simplistic factors (i.e. sentence length or syllable-count) used in these metrics, this is not surprising.

Recent work on ARSS has incorporated sophisticated natural language processing techniques, such as parsing and statistical language modeling, to capture more complex linguistic features and used statistical machine learning tools to build readability metrics [27, 14, 10, 26]. For instance, one metric built using a linear classifier based on unigram language models (models of whether certain words appear in a text) outperformed traditional readability metrics on predicting the readability of technical materials for general readers [27]. Information from language models and syntactic information about the texts (as determined by an automatic parser) were used to train a classifier to predict readability of texts for second-language learners [14]. Simple features from traditional readability metrics, language models, and automatic parser were used to train a support vector machine to identify elementary-school grade levels of texts [26].

A text's readability cannot be judged solely by the content, style, structure, and design of the text itself; it also depends on the reader. The reader's prior knowledge, reading skill, interest, and motivation may affect the readability of a text [13]. The metrics above have generally been targeted at general readers and do not model the particular reading challenges of adults with ID. To measure the ability of state-of-the-art ARSS to predict the difficulty of texts for adults with ID, we conducted a study [10] in which we implemented a readability metric by training a linear regression model using the set of linguistic features used by the state-of-the-art ARSS for readers without disabilities [26]. We used this metric and the popular Flesch-Kincaid grade level index to assign scores to a set of 20 texts. These texts were evaluated for difficulty for adults with ID, and we then measured the correlation of these metrics to the difficulty level of the texts. No significant correlations were measured [10].

An ARSS is needed that can consider the particular linguistic challenges of adults with ID [10] – taking into account our user's interests and their unique cognitive characteristics. For example, these users are better at decoding words than at comprehending text meaning [8]; so, shallow features like "syllable count per word" or models of word occurrence frequency may be less important indicators of reading difficulty. Many of the metrics discussed above have focused on the task of labeling texts with particular elementary school grade levels. Traditional grade levels may not be the ideal way to score texts to indicate how readable they are for adults with ID. Other related work has used models of vocabulary [14]: since we would like to use our tool to give adults with ID access to local news stories, we would prefer a metric that is more topic-independent.

A challenge for our users is to create a cohesive representation of discourse. Due to their impairments in semantic encoding speed, our users may have particular difficulty with texts that place more burden on working memory (items fall out of memory before they can be encoded). Thus, we have begun to experiment with developing an ARSS for adults with ID that considers a count of the number of entities (people, objects, etc.) discussed in a text, and we have had some promising preliminary findings that considering such features helps produce an ARSS that is better tailored to the difficulty of a text for adults with ID [10].

Despite some promising initial work, current ARSS technology has not yet reached a level of accuracy to be used as an evaluation tool for measuring the quality of text output for adults with ID.

2.2 Timed Reading

One method of measuring the difficulty of a text in an experimental setting is to ask a user to silently read the text and to time how long they need to read it. Such an approach was used in a pilot evaluation of the SkillSum system [32], a tool for automatically generating sentences to be read by low-literacy adults (not specifically those with ID). Unfortunately, timed reading proved ineffective at measuring text difficulty. Poor readers tended to skim-read or press the "finished" button without reading the entire text [32]. Even accurate measures of reading time would not directly measure what users understood. Further, for adults with ID, researchers conducting usability studies have argued that timed-task measures are inappropriate for these users, who often complete tasks at their own rate [22].

Another problem with using timed-reading with adults with ID is that most of these users cannot read non-trivial texts independently. Jones et al. [17] measured the literacy of adults with ID with IQ ranges 50-79 (our target users have IQ 55-70). They found that only a small fraction of their participants could understand important linguistic phenomena in a written text: 21% understood the difference between singular vs. plural nouns in a text, 42% singular vs. plural pronouns, 21% comparative vs. superlative adjectives, 11% reversible passive voice, 16% "in" vs. "on," 16% "X but not Y," etc. [17]. Without providing text-to-speech technology, our future news-simplification system would be too difficult to use (even with simplified text). Because some adults with mild ID are good readers, we will also present text onscreen with each word highlighted as it is read aloud to encourage users to read along. It is important to note that merely using text-to-speech software is insufficient for making news articles accessible to these users; linguistic transformations are required to make the text simpler and more understandable.

