# Beyond Readability Metrics: Plain Language Priorities in Disability Advocacy Organizations

ANUKRITI KUMAR, University of Washington, USA
KATE GLAZKO, University of Washington, USA
YUERAN SUN, University of Washington, USA
MARK HARNISS, University of Washington, USA
LUCY LU WANG, University of Washington, USA
JENNIFER MANKOFF, University of Washington, USA

Plain language materials enable people with intellectual and developmental disabilities (IDD) to access critical information about policy, healthcare, and civic participation. Disability advocacy organizations routinely produce these materials, yet we know little about how practitioners approach this work, what standards guide their judgments, or whether current evaluation metrics align with their priorities. Through focus groups and interviews with 11 practitioners across three U.S. disability advocacy organizations, individual walkthroughs where practitioners evaluated AI-simplified documents, and systematic analysis of 33 pairs of original and simplified documents from four organizations using 28 readability metrics, we document plain language production as specialized expertise requiring policy knowledge, community accountability, and multi-stage validation processes. Practitioners who use AI tools report treating outputs as provisional starting points requiring complete human verification rather than as publication-ready content. Organization-produced documents averaged a Flesch-Kincaid Grade Level of 10.2, exceeding all published guideline targets ranging from 3rd to 8th grade, yet practitioners described these materials as successfully meeting community needs. This suggests that published text simplification guidelines may not capture dimensions practitioners and communities consider essential for high-stakes accessibility work. Based on our findings, we propose design principles for text simplification tools that center verification and transparency rather than automation, and call for evaluation frameworks that complement automated metrics with practitioner expertise and community accountability mechanisms.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; *Accessibility systems and tools.*

Additional Key Words and Phrases: Plain language, Text simplification, Intellectual and developmental disabilities, Cognitive accessibility, Automated or linguistic metrics, Disability advocacy, Participatory design

**ACM Reference Format:**
Anukriti Kumar, Kate Glazko, Yueran Sun, Mark Harniss, Lucy Lu Wang, and Jennifer Mankoff. 2026. Beyond Readability Metrics: Plain Language Priorities in Disability Advocacy Organizations. https://doi.org/10.1145/3805689.3806722

## 1 Introduction

Access to information is a prerequisite for meaningful participation in civic, social, and political life. For people with intellectual and developmental disabilities (IDD), however, this access is systematically constrained by the complexity of written materials in domains such as public policy, healthcare, voting, and government services. Plain language, which aims to preserve the full meaning and intent of a text while making it easier to read and understand, has long been recognized as critical for accessibility [15, 46]. Despite legal mandates like the Plain

Authors' Contact Information: Anukriti Kumar, anukumar@uw.edu, University of Washington, Seattle, WA, USA; Kate Glazko, glazko@cs.washington.edu, University of Washington, Seattle, WA, USA; Yueran Sun, yuerans@uw.edu, University of Washington, Seattle, WA, USA; Mark Harniss, mharniss@uw.edu, University of Washington, Seattle, WA, USA; Lucy Lu Wang, lucylw@uw.edu, University of Washington, Seattle, WA, USA; Jennifer Mankoff, jmankoff@uw.edu, University of Washington, Seattle, WA, USA.

Writing Act of 2010, Section 508 and ADA [1, 5, 38], and decades of advocacy, plain language versions of complex documents remain rare, inconsistently implemented, and labor-intensive to produce.

Generative AI systems now offer the possibility of addressing these gaps. People with IDD, family members, self-advocates, and disability advocacy organizations increasingly use large language models (LLMs) [23] to rewrite policies, summarize legislation, and interpret public information. Products marketed for cognitive accessibility have begun incorporating generative AI as core features [2, 44, 45]. Yet this deployment occurs with limited understanding of how such systems align with the expertise and accountability structures of those already responsible for producing accessible content in high-stakes contexts.

Disability advocacy organizations routinely engage in plain-language work as part of their core mission, translating legislation, policies, and public communications for people with IDD [37]. This work is deeply collaborative and shaped by ethical commitments to accuracy, dignity, and accountability to disability communities. Practitioners must balance competing demands: making text accessible while preserving legal precision, meeting elementary reading level targets while maintaining informativeness and completeness, and achieving efficiency while ensuring trustworthiness. Yet little empirical research has examined how these practitioners currently approach plain language production or what systematic gaps emerge when current evaluation paradigms are applied to their work.

