frontiers | Frontiers in Communication

TYPE Brief Research Report
PUBLISHED 03 August 2023
DOI 10.3389/fcomm.2023.1175625

Check for updates

OPEN ACCESS

EDITED BY
Petra Jaecks,
Bielefeld University, Germany

REVIEWED BY
Susana Silva,
University of Porto, Portugal
Katharina Weiland,
Humboldt University of Berlin, Germany
Susan Buell,
University of Dundee, United Kingdom

*CORRESPONDENCE
Andreas Säuberli
✉ andreas@cl.uzh.ch

RECEIVED 27 February 2023
ACCEPTED 06 June 2023
PUBLISHED 03 August 2023

CITATION
Säuberli A, Hansen-Schirra S, Holzknecht F, Gutermuth S, Deilen S, Schiffl L and Ebling S (2023) Enabling text comprehensibility assessment for people with intellectual disabilities using a mobile application. *Front. Commun.* 8:1175625. doi: 10.3389/fcomm.2023.1175625

COPYRIGHT
© 2023 Säuberli, Hansen-Schirra, Holzknecht, Gutermuth, Deilen, Schiffl and Ebling. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Enabling text comprehensibility assessment for people with intellectual disabilities using a mobile application

Andreas Säuberli[1]*, Silvia Hansen-Schirra[2], Franz Holzknecht[3], Silke Gutermuth[2], Silvana Deilen[2], Laura Schiffl[2] and Sarah Ebling[1]

[1]Department of Computational Linguistics, University of Zurich, Zürich, Switzerland, [2]Johannes Gutenberg University Mainz, Germersheim, Germany, [3]University of Teacher Education in Special Needs Zurich, Zürich, Switzerland

In research on Easy Language and automatic text simplification, it is imperative to evaluate the comprehensibility of texts by presenting them to target users and assessing their level of comprehension. Target readers often include people with intellectual or other disabilities, which renders conducting experiments more challenging and time-consuming. In this paper, we introduce *Okra*, an openly available touchscreen-based application to facilitate the inclusion of people with disabilities in studies of text comprehensibility. It implements several tasks related to reading comprehension and cognition and its user interface is optimized toward the needs of people with intellectual disabilities (IDs). We used *Okra* in a study with 16 participants with IDs and tested for effects of modality, comparing reading comprehension results when texts are read on paper and on an iPad. We found no evidence of such an effect on multiple-choice comprehension questions and perceived difficulty ratings, but reading time was significantly longer on paper. We also tested the feasibility of assessing cognitive skill levels of participants in *Okra*, and discuss problems and possible improvements. We will continue development of the application and use it for evaluating automatic text simplification systems in the future.

KEYWORDS

Easy Language, easy-to-read, readability, reading comprehension, text simplification, intellectual disabilities

1. Introduction

The terms “Easy Language”, “Plain Language”, “easy-to-read language”, and “simplified language” all denote varieties of standard language which aim to improve comprehensibility for a wide range of target groups, including people with intellectual disabilities[1] (IDs) or communicative impairments, people who are deaf or hard-of-hearing, or non-native speakers (Maaß, 2020). As efforts to automate the process of simplifying texts are increasing (Schulz et al., 2020; Al-Thanyyan and Azmi, 2021), it also becomes increasingly important to develop and apply accurate and reliable methods for evaluating simplified texts.

Much of the previous work on comprehensibility assessment of simplified texts has focused on comprehension tests and perceived difficulty ratings by experts (e.g., simplified

[1] We use the term *intellectual disability* as an umbrella term to include all forms of cognitive impairment leading to a right to information in Easy Language according to the United Nations Convention on the Rights of Persons with Disabilities (UN CRPD).

language translators) or readers sampled from a general population, which are not necessarily representative of the target group (Alva-Manchego et al., 2021). The reason for this is that target groups are often difficult to access and experiments involving them require significantly more time and expertise (Saggion et al., 2015; Stajner, 2021). Particularly in the field of automatic text simplification, evaluation studies involving the target audience are rare (Stajner, 2021), and most researchers resort to experts or users on crowdsourcing platforms for human evaluation (e.g., Xu et al., 2016; Sulem et al., 2018c; Zhao et al., 2020). In addition, although many people in the target group are active users of digital media and devices (Ramsten et al., 2018), existing tools and platforms for human evaluation are rarely optimized for people with disabilities (Uzor et al., 2021), leading to a high threshold to including the target group in evaluation studies. These impedes digital participation, because people with IDs are excluded from research on improving communication technology targeted at them.

We believe that this situation can be improved by providing tools which enable more efficient, effective, and inclusive evaluation studies with participants from diverse target groups, particularly, people with IDs. Developing digital applications for comprehensibility assessment and adapting them to the needs of these target groups reduces the need for close supervision and increases flexibility in terms of where and when experiments can be conducted. In addition to reducing cost, this also enables a more naturalistic reading environment compared to paper-and-pencil tests in a laboratory setting. In the present work, we introduce and test such a tool and apply it in an initial experiment with participants with ID.

The main contributions of this paper are:

1. We describe the design and implementation of *Okra*, a mobile application for testing text comprehensibility with people with IDs (Section 3).
2. We present results from a small-scale study with *Okra* aiming to detect potential effects of the digital testing modality compared to traditional paper-and-pencil methods, and to test the feasibility of administering low-level cognitive tasks (Section 4).

2. Background and related work

2.1. Human evaluation of text difficulty

Although there is no consensus on best practices, it is generally accepted that evaluating Easy Language with target readers is crucial for obtaining representative results (Alva-Manchego et al., 2020, 2021; Stajner, 2021; Stodden, 2021). However, human evaluation of text difficulty is mostly done with populations such as crowdworkers (Leroy et al., 2013; Redmiles et al., 2019), experts (Sulem et al., 2018a,b), students (Fulmer et al., 2015; Leroy et al., 2022), or target groups that are more easily accessible, such as non-native speakers (Crossley et al., 2014; Vajjala et al., 2016; Vajjala and Lucic, 2019). Exceptions include studies with deaf and heard-of-hearing participants (Alonzo et al., 2021), readers with dyslexia (Rello et al., 2013a,b,c), and people with IDs (Huenerfauth et al., 2009; Fajardo et al., 2014; Saggion et al., 2015; Gutermuth, 2020). Particularly in the field of automatic text simplification, output texts are rarely evaluated with vulnerable populations. The main reasons for this are the difficulty and time involved in accessing these groups and adapting the experiments to the special needs of the participants, as well as ethical issues (Saggion et al., 2015; Deilen and Schiffl, 2020; Stajner, 2021).