2.3 Readability Self-Perception, Likert Scales

Another way to measure a text's difficulty level is to ask research participants subjective questions in which they must evaluate how easy it is to understand a text. This *metacomprehension* of text has been measured using Likert scales in educational studies [24]. However, people with ID are known to struggle with metacomprehension awareness [33], and special training may be necessary for them to learn to evaluate this more reliably [33].

In our pilot study (section 1.2), participants appeared to struggle when answering these Likert-scale questions, and we did not find any significant correlations between participant's responses to these questions and the complexity level of the text. One explanation for this result is that our target users' cognitive impairments may make it difficult for them to form a subjective judgment about how difficult a text is to understand (or to perceive what fraction of the information content of the text they were able to understand). Another potential problem with the Likert scale questions in our pilot study is that the participants may have become overwhelmed with having five choices on the scale. For this reason, in our larger experimental study (section 3), we included three-point Likert-scale questions (e.g. "easy, medium, hard") to check if these were simpler for participants to answer – and more indicative of the complexity level of a text.

There is reason to believe that Likert-scale metacomprehension questions are a poor measure of the difficulty of a text for any users. Cognitive scientists generally recommend using actual comprehension tasks rather than relying on perceived comprehension [21]. In our earlier work on the comprehension of sign language animations by deaf adults (with no intellectual impairments) [16], we found that self-perception of story comprehension (reported on a Likert scale) had low correlation to more direct measures of comprehension, like objective questions.

2.4 Objective Comprehension Questions

We have discussed problems in measuring the readability of a text using metacomprehension (reported on a Likert scale) and timed reading. Therefore, we are left to consider more direct measures of the comprehension of information from a text: objective comprehension questions, such as multiple-choice or yes/no questions, both of which have been used with adults with ID in educational and psychological studies of comprehension [15].

In our pilot study (section 1.2), participants answered some question types (yes/no questions and multiple-choice questions with one-sentence answer choices) at a rate close to random chance (i.e. guessing) and with no significant difference in scores between simple and complex versions of articles. Upon learning of this result, several staff members who work at programs for adults with ID suggested that these users often agree to the first choice or say 'yes' when they do not understand a question. Multiple-choice questions with long answer choices may be difficult for our users to remember and consider at once, and yes/no questions have more content *in the question itself* than do multiple-choice questions (and may be harder to understand).

In our earlier research on measuring the comprehension of sign language animations by deaf adults (without ID), we used a different form of comprehension question: multiple-choice questions with clip-art used to illustrate answer choices [16]. Since we were using clip-art multiple-choice questions in other projects at CUNY, we wondered whether including clip-art or photos in our questions for adults with ID could allow us to better measure these users' comprehension of a text. There are different motivations for using clip-art questions with deaf users or users with ID. We decided to use clip-art questions with deaf users to minimize the amount of English in the experiment environment, which is important to do in sign language comprehension studies [16]. For adults with ID, we hypothesize that the use of clip-art images in the answer choices of multiple-choice questions may allow them to better understand the *question* – thus allow the question to better measure understanding of the text. Researchers working with adults with ID have confirmed that making use of symbols or images can increase comprehension [18].

2.5 Expert Evaluation of Texts

Another possible approach to measuring the complexity of a text for adults with ID is to do so in a more indirect manner. Instead of directly involving users with disabilities in the evaluation process, experts who are familiar with the reading/comprehension capabilities of these users could be asked to evaluate the texts. While we intend to explore this issue in future work (comparing expert opinions about text complexity to user comprehension scores of texts), the goal of the present paper is to identify the best methods for obtaining direct evaluation data from our target users.

3. OUR QUESTION COMPARISON STUDY

We conducted a study to compare alternative types of questions for evaluating the complexity of a text for adults with ID; in this study, we include multiple-choice questions with short answers, multiple-choice questions with clip-art answer choices, yes/no questions, and Likert-scale questions about the user's perception of the text's difficulty. The format of this study was similar to that of our pilot study (previously discussed in section 1.2); adults with ID were shown local news articles displayed onscreen and read aloud via text-to-speech software. Human editors produced simplified versions of the articles to make them easier to understand by adults with ID, and participants saw a mix of these 'Simple' and 'Complex' articles. The goal of this new study was to determine what forms of comprehension questions are best at measuring the complexity of a text for adults with ID.