Most text simplification research evaluates models using offline benchmarks [35, 39], automatic metrics [48, 60], or reader-focused comprehension measures [8]. These approaches rarely ground analysis in the practices of practitioners who produce accessible content for real-world, high-stakes contexts. When simplified text misrepresents or omits content, consequences may include misinformation, loss of agency, or exclusion from decision-making processes [50, 51], affecting disabled people's lives, rights, and resources. Understanding how practitioners navigate these risks and whether current evaluation metrics capture their priorities is essential for developing AI systems that support rather than undermine accessibility goals.

Our work addresses this gap through a community-engaged, mixed-methods study with three U.S.-based disability advocacy organizations that routinely produce plain-language materials and are directly accountable to disabled communities. We examine the following questions:

**RQ1:** How do disability advocacy organizations currently produce plain-language materials for people with IDD? What values and standards guide their work, and how do they perceive and use (or resist) generative AI tools?

**RQ2:** To what extent do current readability and linguistic metrics (e.g., Flesch-Kincaid Grade Level (FKGL), sentence length, word commonality) capture the dimensions practitioners prioritize in their workflows?

**RQ3:** What design principles and accountability mechanisms are needed so AI-assisted plain language tools align with advocates' values and obligations to their communities?

We conducted focus groups and semi-structured interviews with disability advocacy organizations, individual walkthroughs where practitioners evaluated simplified documents, and an analysis of documents from four organizations using automated and linguistic readability metrics. Through this work, we make three contributions:

- **Empirical documentation of practitioner-centered plain-language work** as specialized expertise requiring policy knowledge and disability community accountability. Drawing on focus groups with 11 practitioners across three organizations, we show that plain language work is deeply collaborative, shaped by ethical obligations, community-led standards, and multi-stage validation. Practitioners consistently position AI as a provisional starting point generating content requiring human verification, never as an autonomous producer of publication-ready outputs.
- **Document analysis demonstrating that automated and linguistic metrics inadequately capture dimensions that practitioners prioritize for readers with IDD.** Through quantitative analysis of 33 pairs

of documents (original and human-simplified) collected from four disability advocacy organizations, we show that metrics commonly used to evaluate text simplification (e.g., FKGL, sentence length) may not capture the qualities considered to be most important by practitioners for plain language. Dimensions practitioners prioritize include accuracy, definitional support, information completeness, and structural appropriateness.

- **Design principles centering human expertise, enabling transparency, and supporting disability justice commitments.** Grounded in practitioners' expressed needs and observed breakdowns, we articulate sociotechnical requirements for AI systems that support human expertise in this domain. These principles emphasize transparency, accountability, and human agency.

When systems are optimized for measurable proxies without capturing dimensions practitioners prioritize, there is risk of producing outputs that appear accessible by metric standards while failing to meet actual accessibility needs. We argue for evaluation paradigms that center practitioner expertise and community needs rather than relying solely on automated metrics, with implications for participatory design of AI systems meant to serve accessibility goals.

The remainder of this paper is organized as follows. Section 2 discusses related work in text simplification, accessibility research, and human-AI collaboration. Section 3 describes our study design, participant organizations, and research methods. Section 4 presents findings organized around our three research questions. Section 5 discusses implications for AI development, accessibility practice, and policy, followed by limitations and future directions in Section 6.

## 2 Background and Related Work

### 2.1 Plain Language Standards and Distinctions for the Population with IDD

Several organizations and governmental bodies have established guidelines for producing accessible text. Disability-led guidance such as Inclusion Europe's *Information for all* [31] and the UK government's Easy Read standards [53] provide specific recommendations for creating materials for people with cognitive disabilities, emphasizing concrete language, short sentences, consistent vocabulary, and clear structure. In the United States, the Plain Writing Act of 2010 [54] requires federal agencies to use clear communication that the public can understand. The Federal Plain Language Guidelines [55] provide implementation guidance targeting general audiences at a sixth to eighth grade reading level.