Several different methods have been proposed and used to measure the difficulty of texts. For subjective perception of difficulty, Likert scales are most frequently used (e.g. Leroy et al., 2013, 2022; Fulmer et al., 2015). For measuring actual or objective difficulty, various types of comprehension testing are applied, including multiple-choice questions (Leroy et al., 2013, 2022; Fajardo et al., 2014; Charzyńska and Dębowski, 2015; Alonzo et al., 2021), cloze tests (Charzyńska and Dębowski, 2015; Redmiles et al., 2019), and free recall questions (Leroy et al., 2013, 2022). Some studies also measure different aspects of reading behavior, such as the time taken to read a text (Crossley et al., 2014; Saggion et al., 2015; Alonzo et al., 2021), gaze patterns recorded through eye-tracking (Rello et al., 2013c; Vajjala et al., 2016; Gutermuth, 2020), or scrolling interactions (Gooding et al., 2021).

2.2. Tools for computer-based reading experiments

Many tools used in behavioral and psycholinguistic research support various types of reading tasks, for example, *PsychoPy* (Peirce et al., 2019), *PsyToolkit* (Stoet, 2017), or *jsPsych* (de Leeuw, 2015). Survey platforms such as *Qualtrics* or *SurveyMonkey* provide basic features for multiple-choice or text-based responses, and *Amazon Mechanical Turk* and *Qualtrics* support custom front-end implementations to collect behavioral measurements such as reading time and scrolling behavior, which often involves considerable technical expertise (e.g. Alonzo et al., 2021; Gooding et al., 2021), and making implementations accessible requires user testing. We are not aware of any tools specifically developed for reading experiments with people with IDs. Large-scale digitized testing for this target group is uncommon, and studies designed for participants with IDs are still mostly done using paper-based methods (e.g. Huenerfauth et al., 2009; Fajardo et al., 2014).

2.3. Usage of technology by people with ID

Insights from interviews and surveys have shown that the use of information and communication technologies, and mobile devices in particular, has become widespread among adults with IDs (Ramsten et al., 2018), and may even have significant personal and social benefits (Chadwick et al., 2018; Martin et al., 2021). Use of technology has also been found to be beneficial for people with IDs in education (Maebara et al., 2022) and the development of skills in daily life (Jung et al., 2021), particularly due to the variety of modalities (text, images, video, audio, etc.) supported by the devices. This strongly suggests that participation in digital comprehensibility studies should be possible for this group. However, existing software solutions, including crowdsourcing platforms such as *Amazon Mechanical Turk*, are generally suboptimal in terms of accessibility for many user groups, including users with IDs (Uzor et al., 2021).

Due to this increased use of technology and the growing need of human evaluators from target groups of Easy Language, developing a digital application that is accessible for people with ID is a logical next step. However, the feasibility of such applications and potential effects of the digital modality compared to conventional paper-based methods must be thoroughly tested. Our work presents a first step in this direction.

3. Application description

In response to the increasing demand for and importance of representative human evaluations of text simplification and the lack of suitable tools for one of the main target groups of Easy Language (people with IDs), we present a prototype of a mobile application for touchscreen-based assessment of reading comprehension. Its main goal is to create a simple way for researchers to set up and configure experiments, which can then be presented to participants in an accessible way, either on their own device, or a device provided to them by the researcher (in a laboratory setting).

3.1. Requirements

Based on the specific needs and difficulties of the target users and the shortcomings of existing tools for collecting reading comprehension data described in Section 2, we formulate the following requirements for our application:

From a participant's perspective, the application should:

- Provide an easy-to-understand and easy-to-use interface, specifically for participants with mild to moderate IDs or limited language skills.
- Support independent use as best as possible, i.e. on a personal device, without supervision.
- Keep up the user's motivation.

From a researcher's perspective, the application should:

- Collect all data which is potentially useful for evaluating Easy Language.
- Allow conducting both remote and in-lab experiments.
- Provide a simple and reproducible way of setting up customized experiments.

3.2. Design and implementation

To allow conducting experiments both in a lab and remotely using participants' personal devices, we chose a client-server implementation. The client application is installed on a touchscreen device and used by the participant to complete tasks. On the server side, we implemented a web application which includes a dashboard where researchers can configure experiments and download results, and an application programming interface (API) to communicate with registered clients.

To address the requirements described in the previous section, we designed the graphical user interface to reduce the amount of information visible on screen simultaneously and provide clear indicators of the next steps at every point in time. As it is safe to assume that most participants are at least somewhat familiar with modern Android or iOS applications (Ramsten et al., 2018), we follow Material Design specifications[2] to implement components and navigation behavior reminiscent of widely used apps. When participants open *Okra*, they are asked to scan a QR code given to them by the researcher, which registers their device and allows them to receive experiments to participate in. Each experiment starts with a screen with instructions written in Easy Language, followed by a practice task and a number of main tasks. After each task, an encouraging message is shown for positive reinforcement, and the participant is allowed to take a break and continue at their own pace. Where easily possible, we included gamification elements such as colorful pictures and animations (see Figure 1 for sample screenshots). During tasks, user interactions (i.e., scrolling and touch events) are recorded, and the log is sent to the server after the task is finished.

No personal information is collected or stored in the client application, and participants are only identified by randomly generated identifiers. The researcher is responsible for collecting personal information and mapping them to participant identifiers. This means that data confidentiality can be handled by the researcher according to individual requirements.

The client application is implemented using the cross-platform user interface (UI) toolkit *Flutter*[3], meaning that it can be compiled into a native Android/iOS app or a Progressive Web App (PWA) which can be installed directly from a web browser. The server is a *Django*[4] app and contains a dashboard for registering participants and configuring experiments, and the API for communicating with clients.