For this study, we assume that there is a real difference between Simple and Complex stories – it is not a hypothesis being tested. What we care about is how well the different types of questions reveal this assumed-to-exist difference. We care about the *magnitude* of the Complex-Simple *difference* for each question type. The text simplification software system we plan on developing in future work would start with a Complex version of an article and modify the text to become closer to the style of the Simple articles in this experiment. Since a human produced the Simple version of the articles in this experiment, they may be better than what we'd expect a computer to produce. So, in the future, we may want to evaluate the reading difficulty of texts that are somewhere on this *range* from Complex to Simple. We want to find a question type that has the ability to reveal whether a text is closer to the Complex or the Simple end of this spectrum effectively. Our hypothesis in this study is that short-answer multiple-choice questions in which clip-art accompanies each answer choice would be more effective at distinguishing Simple and Complex stories than yes/no or Likert-scale questions.

3.1 Participant Recruitment, Demographics

As in our earlier pilot study, after obtaining IRB approval, this study was advertised to adults with mild ID who participate in day habilitation programs in New York City. The programs to which the study was advertised have their own clinical screening criteria, and so we were not required to perform intelligence-test screening. Twenty adults with ID participated in the study (8 women, 12 men); participants were ages 22 to 50. Some participants had difficulties focusing on the experimental task that day, and one had visual impairments that made it more difficult to see the onscreen text or clip-art answer choices (even after bigger font sizes and image enlargements were provided). Because some individuals were not able to effectively participate in the study, we therefore established an inclusion criterion: Data from participants whose accuracy on the comprehension questions was very low (within 10% of random guessing, i.e. 43.33%) were omitted from the study. Six (of 20) participants were omitted. By determining whom to include in the study based on overall performance (on all articles, Simple and Complex combined), we did not need to establish a policy for excluding participants for idiosyncratic reasons of inattentiveness or other participation difficulties.

3.2 Eleven Articles, Simple vs. Complex

Eleven articles were collected from online sources of local news: New York Daily News, New York Post, New York Times, and NBC New York News. Nine articles focused on various current events around the greater New York area, including some articles on transportation, local sports, and human-interest articles about people active in their communities. Two articles on entertainment news were also included (with less of a local focus).

Humans performed the simplifications on the articles, restricting themselves to operations on the texts that are within the near state-of-the-art of computational linguistic techniques. Operations included breaking apart complex sentences, un-embedding information in complex prepositional phrases and reintegrating it as separate sentences, replacing infrequent vocabulary items with more common equivalents, and omitting sentences and phrases from the story that mention entities and concepts extraneous to the main theme of the article. For instance, the original sentence "Bike theft and its cousin, bike abandonment, are aspects of urban cycling as sure as broken glass, potholes and the lurking threat of being doored." was transformed into "Bike theft and bike abandonment are aspects of urban cycling." Preliminary results of our linguistic research on ARSS for adults with ID suggest that the number of entities mentioned in a text is a useful factor to consider when assigning a readability score to a text [10]. We noticed that many of the news articles included a sentence or two in which a random 'person on the street' or 'expert from an organization' were quoted. Typically, this sentence was the only location in the text in which this person or organization was mentioned; such sentences were omitted from the articles.

Originally, the articles ranged from approximately 100 to 800 words in length, with most of the articles in the 300-400 word range. After simplification, they ranged from approximately 80 to 300 words, most 150-250 words. Because it was important for this study that there be a true difference in the complexity of these texts for adults with ID, a psychologist from a social-services agency for adults with ID was asked to review the simplified articles to check their level of difficulty for our target users.

3.3 Three Comprehension Question Types

In our pilot study, we tested several comprehension question types: yes/no questions, multiple-choice questions with one-sentence answers, multiple-choice questions with noun-phrase or verb-phrase answers, etc. The preliminary findings of that study indicated that long-answer multiple-choice and yes/no questions did a poor job at measuring participant's comprehension of a text. In this study, we planned on comparing three forms of comprehension questions (examples shown in Figure 1):

- MultipleChoice: These are multiple-choice questions in which the participant needs to select from three choices, each of which is a single word or a short phrase (not full sentences).
- ClipArt: These are the same question as the MultipleChoice (with the same answer choices given), but for each answer choice, a clip art image or a photograph was shown to illustrate that answer choice. English text captions were placed below each clip-art answer choice (the same text as the answer choices for the corresponding MultipleChoice question).
- TrueFalse: These questions are actually yes/no questions with three answer choices: Yes, No, It didn't say. The "It didn't say" option was included so that these questions would have the same number of answer choices as the others in the study.