Throughout this paper, we use "plain language" as our partner organizations do: to describe materials designed to make complex text accessible to people with IDD. The distinction between plain language for general audiences and accessible text for people with IDD is important. ISO 24495-1:2023 [32] explicitly notes that plain language differs from "easy language" or "Easy Read" materials designed for people with cognitive disabilities. Research has shown that text simplified according to general plain language principles may not be sufficiently accessible for readers with IDD. Yaneva et al. [61] found that Simple Wikipedia, often used as a simplification corpus, is more complex than materials specifically written and validated for readers with cognitive disabilities. Fajardo et al. [19] demonstrated that commonly proposed simplification strategies, such as adding connectives or substituting high-frequency words, do not reliably improve inferential comprehension for readers with intellectual disability. Effects depend on the type and familiarity of linguistic features and the specific task context.

Another alternative for readers with IDD is the Minimum Text Complexity Guidelines [20], which are more targeted, specifying reading levels at or below third grade, use of at least 92% high-frequency words, and constraints on sentence structure and vocabulary. These guidelines reflect substantially different requirements compared to plain language guidance targeting general audiences.

## 2.2 Automatic Text Simplification

Natural language processing research on automatic text simplification has evolved from rule-based systems to neural approaches [9, 39, 60]. Recent transformer-based approaches achieve high fluency [35, 36], with Kew et al. [35] demonstrating that large language models (LLMs) like GPT-4 perform comparably to supervised baselines without task-specific training. Srikanth and Li [47] introduced "elaborative simplification," finding that 78% of simplifications in the Newsela corpus [59] involved adding explanations rather than only removing complexity. This approach differs from reduction-focused simplification and may be particularly relevant for readers who benefit from additional context. Recent work examining text simplification in municipal and governmental contexts [12, 56] has found that professional content creators working with high-stakes communication require tools that support rather than replace humans.

Evaluation of text simplification has primarily relied on *n*-gram-based automatic metrics like SARI [60] or BLEU [40], or readability formulas like FKGL. However, several studies have identified significant limitations in current evaluation approaches. Guo et al. [27] conducted a meta-evaluation of 14 automated metrics for plain language summarization over four quality criteria: informativeness, simplification, coherence, and faithfulness, finding that no single metric captures all criteria simultaneously. Joseph et al. [34] examined factuality in plain language summarization of medical evidence through their FactPICO benchmark, finding that existing metrics correlate poorly with expert judgments of accuracy. Agrawal and Carpuat [8] used reading comprehension questions to evaluate meaning preservation, showing that even the best systems struggle with a substantial portion of questions. Most relevant to this work, Cripwell et al. [17] found that metrics like SARI and BLEU poorly predict simplification quality for readers with intellectual disabilities.

## 2.3 Generative AI for Text Simplification: Adoption and Concerns

Persistent gaps in the availability of accessible materials have driven interest in using generative AI for text simplification. Several studies have examined LLM-based approaches specifically for people with intellectual disabilities. Weijers et al. [58] evaluated LLMs including GPT-4o and Llama3 for Dutch text simplification, finding that producing simplified text is only one component of successful simplification, with user comfort and personalization also playing important roles. Gao et al. [22] incorporated preference feedback from people with IDD during model training, demonstrating the value of participatory approaches in developing these systems.

Despite technical advances, substantial concerns remain about using generative AI for plain language production in high-stakes contexts. The Autistic Self Advocacy Network issued a public statement [11] arguing against using generative AI to write plain language materials, citing concerns that model-generated rewrites can introduce errors or meaning changes, omit important details, and create text that appears simpler but is less accurate or usable for disabled readers. Glazko et al. [25] documented instances where generative AI produced incorrect or ableist conclusions when summarizing disability-related materials, noting that such errors can be particularly difficult for disabled users to verify when the original documents themselves are inaccessible.

Research on human-AI collaboration has shown that verification of AI outputs presents challenges. Studies have found that people can defer to AI recommendations and exhibit overreliance even when systems produce errors [28, 41]. Little empirical research has examined how professionals who produce plain language materials in high-stakes contexts currently approach this work and emerging AI tools. Our work addresses these gaps through empirical investigation with disability advocacy organizations that routinely produce plain language materials for people with IDD.