3.3. Tasks

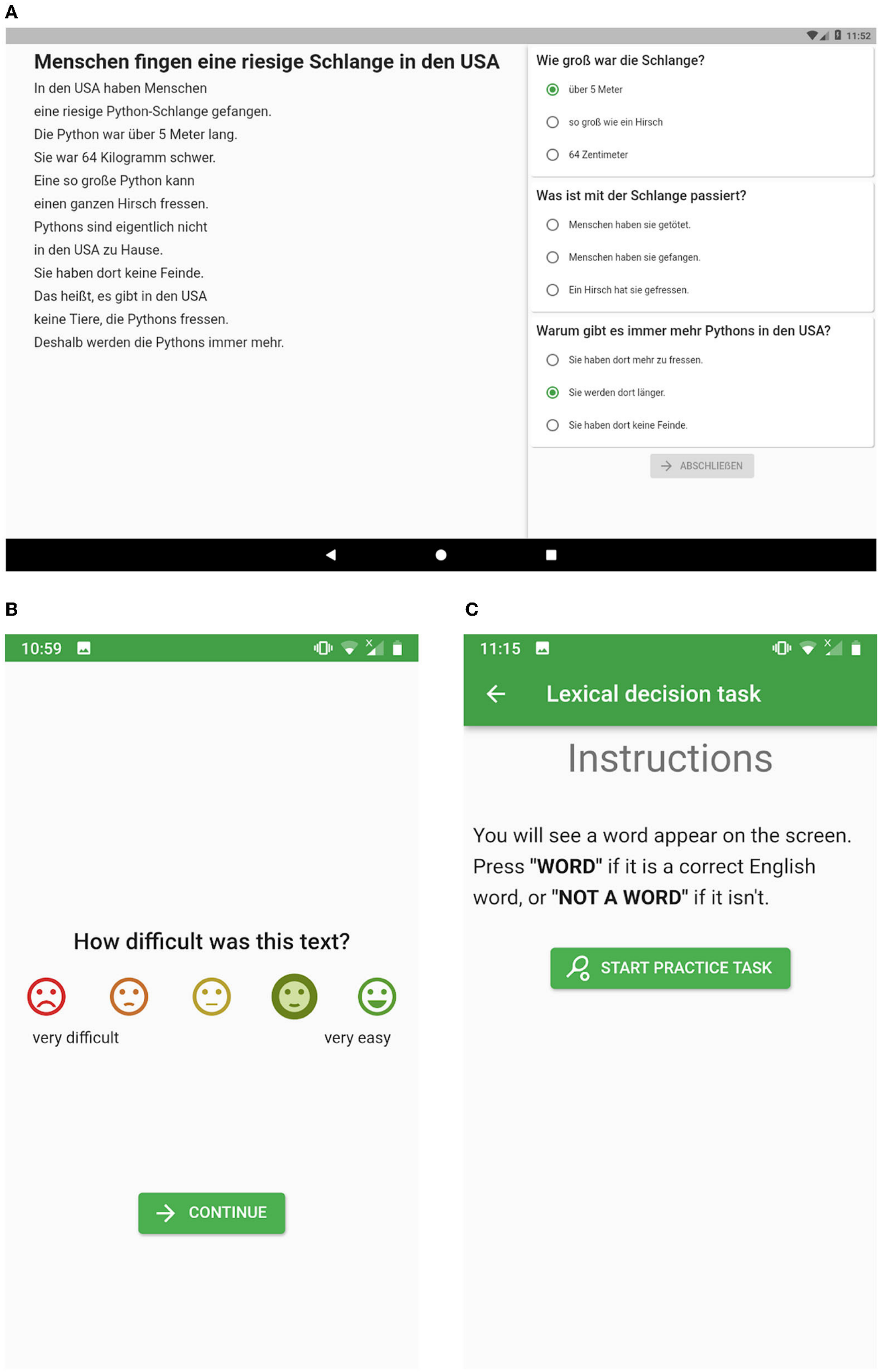
We identified tasks which can be made accessible to target users while remaining useful for Easy Language research and evaluation of text simplification. In a typical study, measuring low-level cognitive skills may also be relevant for screening or comparing to a control group. Therefore, apart from reading comprehension tasks, we also include tasks for measuring skills such as working memory and visual attention. The following types of tasks are currently implemented in the prototype:

- Reading tasks with multiple-choice questions and Likert-scale or slider ratings [screenshots (A) and (B) in Figure 1].
- Multiple-choice cloze tests, where a short segment of text with a single gap is shown at a time.
- Lexical decision tasks, where the user judges whether a string of characters is a word or a non-word.

[2] https://material.io/

[3] https://flutter.dev/

[4] https://www.djangoproject.com/



Menschen fangen eine riesige Schlange in den USA

In den USA haben Menschen
eine riesige Python-Schlange gefangen.
Die Python war über 5 Meter lang.
Sie war 64 Kilogramm schwer.
Eine so große Python kann
einen ganzen Hirsch fressen.
Pythons sind eigentlich nicht
in den USA zu Hause.
Sie haben dort keine Feinde.
Das heißt, es gibt in den USA
keine Tiere, die Pythons fressen.
Deshalb werden die Pythons immer mehr.

Wie groß war die Schlange?

● über 5 Meter
○ so groß wie ein Hirsch
○ 64 Zentimeter

Was ist mit der Schlange passiert?

○ Menschen haben sie getötet.
○ Menschen haben sie gefangen.
○ Ein Hirsch hat sie gefressen.

Warum gibt es immer mehr Pythons in den USA?

○ Sie haben dort mehr zu fressen.
● Sie werden dort länger.
○ Sie haben dort keine Feinde.

How difficult was this text?

Instructions

You will see a word appear on the screen. Press "**WORD**" if it is a correct English word, or "**NOT A WORD**" if it isn't.

FIGURE 1
Screenshots of *Okra*. **(A)** Reading task with comprehension questions on a tablet screen (in German), as it was presented to participants (cf. Section 4.2). **(B)** Difficulty rating on a phone screen. **(C)** Instructions for a lexical decision task on a phone screen.

- *n-back* tasks for testing working memory, first introduced by Kirchner (1958).
- Digit span tasks, where participants need to remember and recall sequences of digits of increasing length.
- Word-picture-matching tests, where participants choose the matching picture for the displayed word, as described by Deilen (2020).
- Reaction time tests, where an image appears on screen and participants tap it as quickly as possible.
- Trail Making Tests for testing visual attention (Reitan and Wolfson, 1993).
- An adaptation of the electronic short-term memory skill game *Simon*, where participants remember an increasingly long sequence of buttons to press.

Implementations of these tasks are contained in the client application installed on participants' devices. Instructions, stimulus data, and procedure details (number of trials, size of UI elements, timing etc.) can be configured by the researcher through a web application. The client is currently available in German and English.

3.4. Availability

The source code for both client and server implementations are available under free and open source licenses at https://github.com/saeub/okra and https://github.com/saeub/okra-server. The client application is currently not available through any official app store.

4. Experiment: effect of testing modality and feasibility analysis

We used *Okra* in a small-scale experiment with participants with IDs. The goal of this experiment was to gather initial evidence for the following two questions:

- Is there a measurable difference between reading comprehension and perceived difficulty rating tasks performed in *Okra* compared to paper-and-pencil testing?
- Is it feasible to test low-level cognitive skills with people with ID using *Okra*?

The latter question is relevant because in future studies, these cognitive tasks will be useful for characterizing the target group, screening participants, or correlating reading behavior to certain cognitive skills.

A selection of results of this study has been reported in Säuberli (2021).

4.1. Participants

After institutional review board (IRB) approval and a pilot study with two participants, 16 participants took part in the main study. They were recruited directly through their instructor in an educational program for people with learning difficulties and disabilities in Austria. There were no additional inclusion criteria. They took part on a voluntary basis and were compensated monetarily. Participants were not screened for disability specifically, but all participants in the educational program have some form of cognitive impairment or learning disorder (the most common being Autism Spectrum Disorder, Down Syndrome, and developmental delay) and a degree of disability of at least 50% according to Austrian legislation.[5] They were aged between 18 and 38 (median: 26) at the time of the first session. Eight of them identified as female, eight as male. All were native German speakers. According to their survey responses from the first session, 14 of them use a smartphone on a daily basis, two only weekly. This is in line with previous research of technology usage among people with ID (Ramsten et al., 2018) and validates our assumptions for the design of the application (cf. Section 3.2). Self-reported reading frequency ["How often do you read texts (for example, in newspapers, books, or the internet)?"] was distributed between *every day* ($n = 4$), *once per week or more* ($n = 8$), and *less than once per week* ($n = 4$). All of them had at some point read texts in Easy Language before.