Questions were shown to participants on paper, and a researcher read each question and each answer choice aloud (reading the English caption text for the ClipArt question answer choices).

3.4 Six Facts Per Article, Different Categories

Our goal is to find out which question type is best able to measure a difference in comprehension between the Simple and Complex versions of the articles. If questions focus on different content from the article, then some may focus on details that were easier to understand in the Simple vs. Complex version. Such questions would appear to be "good" in this study because they would appear to reveal a difference between Simple and Complex. To prevent such differences in content from affecting our comparison of different question types (ClipArt, MultipleChoice, TrueFalse), we selected six basic facts of information in each article to ask comprehension questions about. We asked about each fact using questions of each type. Thus, we produced three different types of question for each of the six facts for all eleven articles: a total of 198 questions (11 articles × 6 facts × 3 question types).

For each article, one question was: "What is this article about?" The correct answer to this question was always a concrete entity (person, place, thing) in the story. The other facts for each article were of mix of different semantic categories: entities, abstract concepts, verbs or action events, locations, how/why questions, numbers, and non-referring words. (Answers to questions in this final category are *words themselves*, e.g. the brand name of a product. These answer choices do not refer to an item or event; the *word itself* is the answer.) For number and non-referring-word questions, it can be difficult to identify ClipArt images. For small numbers (less than 6), copies of an object can be shown, but few of the number questions in this study met this criterion.

MultipleChoice Version:
What product does the Kryptonite company make?
a. cars
b. bikes
c. locks and chains

ClipArt Version:
What product does the Kryptonite company make?

a. cars

b. bikes

c. locks and chains

TrueFalse Version:
Does the Kryptonite company make bikes?
a. yes
b. no
c. it didn't say

Figure 1. MultipleChoice, ClipArt, and TrueFalse versions of a question about the same fact from an article on bike theft.

For example, one non-referring-word question asked for the brand name that a company planned to use for a product. The ClipArt question included photos of signs containing only the word of each answer choice. A similar strategy was used to obtain images for use in ClipArt questions in which the answer choices were numbers. Such use of images (in which the image is merely a photo of a printed word on a sign) is likely of little benefit to users beyond simply displaying text. In fact, differences in appearance could distract the participant. Because we were not entirely pleased with our ClipArt for these cases, we analyze the results of questions about these facts separately in section 4.1.

3.5 Experiment Design

While each participant saw all eleven articles, we needed to select a sequence for presenting the articles to each participant and whether each article would be presented as a Simple or Complex version. We produced ten random permutations of the 11 articles, and we reused each permutation two times during the study. One participant saw this permutation in this manner: simple, complex, simple, complex, etc. Another participant should saw that same permutation in this manner: complex, simple, complex, simple, etc. Thus, we had 20 unique sequences of experiment stimuli, one for each participant in the study. No participant saw both a Simple and a Complex version of the same article.

Each participant was asked about all six facts for each article, but the order in which they were asked about the facts and the question-type (ClipArt, MultipleChoice, or TrueFalse) for each fact had to be selected. We rotated the order in which each of the six facts asked about each article, and we ensured that no participant was asked about the same fact more than once. For each article, we wanted each participant to have two ClipArt questions, two MultipleChoice, and two TrueFalse. We also ensured that across the entire study, the number of questions of each type (ClipArt, MultipleChoice, TrueFalse) was balanced across each participant, each fact, each article, and each version (Simple, Complex). For each question, we also randomized the order of the answer choices – to compensate for the tendency of some participants to select the first or last choice presented.