Our work complements a growing body of disability-focused scholarship at FAccT, including work on disability bias in large language models [21], AI-mediated hiring discrimination against disabled people [24], AI attitudes among disabled populations [29], ableism detection [42], and demographic stereotype amplification [13]. While this prior work examines bias, representation, and discrimination; it does not directly address cognitive accessibility, plain language production, or the accountability challenges specific to practitioners serving people with

| Phase | Time | Format | N | Activities |
|---|---|---|---|---|
| **Phase 1: Focus Groups** | August–September 2025 | Group discussions | 11 people (Organization A = 5, Organization B = 4, Organization C = 2) | Focus group discussions exploring current plain language practices, AI experiences, values, challenges, and validation processes |
| **Phase 2: Individual Walk-throughs** | August–September 2025 | Group sessions | 5 people (Organization A = 2, Organization B = 2, Organization C = 1) | Think-aloud walkthroughs with real-world documents, evaluating AI-generated plain language outputs |
| **Phase 3: Document Analysis** | October–November 2025 | Document analysis | 33 documents (Organization A = 8, Organization B = 11, ASAN = 5, CDT = 9) | Systematic quantitative analysis comparing original and organization-crafted plain language documents |

Table 1. Overview of study phases, timeline, participants, and key activities.

IDD. Our paper contributes a complementary perspective: the evaluation validity problem in accessibility-critical AI workflows, and the verification labor borne by practitioners accountable to disability communities.

## 3 Methodology

### 3.1 Study Design Overview

Table 1 presents study phases, timing, and participants. We proceeded in three sequential phases: focus group discussions exploring current practices, values, and quality criteria (Phase 1); individual walkthroughs where practitioners reviewed simplified materials, surfacing priorities through think-aloud observations (Phase 2); and document analysis comparing organizational plain language to original texts using automated metrics (Phase 3).

*Study Scope and Boundaries.* This study centers practitioners who produce plain language materials, not people with IDD who use these materials. While practitioners' expertise reflects accumulated community feedback, direct investigation of how IDD users assess quality and whether practitioner priorities predict comprehension outcomes remains essential future work. Our findings characterize professional plain language practices in U.S. advocacy contexts and metric-practice alignment, but cannot make claims about end-user comprehension or preferences.

This study was reviewed by the University of Washington Institutional Review Board (IRB) and deemed exempt under applicable federal regulations. Consent was obtained from all participants; participation in this study was on a voluntary basis and not compensated.

### 3.2 Partner Organizations

We partnered with three U.S. disability advocacy organizations selected through purposive sampling based on four criteria: (1) routine production of plain language materials for people with IDD, (2) diverse geographic scope and constituencies, (3) different document types (policy, legislative, research), and (4) willingness to reflect critically on their practices and priorities. They happened to vary in levels of AI adoption, which allowed us to capture diverse perspectives on emerging technology use in plain language production. Organizations are anonymized per IRB requirements. All participants provided informed consent and were informed they could withdraw at any time.

**Organization A** is a federally-mandated state-level developmental disabilities council, producing plain language versions of council meeting materials, policy documents, budget reports, and educational resources for council members (including self-advocates with IDD) and the broader disability community.

**Organization B** is a regional advocacy organization serving disabled people across over a dozen U.S. states, specializing in plain language translation of legislative bills, ballot measures, and policy documents to facilitate political participation by disabled constituents.

**Organization C** is a national network of university-based interdisciplinary centers advancing policy and practice for people with developmental disabilities, producing plain language versions of research reports, policy newsletters, academic posters, and educational materials for national audiences.

## 3.3 Phase 1: Focus Group Discussions

Focus groups served as our primary data source for RQ1, documenting the expertise, values, and accountability structures underlying professional plain language practice.

*3.3.1 Participants.* Eleven practitioners participated across three organizations:

- **Organization A (N=5):** Senior staff with decades of experience in disability advocacy, council coordinators responsible for meeting materials, and staff members who regularly interface with self-advocates
- **Organization B (N=4):** Policy analysts responsible for legislative tracking and translation, advocacy specialists, and organizational leadership
- **Organization C (N=2):** Senior staff member and a member with IDD responsible for accessibility across communication channels, including social media, policy materials, and academic outputs

We recruited participants through organizational leadership who connected us with staff actively involved in plain language work. Participants had varying levels of prior AI experience. Some used ChatGPT or similar tools daily; others had experimented with AI but remained primarily reliant on manual processes; Organization B had explored AI but discontinued use due to quality concerns. All participants had expertise in plain language principles, familiarity with established guidelines (e.g., Plain Language Action and Information Network, Autistic Self Advocacy Network resources), and direct accountability relationships with disability communities.