4.2. Procedure, tasks and variables

There were two sessions per participant. Each session was administered one-on-one by an employee at the facility where the participants' educational program took place. The experiment consisted of a reading task, which was split across the two sessions, and three different low-level cognitive tasks at the beginning of the second session. Each task was preceded by written instructions and a practice trial. These instructions and the remaining text material were checked by a professional in Easy Language to ensure that they adhere to guidelines designed for the target group. In addition, the session administrator constantly monitored the participants' screens during the experiment and, if necessary, added oral instructions, in order to prevent misunderstanding of the tasks.

For the reading task, we selected eight newspaper articles written in German Easy Language taken from the APA (Austrian Press Agency) corpus (Säuberli et al., 2020), ranging between 63 and 122 words in length. For each text, we wrote three multiple-choice comprehension questions with three answer choices. After initially reading the text (without seeing the questions yet), participants had to rate the difficulty of the text on a 5-point rating scale (1 = *very difficult*, 5 = *very easy*). The text was then shown again, together with the comprehension questions, and participants had unlimited time to answer them. This was followed by two more 5-point ratings on the difficulty of the questions (1 = *very difficult*, 5 = *very easy*) and enjoyment ("How much did you enjoy this task?"; 1 = *not at all*, 5 = *very much*). Each participant read four texts on an Apple iPad 2018 (9.7 inches) using *Okra*[6], and

[5] *Verordnung des Bundesministers für Arbeit, Soziales und Konsumentenschutz betreffend nähere Bestimmungen über die Feststellung des Grades der Behinderung (Einschätzungsverordnung)*, BGBl. II Nr. 261/2010.

[6] *Okra* was built as a PWA from the code in public repository (https://github.com/saeub/okra) at commit hash `b56c7a7` and run in the Safari web browser.

four on paper, using a pen to mark their answers. Care was taken that the visual presentation (font size, layout, etc.) was the same in both conditions. In the paper condition, the administrator used a stopwatch to measure the initial reading time.[7]

In the reaction time (RT) task, a red balloon was visible and the participant was instructed to tap it as quickly as possible. After popping the balloon, the next one appeared after a random delay between 0 and 1 second. In the lexical decision task, a string of letters was shown on screen and the participant was instructed to tap the correct button (labeled "WORD" or "NOT A WORD") as quickly as possible. We selected ten words from a list of the 5000 most frequent German words (Perkuhn et al., 2009), and generated ten pseudowords using *Wuggy* (Keuleers and Brysbaert, 2010). In the short-term memory task, participants had to observe four differently colored buttons light up in a specific sequence, starting with a sequence of length 1. They then had to repeat this sequence by tapping the buttons in the correct order. The sequence was then extended by an additional button press and presented again, and so on. The trial ended as soon as the participant pressed an incorrect button. Since the three cognitive tasks heavily rely on precise stimulus timing and touch-based user interaction, they could only be performed on the iPad. The main reason for including them is to test their feasibility with the target group.

4.3. Analysis

We used item response theory (IRT) to answer the question on the difference between modalities. IRT models are used to study how underlying latent traits (i.e., unobservable traits such as reading ability) are linked to observed performances (i.e., scores on a reading test or questionnaire responses on reading difficulty) (see also Ockey, 2021). One particular method of IRT is many-facet Rasch measurement (MFRM; Linacre, 1994), which allows researchers not only to investigate the link between latent traits and observable performances, but also how other factors (so-called "facets") influence the performances (Eckes, 2015). As the factor we were particularly interested in is the condition (paper-and-pencil or *Okra*), we constructed a MFRM model consisting of three facets (participant, item, and condition) and used MFRM bias analyses to study differences between the item and condition facet. For the analysis, we first coded the answers to the items dichotomously as either correct or incorrect. For the three ratings, we applied separate MFRM models with three facets (participant, text, and condition) using the 5-point rating scale responses.

To test the difference in reading time between modalities, we applied a linear mixed-effect model with participants and texts as random effects using the R package *lme4* (Bates et al., 2015) and the formula `reading.time~condition + (1 | participant) + (1 | text)`.

4.4. Results

4.4.1. Reading task

Out of the 128 data points obtained (16 participants × 8 texts), one measurement was lost due to a software bug (which was immediately fixed), leading to a total of 127 data points.

Figure 2 shows the distribution of measurements for participants and items (questions). Question 2 of text G was answered correctly by all participants. For the remaining data points, mean-square infit statistics range between 0.70 and 1.44 for participants and between 0.75 and 1.38 for items, indicating an acceptable model fit. The model could not statistically separate the two elements in the condition facet (separation = 0.00), indicating that participants performed equally well in the two conditions. The bias analysis showed no significant difference between the two conditions for any of the items (all $p > 0.17$). Overall, the ratio of correct answers was quite high, with an average of 17.5 out of 24 correctly answered questions per participant (s.d.: 3.3).

Figure 3 shows mean rating responses for each text. For most participants, there was a strong tendency toward very positive responses, and some of the participants gave the same responses for all texts (four participants in the case of text and question difficulty ratings). All three rating dimensions are also highly correlated with each other (Pearson's $r > 0.55$, $p < 0.001$). For all ratings, the MFRM analyses resulted in 0.00 separation of the condition facet, suggesting that there was no difference in perceived difficulty and enjoyment/motivation between modalities.

Average reading time was noticeably shorter on the iPad than on paper for almost all texts, as Figure 4 shows. According to the linear mixed-effect model, this effect is 9.97 seconds with a standard error of 2.22 seconds ($p < 0.001$). The model also shows considerable variance between individuals, with a standard deviation of 17.53 seconds for the random effect of participants, and less variance between texts (s.d.: 7.83 seconds).

4.4.2. Cognitive tasks

Since the three cognitive tasks heavily rely on precise stimulus timing and touch-based user interaction, they could only be performed on the iPad. A summary of the most relevant measurements is presented in Table 1.

The RT task resulted in a relatively low variance (mean: 0.68 sec, s.d.: 0.10 sec), and there is no significant correlation with any of the other measurements. This suggests that the effect of differences in motor response speeds between participants on other tasks is minimal.

Results from the lexical decision task are in line with psycholinguistic expectations, with pseudowords generally causing a longer RT than words. However, three participants (3, 5, and 10) gave the same response "WORD" to all trials and did not exhibit any difference in RT between words and pseudowords. Responses by participant 11 were also equal to random guessing and showed no difference in RT.