After participants saw the article and the six comprehension questions, we presented a set of Likert-scale questions for the article. Our pilot study (section 1.2) suggested that participants answered five-point Likert scales with little regard to text complexity. So, in this study, we included three-point Likert scales (to see if they would be more manageable):

- Did this story feel short or long? Short, Medium, Long
- Was it easy or hard to understand? Easy, Medium, Hard
- Was it interesting or boring? Interesting, Medium, Boring

4. RESULTS OF EVALUATION STUDY

Statistical tests to be performed were planned prior to data collection. Since no subject saw both Complex and Simple versions of the same article, a Mann-Whitney U-test was used to check for significant differences between per-question responses for Simple and Complex versions of articles for each question-type. A non-parametric significance test was selected because the comprehension data were not known to be normally distributed. A significance level of alpha = 0.05 was chosen for this study.

4.1 Distinguishing Simple vs. Complex

We care about the ability of each question type to reveal a *difference* between the Simple vs. Complex stories. Figure 2 displays average comprehension question success rate for each question type (ClipArt, MultipleChoice, TrueFalse) and article-version (Simple, Complex). ClipArt questions showed a larger difference between the Simple and Complex versions of articles (Complex/Simple ratio for ClipArt = 1.17, MultipleChoice = 1.06, TrueFalse = 1.01). The TrueFalse questions showed little difference between the Simple and Complex versions of the texts. Success on TrueFalse was lower than the other types of questions (significant). Of course, even hard questions could be good ones if they can reveal the difference between Complex and Simple; unfortunately, TrueFalse does a poor job at distinguishing these.

Section 3.4 discussed concerns about some ClipArt questions about numbers or non-referring words. To isolate these questions from the rest of the data, we construct a response meta-variable called "ClipArtOptional." We re-examine the data to collapse the responses from the MultipleChoice and ClipArt questions into a single column according to this policy: for number or non-referring-word facts, use a MultipleChoice question; otherwise, use a ClipArt question. The right side of Figure 2 displays results for ClipArtOptional, and we see that ClipArtOptional has a larger difference between Complex and Simple articles (in fact, the difference is statistically significant). Because a true difference is assumed between Simple and Complex, ClipArtOptional's ability to indicate this difference suggests that this may be the best design for comprehension questions in future studies. The way in which we selected clip-art and photos for the number and non-referring-word questions in this study apparently introduced noise into the responses collected with ClipArt questions. Of course, ClipArtOptional was synthesized after-the-fact from this study's data; a better evaluation of this question-type-selection policy would be to conduct an additional study in which the questions given to some participants are chosen based on this policy.

The responses to the Likert-scale questions were coded as 0 (easy, short, interesting), 0.5 (medium), and 1 (hard, long, boring). Figure 3 shows the results for these values. Differences between Complex and Simple articles were not statistically significant. Interestingly, despite low comprehension question results, many participants rated the articles as "Easy" to understand.

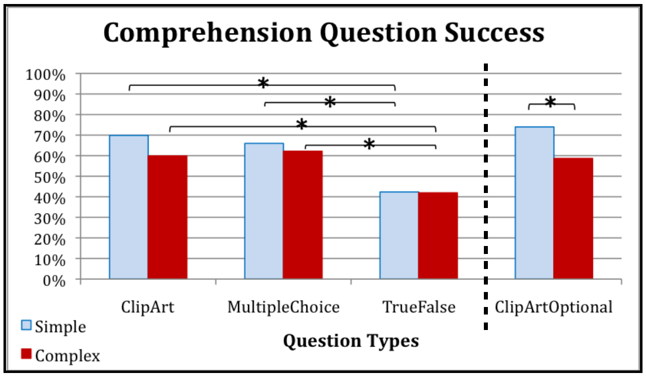

Figure 2. Percentage correct responses for each question type, statistically significant differences marked with an asterisk in the graph.

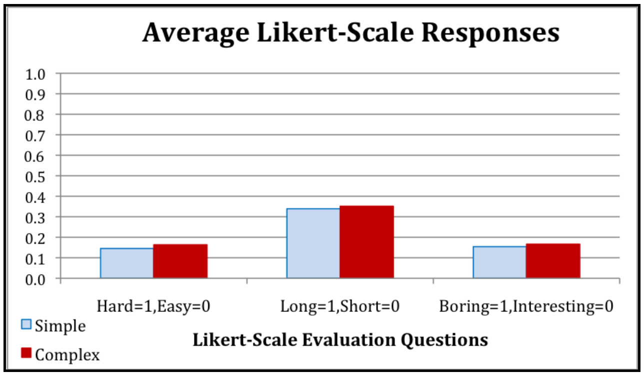

Figure 3. Average response scores for Likert-scale questions, no statistically significant differences observed between the Simple and Complex versions of articles.

5. CONCLUSION AND FUTURE WORK

This study has identified important factors in the design of experimental studies with adults with ID for evaluating readability and simplification technologies. In particular, this study has identified types of comprehension questions that are most effective at highlighting the difference between complex and (human) simplified texts for these users. This work has laid an important framework for future experimental evaluation studies of text simplification and read-aloud software for these users. The most effective question type used in this study was a question with three multiple-choice answers, each illustrated by clip-art of a photograph. We found that photos or clip-art images are less useful for answer choices containing numbers or non-referring words. While the results for ClipArtOptional questions were significant, the difficulty in measuring text complexity with these users is not entirely surprising – especially considering all the factors that may affect performance on a comprehension question:

1. The difficulty level of the text. (We want to measure this.)
2. The difficulty level of the specific fact being tested.
3. Attention and engagement of the research participant.
4. Participant's interest in the specific topic of that article.
5. Participant's memory of that specific fact from the text.
6. Participant's skill at the story-reading/listening task.
7. Participant's skill at the question-answering task.
8. Participant's understanding of the question being asked.
9. Participant's understanding of the answer choices presented.
10. Participant's ability to view and understand clip-art pictures.
11. Participant's outside knowledge of the information topic.
12. Participant's familiarity with entities in the answer choices.
13. Participant's accidental error in selecting an answer choice.

We are interested in measuring factor #1 on this list; the other factors act as "noise." Our results for ClipArtOptional indicate that this question type can indicate the difficulty level of a text – but in a very noisy manner. Careful experiment design (with a variety of article topics and questions) and a large sample size are needed to successfully measure the complexity level of a text (through all of this noise). We believe the benefit of clip-art/photos arises from their helping participants understand the answer choices (factor #9) and perhaps from keeping participants more engaged in the experiment (factor #3). However, clip-art and photos have a risk of introducing noise from the participant's understanding of the images (factor #10) and perhaps some additional risk of participants selecting answer choices with which they are familiar (factor #12). Thus, we needed to empirically measure the use of these images in this study.

While we focused on text comprehension by adults with ID, our findings may have applications for the design of experiments to measure comprehension by other groups of users with literacy impairments. Further, while we measured text understandability and users' opinions about whether a text was easy/hard, interesting/boring, or long/short, the results of our study may also be of interest to researchers conducting usability studies or other surveys with adults with ID. Given our challenges in using TrueFalse or Likert-scale questions with these users, caution may be warranted in the use of such question types with these users.

A goal of many accessibility and assistive technology researchers is to involve users in the design and evaluation process to give them a *voice* as to what they want and need from technology. Studies like the one in this paper illustrate that with some users, it is not enough to merely invite them to participate in studies. We must know the right questions to ask and the right way to interpret their response – to provide them with a real voice in the process.

5.1 Future Work

In future work, we intend to study further how to best conduct evaluations of computational linguistic software for these users. In addition to conducting studies using the ClipArtOptional policy developed in this paper, we also plan on incorporating alternative comprehension question designs – perhaps using two or four answer choices per question. More diverse possibilities we have considered include using physical objects to represent the answer choices available to the participant or asking the participant to perform a task in which memory of information content from the article would affect their behavior. While these questions can help to measure comprehension, we also want to uncover methods for better understanding these users' satisfaction or enjoyment of the system. The Likert-scale questions in this study proved insufficient at this task, and we have found that these users struggle to articulate answers to open-ended survey questions.

Section 2.1 discussed Automatic Readability Scoring Systems (ARSS) and highlighted their current limitations at evaluating the complexity of texts for adults with ID. While we don't intend to use ARSS to evaluate our text-simplification system, it is still important for us to pursue research in this area. This is because we envision ARSS as an important internal component of our future text-simplification system. An ARSS tailored to adults with ID could be used to select easier-to-read versions of news articles (as a starting point for our system), and the ARSS could be used as a guiding metric for the text-simplification software. It could help it decide which transformations to make to a text and decide when it has simplified a text enough. To use an ARSS both as an internal component and as an evaluation metric of our text-simplification system would be circular. Thus, we prefer to use evaluation techniques in which actual users evaluate the output of our simplification system via comprehension questions.