Since the short-term memory task consisted of a single main trial which stopped immediately after the first incorrectly pressed button, we used the maximum score out of practice and main trials

[7] Reading time was measured as the duration between the start of the initial text presentation and the end of the first rating after reading the text (in both conditions). Because the text difficulty rating was on the same page as the text in the paper condition, it was not possible to measure the end of reading precisely.

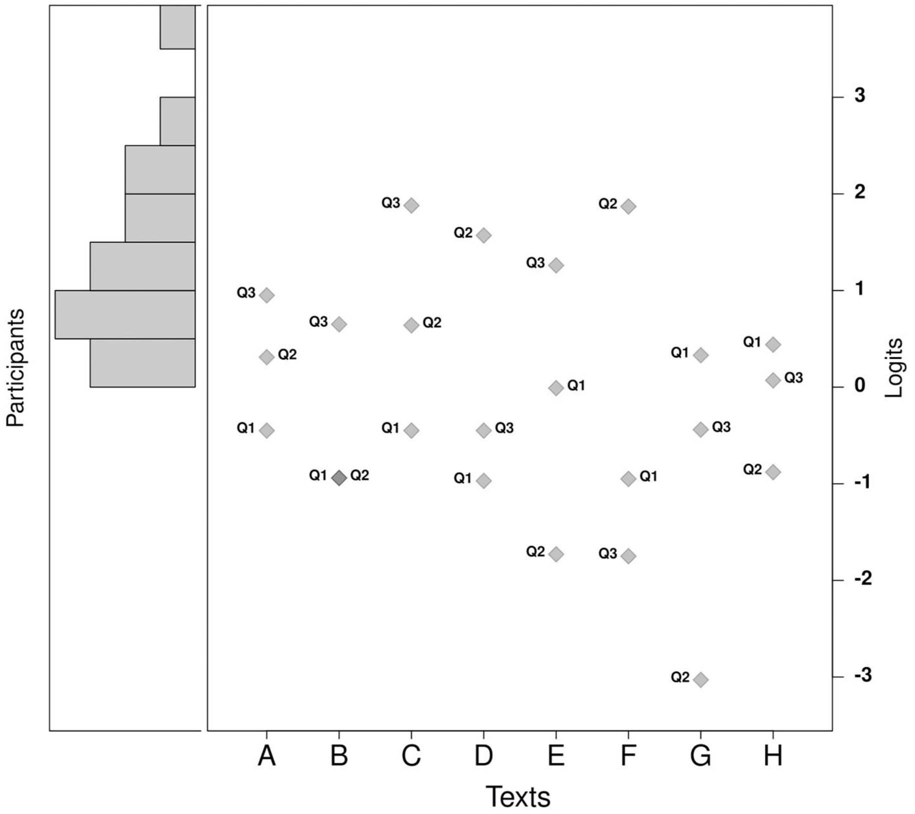


FIGURE 2
Wright map (Many-Facet Rasch Measurement) of performance in comprehension question responses. Participants and comprehension questions (Q1–3 for texts A–H) are projected onto a common logit scale. The higher a participant's logit value, the better their performance, and the higher a question's logit value, the higher its difficulty. A participant has an estimated chance of 50% of correctly answering a question with the same logit value as theirs.

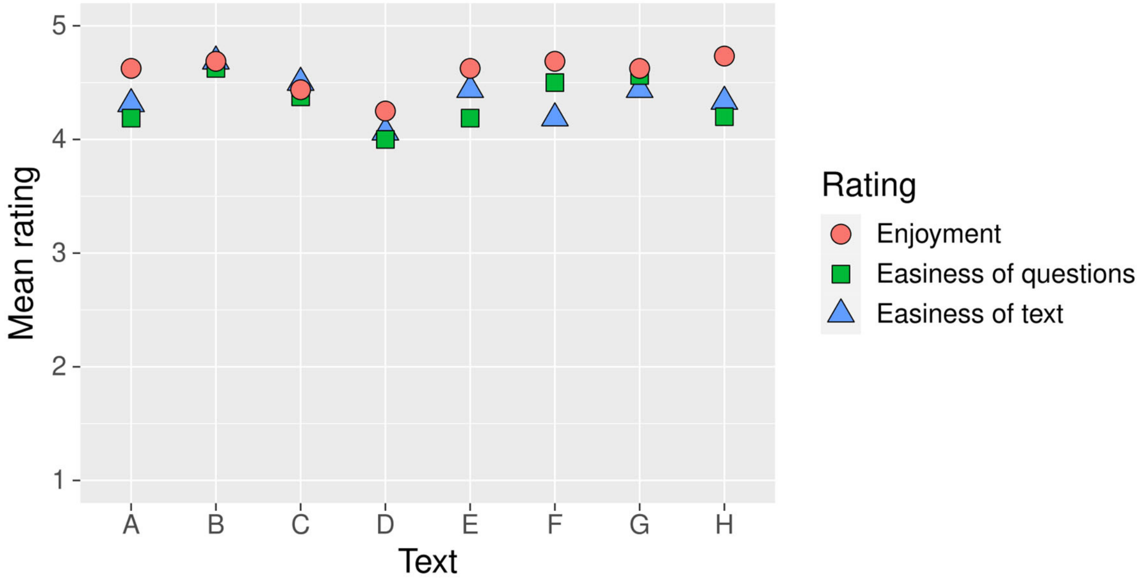


FIGURE 3
Mean rating responses for each text. 1 is the lowest (most negative), 5 is the highest (most positive) response. The questions were "How much did you enjoy this task?" (1 = *not at all*, 5 = *very much*), "How difficult were the questions?" (1 = *very difficult*, 5 = *very easy*, "How difficult was the text?" (1 = *very difficult*, 5 = *very easy*) (presented to participants in German, here translated to English by the authors).

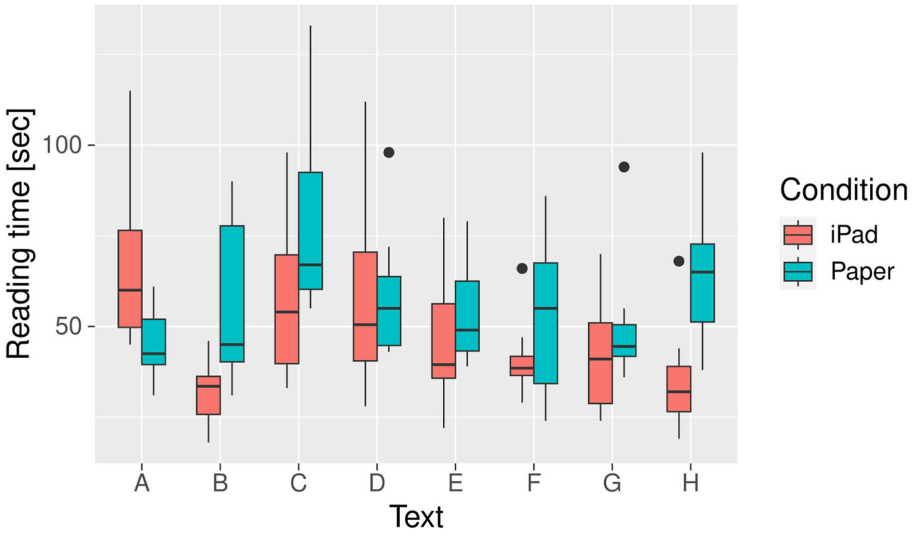


FIGURE 4
Comparison of initial reading times between the two modalities for each text. Reading times are not normalized by text length, as they include the time taken for both reading and the text difficulty rating.

TABLE 1 Summary of aggregated measurements for all participants and tasks.

	Reading		RT	Lexical decision			Memory
Participant	Avg. correct responses	Reading time [s]	Reaction time [s]	Ratio of correct responses	Correct word RT [s]	Correct pseudoword RT [s]	Longest sequence
1	2.63	45.0	0.68	0.85	1.70	2.79	7
2	2.29	43.1	0.55	0.90	1.59	2.13	5
3	2.38	45.5	0.67	*0.50*	*1.89*	—	4
4	2.13	46.4	0.77	0.95	2.44	4.11	3
5	2.13	43.3	0.70	*0.50*	*1.23*	—	5
6	2.63	52.9	0.65	0.95	2.27	2.43	8
7	2.00	97.8	0.52	0.75	4.11	14.57	2
8	1.63	51.4	0.67	0.85	1.86	5.17	7
9	2.88	37.0	0.69	0.95	1.58	2.04	11
10	1.88	33.8	0.68	*0.50*	*0.88*	—	4
11	2.50	59.3	0.72	*0.50*	*0.97*	*0.88*	4
12	2.13	70.1	0.61	0.75	2.67	4.64	4
13	2.75	50.4	0.69	1.00	1.39	2.32	20
14	1.63	31.4	0.60	1.00	1.60	1.76	5
15	1.88	85.9	0.95	0.75	1.64	3.43	5
16	1.63	55.3	0.80	1.00	2.27	2.43	9
Mean	2.19	53.0	0.68	0.79	1.88	3.75	6.4
±s.d.	±0.41	±18.0	±0.10	±0.19	±0.78	±3.47	±4.3

Time measurements are in seconds. Measurements in *italics* were excluded from further analysis due to chance-level performance.

to get a more reliable measurement. Still, we can observe a very large variance between participants.

4.5. Discussion

4.5.1. Effect of testing modality

Both in terms of accuracy of responses to the comprehension question and in terms of subjective perception ratings, we found no evidence of any difference between the two modalities. However, Figures 2, 3 suggest that there is a ceiling effect due to low text and/or question difficulty. This underlines the need for a sufficiently large sample size in the pilot study, since variance between participants is difficult to predict in such a diverse target group. The relatively small sample size is another obvious limitation. At the least, the results allow us to exclude large effect sizes from modality for this target group. This confirms our expectations, given the frequency of technology use reported by the participants and the population of people with IDs in general (Ramsten et al., 2018).

The large difference in reading time is more difficult to explain. One possibility is that actual reading speed was faster when reading on the iPad than on paper, which contradicts previous research which found differences in comprehension but not in reading speed (Kong et al., 2018). Another explanation could be that participants are less inhibited to make the conscious decision that they have finished reading and push the "CONTINUE" button in the application, compared to the paper modality, where the end of the initial reading stage was indicated by participants using the pencil to mark an option on the rating scale. In any case, since the difference in reading time did not appear to affect comprehension, we consider it unproblematic.

4.5.2. Feasibility of cognitive tasks

In order to be feasible in studies with people with ID, the administered tasks must be understood by participants, and maintain participants' attention by avoiding excessive strain or boredom. At least in the RT and lexical decision tasks, the high performance and relatively low variance show that most of the participants have correctly understood the tasks. Moreover, based on comments by some participants, the cognitive tasks were perceived as games (the short-term memory task in particular), which may have supported motivation and attention (cf. Bratu et al., 2022).

However, given the random-guessing accuracy of several participants in the lexical decision task and the large variance of performance in the short-term memory task, which cannot be plausibly explained by differences in memory capacity alone, there are clearly still problems with some of the tasks. Particularly in the memory task, we suspect that performance was heavily influenced by task familiarity and individual learning curves. Some participants had to repeat the practice trial several times, while one participant, who performed very highly, remarked that they often played similar games. Choosing tasks with a high error tolerance (which the memory task was not) or using a larger number of trials may also yield more reliable results. Regarding the lexical decision task, it is unclear whether the three participants who always gave positive responses without any difference in RT between words and pseudowords misunderstood the task or lost motivation, since two of them did give some negative responses during the practice task. Further testing is necessary to determine how this task can be improved.

In this study, we refrained from displaying any feedback about correct or incorrect responses in the application, in order to avoid discouraging participants. However, depending on the difficulty of the task, it may be better to show feedback, especially if there is little to no personal supervision, to avoid misunderstanding and strengthen extrinsic motivation (cf. Rodríguez et al., 2022). In the future, we would also like to further develop the gamification elements and put more measures in place to monitor motivation or misunderstanding of instructions.

5. Conclusion and outlook

We presented *Okra*, a prototype mobile application for conducting reading experiments with people with IDs. Our primary goal was to provide a tool for researchers to enable digitized comprehensibility evaluation with target readers (instead of experts or general populations) by making use of the increased technological literacy among people with IDs, and ultimately lowering the threshold to including target groups in research on Easy Language and text simplification.

Therefore, our mobile application contributes to participation *in* digital technologies (Bosse, 2016) of persons with disabilities. At the same time, automatic text simplification as an assistive technology increases participation *through* digital technologies; here, more representative evaluations of texts in Easy Language of the kind made possible through our mobile application are capable of improving the quality of automatic text simplification models.

We also conducted a study with people with ID, testing the effect of modality (paper vs. iPad) on reading comprehension and subjective ratings and the feasibility of assessing cognitive skills in *Okra*. Although there was no evidence of a modality effect, we found that reading times were significantly longer on paper than on the iPad. Observations from this initial study confirm that it is feasible to use the application for evaluating Easy Language and basic cognitive assessment with this target group. However, we have identified several issues concerning usability and reliability of results, which we are going to address in future versions of the application. An additional limitation of our study is that we did not conduct any standardized testing of language competence or a detailed survey of reading habits. As a next step, we will conduct more systematic usability testing and use *Okra* to evaluate the output of human and automatic text simplification with people with ID.

While the experiments described in this paper were conducted in a highly controlled environment and with close supervision, we will also work to improve the usability and accessibility of the application to allow participants to use it more independently (ideally, outside of laboratory conditions), and to implement and test a wider range of task types. As a long-term goal, the user interface should also be made accessible for other target groups of Easy Language. Thus, we hope that it will become a tool for researchers to simplify and encourage the inclusion of people with disabilities.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by Ethics Committee of the Faculty of Arts and Social Sciences, University of Zurich. The patients/participants provided their written informed consent to participate in this study.