To build an ARSS using statistical computational linguistic techniques, it is important to have a large corpus of texts labeled with readability judgments from adults with ID (to serve as training data for machine learning techniques). An ideal corpus would contain texts that have been written specifically for adults with ID, in particular if such texts were "paired," with alternate versions of each text written for a general audience. Such data would allow statistical ARSS models to learn which linguistic features of texts are predictive of their readability for these users (independent of the topic of the text). The best way to accumulate such a paired and evaluated corpus is to conduct experimental studies. Of course, first it must be determined what form of comprehension questions to ask in such studies. Thus, learning how to best experimentally evaluate text readability is not only important for our research from an *evaluation* perspective, but it will also allow us to create a linguistic resource (a corpus of texts labeled with readability judgments) needed to *build* our system.

We will also continue the design of our read-aloud/read-along text simplification system for presenting local news articles to adults with ID. Significant computational linguistic work remains: creating a reliable ARSS tailored to adults with ID and building simplification software to modify a text for these users. As we conduct more experiments in which adults with ID evaluate complex or simplified articles, we will begin to accumulate a corpus of texts, each labeled with a comprehension score by adults with ID. This resource will be useful in the design of the linguistic components of our project, but we also believe it will be useful to researchers studying text simplification tools, readability metrics, or literacy/educational issues for adults with ID. After addressing linguistic issues, we will focus on the HCI design issues of this application, drawing on the evaluation techniques developed by other researchers (section 1.1). We believe that our future system would promote community awareness, access to relevant daily living information, awareness of conversational topics to promote social interactions, and reading practice time.

6. ACKNOWLEDGMENTS

This work was supported by a grant from the Research Enhancement Committee at CUNY Queens College and from the PSC-CUNY Research Award Program. We thank the Weekly Reader Corporation for making its corpus available for our research. We are grateful to Paola Garcia, Allen Harper, Pengfei

Lu, Karthik Natarajan, Josh Waxman, and Henry Wiyanto for their help during the comprehension studies with adults with ID.

7. REFERENCES

[1] Carroll, J., Minnen, G., Pearce, D., Canning, Y., Devlin, S., Tait, J. 1999. Simplifying text for language-impaired readers. In *Proceedings of EACL'99 Poster*, p. 269.

[2] Chall, J.S., and Dale, E. 1995. *Readability revisited: the new Dale-Chall readability formula*. Cambridge: Brookline.

[3] Cohen, S.A., Steinberg, J.E. 1983. Effects of Three Types of Vocabulary on Readability of Intermediate Grade Science Textbooks: An Application of Finn's Transfer Feature Theory. *Read Res Q*, 19(1):86-101.

[4] Davies, D., Stock, S. and Wehmeyer, M. 2001. Enhancing Independent Internet Access for Individuals with Mental Retardation through Use of a Specialized Web Browser: A Pilot Study. *Educ Train Ment Retard Dev Disabil*, 36(1):107-113.

[5] Davies D., Stock S., King L., and Wehmeyer M. 2008. "Moby-Dick Is My Favorite:" Evaluating a Cognitively Accessible Portable Reading System for Audiobooks for Individuals with Intellectual Disability. *Intellectual and Developmental Disabilities.* 46(4):290-298.

[6] Davison, A., and Kantor, R. 1982. On the failure of readability formulas to define readable texts: A case study from adaptations. *Read Res Q*, 17(2):187-209.

[7] Dawe, M. 2007. Understanding mobile phone requirements for young adults with cognitive disabilities. In *Proceedings of ASSETS'07*, New York: ACM Press, pp. 179-186.

[8] Drew, C.J., and Hardman, M.L. 2004. *Mental retardation: A lifespan approach to people with intellectual disabilities (8th ed.)*. Columbus, OH: Merrill.

[9] Feng, J., Lazar, J., Kumin, L., and Ozok, A. 2008. Computer usage by young individuals with down syndrome: an exploratory study. In *Proceedings of ASSETS'08*, New York: ACM Press, pp. 35-42.

[10] Feng, L., Elhadad, N., Huenerfauth, M. 2009. Cognitively Motivated Features for Readability Assessment. In *Proceedings of EACL'09*, pp. 229-237.