Author contributions

AS implemented the application, designed the experiment, prepared stimulus data, performed analyses, and wrote the manuscript. SE contributed to conceptualization and study design. SH-S, SG, SD, and LS contributed to study design and served as expert testers for the application. FH contributed to study design and data analysis. All authors contributed to stimulus material and manuscript revision, read, and approved the submitted version.

Funding

This work was carried out as part of the project capito automatisiert funded by the Austrian Research Promotion Agency General Programme under grant agreement number 881202.

Acknowledgments

We are greatly indebted to CFS GmbH (capito) for recruiting the participants, running the experiment, and helping to translate the material into Easy Language. We would also like to thank APA for providing the texts.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alonzo, O., Trussell, J., Dingman, B., and Huenerfauth, M. (2021). "Comparison of methods for evaluating complexity of simplified texts among deaf and hard-of-hearing adults at different literacy levels," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (New York, NY: Association for Computing Machinery). doi: 10.1145/3411764.3445038

Al-Thanyyan, S. S., and Azmi, A. M. (2021). Automated text simplification: a survey. *ACM Comput. Surv*. 54, 695. doi: 10.1145/3442695

Alva-Manchego, F., Scarton, C., and Specia, L. (2020). Data-driven sentence simplification: survey and benchmark. *Comput. Linguist*. 46, 135–187. doi: 10.1162/coli_a_00370

Alva-Manchego, F., Scarton, C., and Specia, L. (2021). The (un)suitability of automatic evaluation metrics for text simplification. *Comput. Linguist*. 47, 861–889. doi: 10.1162/coli_a_00418

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw*. 67, 1–48. doi: 10.18637/jss.v067.i01

Bosse, I. (2016). *Teilhabe in einer digitalen Gesellschaft - Wie Medien Inklusionsprozesse bef6rdern können*. Available online at: http://www.bpb.de/gesellschaft/medien/medienpolitik/172759/medien-und-inklusion (accessed May 22, 2013).

Bratu, M., Stan, S., and Muntean, C. H. (2022). "Benefits and limitations of using modern technologies for teaching STEM subjects to students with intellectual disabilities," in *2022 International Conference on Advanced Learning Technologies (ICALT)* (Bucharest: IEEE), 259–261. doi: 10.1109/ICALT55010.2022.00084

Chadwick, D. D., Chapman, M. J., and Caton, S. (2018). *Digital inclusion for people with an intellectual disability*. The Oxford Handbook of Cyberpsychology. doi: 10.1093/oxfordhb/9780198812746.013.17

Charzyńska, E., and Dębowski, Ł. J. (2015). Empirical verification of the polish formula of text difficulty. *Cognit. Stud*. 15. doi: 10.11649/cs.2015.010

Crossley, S. A., Yang, H. S., and McNamara, D. S. (2014). What's so simple about simplified texts? A computational and psycholinguistic investigation of text comprehension and text processing. *Read. Foreign. Lang*. 26, 92–113.

de Leeuw, J. (2015). jspsych: a javascript library for creating behavioral experiments in a web browser. *Behav. Res. Methods*. 47, 1–12. doi: 10.3758/s13428-014-0458-y

Deilen, S. (2020). "Visual segmentation of compounds in Easy Language: Eye movement studies on the effects of visual, morphological and semantic factors on the processing of German noun-noun compounds," in *Easy Language Research: Text and User Perspectives*, Hansen-Schirra, S., and Maaβ, C. (eds). Berlin: Frank & Timme. p. 241–256.

Deilen, S., and Schiffl, L. (2020). "Using eye-tracking to evaluate language processing in the easy language target group," in *Easy Language Research: Text and User Perspectives*, Hansen-Schirra, S., and Maaβ, C., (eds). Berlin: Frank & Timme. p. 273–281.

Eckes, T. (2015). *Introduction to Many-Facet Rasch Measurement*. Bern: Peter Lang.

Fajardo, I., Ávila, V., Ferrer, A., Tavares, G., Gómez, M., and Hernández, A. (2014). Easy-to-read texts for students with intellectual disability: linguistic factors affecting comprehension. *J. Appl. Res. Intellect. Disabil*. 27, 212–225. doi: 10.1111/jar.12065

Fulmer, S. M., D'Mello, S. K., Strain, A., and Graesser, A. C. (2015). Interest-based text preference moderates the effect of text difficulty on engagement and learning. *Contemp. Educ. Psychol*. 41, 98–110. doi: 10.1016/j.cedpsych.2014.12.005

Gooding, S., Berzak, Y., Mak, T., and Sharifi, M. (2021). "Predicting text readability from scrolling interactions," in *Proceedings of the 25th Conference on Computational Natural Language Learning*. Toronto: Association for Computational Linguistics. p. 380–390. doi: 10.18653/v1/2021.conll-1.30

Gutermuth, S. (2020). *Leichte Sprache für alle? Eine zielgruppenorientierte Rezeptionsstudie zu Leichter und Einfacher Sprache [Easy language for everyone? A target* *group oriented reception study of "Leichte Sprache" and "Einfache Sprache"*]. Berlin: Frank & Timme.

Huenerfauth, M., Feng, L., and Elhadad, N. (2009). "Comparing evaluation techniques for text readability software for adults with intellectual disabilities," in *Proceedings of the 11th International ACM SIGACCESS Conference on Computers and Accessibility*. p. 3–10. doi: 10.1145/1639642.1639646

Jung, S., Ousley, C. L., Mcnaughton, D., and Wolfe, P. S. (2021). The effects of technology supports on community grocery shopping skills for students with intellectual and developmental disabilities: a meta-analysis. *J. Spec. Educ*. 37, 351–362. doi: 10.1177/0162643421989970

Keuleers, E., and Brysbaert, M. (2010). Wuggy: a multilingual pseudoword generator. *Behav. Res. Methods* 42, 627–633. doi: 10.3758/BRM.42.3.627

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *J. Exp. Psychol*. 55, 352–358. doi: 10.1037/h0043688

Kong, Y., Seo, Y. S., and Zhai, L. (2018). Comparison of reading performance on screen and on paper: A meta-analysis. *Computers & Education* 123:138–149. doi: 10.1016/j.compedu.2018.05.005

Leroy, G., Endicott, J. E., Kauchak, D., Mouradi, O., and Just, M. (2013). User evaluation of the effects of a text simplification algorithm using term familiarity on perception, understanding, learning, and information retention. *J. Med. Internet Res*. 15, e144. doi: 10.2196/jmir.2569

Leroy, G., Kauchak, D., Haeger, D., and Spegman, D. (2022). Evaluation of an online text simplification editor using manual and automated metrics for perceived and actual text difficulty. *JAMIA Open*. 5, ac044. doi: 10.1093/jamiaopen/ooac044

Linacre, J. M. (1994). *Many-facet Rasch Measurement*. San Diego, CA: MESA Press.

Maaβ, C. (2020). *Easy Language-Plain Language-Easy Language* Plus: Balancing comprehensibility and acceptability. Berlin: Frank & Timme. doi: 10.26530/20.500.12657/42089

Maebara, K., Yamaguchi, A., Suzuki, T., and Imai, A. (2022). A qualitative study on the function of information and communication technology utilization in teaching students with intellectual disabilities: implications for techniques of teaching/job coaching. *J. Intell. Disabil. – Diag. Treat*. 23, 209–227. doi: 10.6000/2292-2598.2022.10.01.2

Martin, A. J., Strnadová, I., Loblinzk, J., Danker, J., and Cumming, T. M. (2021). The role of mobile technology in promoting social inclusion among adults with intellectual disabilities. *J. Appl. Res. Intellect. Disabil*. 34, 840–851. doi: 10.1111/jar.12869

Ockey, G. J. (2021). "Item response theory and many-facet Rasch measurement," in *The Routledge Handbook of Language Testing*. Oxfordshire: Routledge. p. 462–476. doi: 10.4324/9781003220756-36

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., et al. (2019). PsychoPy2: Experiments in behavior made easy. *Behav. Res. Methods*. 51, 195–203. doi: 10.3758/s13428-018-01193-y

Perkuhn, R., Belica, C., Kupietz, M., Keibel, H., and Hennig, S. (2009). *DeReWo: Korpusbasierte Wortformenliste*. Kostroma: DeReWo.

Ramsten, C., Martin, L. K., Dag, M., and Hammar, L. M. (2018). Information and communication technology use in daily life among young adults with mild-to-moderate intellectual disability. *J. Intell. Disabil*. 24:289–308. doi: 10.1177/1744629518784351

Redmiles, E., Maszkiewicz, L., Hwang, E., Kuchhal, D., Liu, E., Morales, M., et al. (2019). "Comparing and developing tools to measure the readability of domain-specific texts," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4831-4842, Hong Kong, China*. Hong Kong, China: Association for Computational Linguistics.

Reitan, R., and Wolfson, D. (1993). *The Halstead-Reitan neuropsychological Test Battery: Theory and Clinical Interpretation*. Herndon, VA: Neuropsychology Press.

Rello, L., Baeza-Yates, R., Bott, S., and Saggion, H. (2013a). "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: Association for Computing Machinery). doi: 10.1145/2461121.2461126

Rello, L., Baeza-Yates, R., Dempere-Marco, L., and Saggion, H. (2013b). "Frequent words improve readability and short words improve understandability for people with dyslexia," in *Human-Computer Interaction – INTERACT 2013*, eds P. Kotze, G. Marsden, G. Lindgaard, J. Wesson, and M. Wincler (Berlin; Heidelberg: Springer), 203–219. doi: 10.1007/978-3-642-40498-6_15

Rello, L., Bautista, S., Baeza-Yates, R., Gervás, P., Hervás, R., and Saggion, H. (2013c). "One half or 50%? an eye-tracking study of number representation readability," in *Human-Computer Interaction – INTERACT 2013*, eds P. Kotze, G. Marsden, G. Lindgaard, J. Wesson, and M. Wincler (Berlin; Heidelberg: Springer), 229–245.

Rodríguez, F., de Blume, A. G., and Soto, C. (2022). Effects of reading motivation and meta-comprehension on the reading comprehension of students with intellectual disabilities. *Elect. J. Res. Educ. Psychol*. 30.

Saggion, H., Štajner, S., Bott, S., Mille, S., Rello, L., and Drndarevic, B. (2015). Making it Simplext: Implementation and evaluation of a text simplification system for Spanish. *ACM Trans. Access. Comput*. 6, 8046. doi: 10.1145/2738046

Säuberli, A. (2021). "Measuring text comprehension for people with reading difficulties using a mobile application," in *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS '21*. New York, NY: Association for Computing Machinery.

Säuberli, A., Ebling, S., and Volk, M. (2020). "Benchmarking data-driven automatic text simplification for German," in *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*. Marseille, France: European Language Resources Association. p. 41–48.

Schulz, R., Degenhardt, J., and Czerner-Nicolas, K. (2020). "Easy language interpreting," in *Easy Language Research: Text and User Perspectives*, Hansen-Schirra, S. and Maaß, C. (eds). Berlin: Frank & Timme. p. 163–178.

Štajner, S. (2021). "Automatic text simplification for social good: Progress and challenges," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Toronto: Association for Computational Linguistics. p. 2637–2652.

Stodden, R. (2021). "When the scale is unclear - analysis of the interpretation of rating scales in human evaluation of text simplification," in *Proceedings of the First Workshop on Current Trends in Text Simplification (CTTS 2021) co-located with the 37th Conference of the Spanish Society for Natural Language Processing (SEPLN2021)*, Saggion, H., Štajner, S., Ferrés, D., and Sheang, K. C. (eds). Málaga: CEUR Workshop Proceedings. CEUR-WS.org.

Stoet, G. (2017). PsyToolkit: a novel web-based method for running online questionnaires and reaction-time experiments. *Teach. Psychol*. 44, 24–31. doi: 10.1177/0098628316677643

Sulem, E., Abend, O., and Rappoport, A. (2018a). "BLEU is not suitable for the evaluation of text simplification," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics. p. 738–744.

Sulem, E., Abend, O., and Rappoport, A. (2018b). "Semantic structural evaluation for text simplification," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics. p. 685–696.

Sulem, E., Abend, O., and Rappoport, A. (2018c). "Simple and effective text simplification using semantic and neural methods," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics. p. 162–173.

Uzor, S., Jacques, J. T., Dudley, J. J., and Kristensson, P. O. (2021). "Investigating the accessibility of crowdwork tasks on mechanical turk," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21*. New York, NY: Association for Computing Machinery.

Vajjala, S., and Lucic, I. (2019). "On understanding the relation between expert annotations of text readability and target reader comprehension," in *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Florence, Italy: Association for Computational Linguistics. p. 349–359.

Vajjala, S., Meurers, D., Eitel, A., and Scheiter, K. (2016). "Towards grounding computational linguistic approaches to readability: Modeling reader-text interaction for easy and difficult texts," in *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*. Osaka, Japan: The COLING 2016 Organizing Committee. p. 38–48.

Xu, W., Napoles, C., Pavlick, E., Chen, Q., and Callison-Burch, C. (2016). Optimizing statistical machine translation for text simplification. *Trans. Assoc. Computat. Ling*. 4, 401–415. doi: 10.1162/tacl_a_00107

Zhao, Y., Chen, L., Chen, Z., and Yu, K. (2020). Semi-supervised text simplification with back-translation and asymmetric denoising autoencoders. *Proc. Innov. Appl. Artif. Intell. Conf*. 34, 9668–9675. doi: 10.1609/aaai.v34i05.6515