[11] Fickas, S., Pataky, C., and Chen, Z. 2006. DuckCall: tackling the first hundred yards problem. In *Proceedings of ASSETS'06*. New York: ACM Press, pp. 283-284.

[12] Fowler, A.E. 1998. Language in mental retardation. In Burack, Hodapp, and Zigler (Eds.), *Handbook of Mental Retardation and Development*. Cambridge, UK: Cambridge Univ. Press, pp. 290-333.

[13] Gray, W. S. and B. Leary. 1935. *What makes a book readable.* Chicago: Chicago University Press.

[14] Heilman, M., Collins-Thompson, K., Callan, J., and Eskenazi, M. 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Proceedings of NAACL'07*, pp. 460-467.

[15] Hickson-Bilsky, L. 1985. Comprehension and mental retardation. *International Review of Research in Mental Retardation*, 13:215-246, St. Louis: Elsevier.

[16] Huenerfauth, M. 2008. Evaluation of a Psycholinguistically Motivated Timing Model for Animations of American Sign Language. In *Proceedings of ASSETS'08*, New York: ACM Press, pp. 129-136.

[17] Jones, F.W., Long, K., Finlay, W.M. 2006. Assessing the reading comprehension of adults with learning disabilities. *J Intel Disabil Res*, 50(6):410-418, Malden, MA: Blackwell.

[18] Jones, F.W., Long K., and Finlay W.M.L. 2007. Symbols can improve the reading comprehension of adults with learning disabilities. *J Intel Disabil Res*, 51(7):545-550.

[19] Katims, D.S. 2000. Literacy instruction for people with mental retardation: Historical highlights and contemporary analysis. *Educ Train Ment Retard Dev Disabil*, 35(1):3-15.

[20] Kincaid, J., Fishburne, R., Rodgers, R., and Chisson, B. 1975. Derivation of new readability formulas for navy enlisted personnel. Technical report, Research Branch Report 8-75, U.S. Naval Air Station.

[21] Kintsch, W. 1998. Comprehension: A paradigm for cognition. New York: Cambridge University Press.

[22] Lepistö, A., and Ovaska, S. 2004. Usability evaluation involving participants with cognitive disabilities. In *Proc of NordiCHI-2004*, New York: ACM Press, pp. 305-308.

[23] LoPresti, E., Kirsch, N., Simpson, R., and Schreckenghost, D. 2005. Solo: interactive task guidance. In *Proceedings of ASSETS'05,* New York: ACM Press, pp. 190-191.

[24] Maki, R.H., and Berry, S.L. 1984. Metacomprehension of text material. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 10(4):663-679.

[25] Perfetti, C., Lesgold, A. 1977. Discourse comprehension and sources of individual differences. In M. A. Just, & P. A. Carpenter (eds.), *Cognitive processes in comprehension*. Hillsdale, NJ: Lawrence Erlbaum Assoc.

[26] Petersen, S.E., Ostendorf, M. 2009. A machine learning approach to reading level assessment. *Computer Speech and Language*, 23: 89-106, St. Louis: Elsevier.

[27] Si, L., and Callan, J. 2001. A statistical model for scientific readability. In *Proc CIKM'01*, ACM Press, pp. 574-576.

[28] Small, J., Schallau, P., Brown, K., and Appleyard, R. 2005. Web Accessibility for People with Cognitive Disabilities. In *CHI '05 Extended Abstracts*, New York: ACM, 1793-1796.

[29] SpheNet. 2007. TTSReader, (accessed December 8, 2007), http://www.sphenet.com/TTSReader/

[30] TheArcLink. 2003. The Desk Launches in 11 States for People with Disabilities, (accessed October 29, 2008), http://www.thearclink.org/news/article.asp?ID=591

[31] U.S. Census Bureau. 2000. *Projections of the total resident population by five-year age groups* and *sex, with special age categories: Middle series 2025-2045*. Washington: Census Bureau, Populations Projections Prog., Population Division.

[32] Williams, S., and Reiter, E. 2005. Generating readable texts for readers with low basic skills. In *Proc ENLG'05*, pp. 140-147.

[33] Wong, B., Jones, W. 1982. Increasing Metacomprehension in Learning Disabled and Normally Achieving Students through Self-Questioning Training. *Learn Disabil Q*, 5(3):228-240.