*3.3.2 Procedure.* We conducted 60-90 minute focus group discussions with each organization during Aug-Sep 2025 via video conference. Focus groups followed a semi-structured protocol (Appendix A) covering six domains:

- **Current plain language practices:** frequency, document types, workflows, tools used (manual and technological)
- **Guidelines and standards:** resources guiding work, validation approaches, target reading levels
- **Values and principles:** concerns about output quality, accountability to communities
- **AI usage considerations:** whether and how they use AI, prompting strategies, perceived benefits, and limitations
- **Validation processes:** what is assessed when evaluating simplified text, how they know when materials are ready, feedback mechanisms from disability communities
- **Challenges and needs:** Resource constraints, workflow pain points, desired tool

All sessions were recorded with permission and transcribed using TurboScribe [52] with manual correction. At least two researchers attended each session to enable real-time clarification and follow-up probes.

*3.3.3 Qualitative Data Analysis.* The first and second authors analyzed focus group transcripts using reflexive thematic analysis following Braun and Clarke's six-step approach [16], resulting in 10 themes. We chose this inductive, data-driven method to identify patterns in practitioner perspectives without imposing predetermined

frameworks. Throughout analysis, we practiced reflexivity by discussing how our assumptions shaped interpretation and memoing about findings that challenged our expectations. Complete details of our analytical process appear in Appendix B.

## 3.4 Phase 2: Individual Walkthroughs

Individual walkthroughs provided detailed insight into how practitioners assess simplified text quality, yielding rich qualitative observations for RQ2 and RQ3.

*3.4.1 Participants and Materials.* From the 11 focus group participants, 5 practitioners who engaged in daily simplification work participated in individual walkthroughs: 2 from Organization A, 2 from Organization B, and 1 from Organization C. Each selected 1-5 documents their organization had previously simplified, yielding 12 documents (3 from Organization A, 4 from Organization B, and 5 from Organization C).

*3.4.2 Procedure.* Walkthroughs occurred via video conference (30-45 minutes per organization) following the protocol in Appendix C. For each document, we presented AI-simplified versions generated using GPT-4o with a prompt incorporating disability-led plain language guidelines (full prompt in Appendix D). These materials served as probes to elicit practitioner thinking about quality and surface what practitioners attend to when evaluating simplified text.

Practitioners reviewed each simplified version using think-aloud protocol, verbalizing their thoughts while examining the materials. We used minimal interviewer intervention with neutral prompts ("What are you thinking now?" "Can you tell me more?") when participants fell silent. After each document, practitioners provided open-ended feedback responding to: "If you would edit this text, what would be your main reasons?" Sessions were recorded and transcribed.

*3.4.3 Data Analysis.* Two researchers independently coded practitioner open-ended responses and think-aloud observations to identify recurring quality concerns. We developed a framework categorizing:

- *Missing definitions*: Technical terms, jargon, or domain-specific language appearing without explanation
- *Information erasure*: Summarization, condensation, or omission of details practitioners deemed necessary
- *Accuracy issues*: Subtle meaning changes, misrepresentation of source intent, or altered legal/policy language
- *Structural misalignments*: Formatting, organization, or presentation inconsistent with organization-specific conventions

We calculated inter-rater reliability for identifying these observation types (Cohen's kappa = 0.82) and resolved disagreements through discussion. We identified dimensions practitioners consistently emphasized that existing automated and linguistic metrics do not capture; these dimensions informed our analytical approach in Phase 3.

## 3.5 Phase 3: Document Analysis

We then conducted quantitative analysis of 33 documents comparing original complex texts with organization-crafted plain language versions across 28 automated and linguistic metrics.

*3.5.1 Document Corpus.* We assembled a corpus of 33 documents from four sources, each providing both original complex text and organization-crafted plain language versions:

- **Organization A (N=8):** Meeting materials, policy documents, budget reports provided by partner organization
- **Organization B (N=11):** Legislative bill summaries provided by partner organization
- **Autistic Self Advocacy Network (ASAN) (N=5):** Publicly available policy fact sheets [3]
- **Center for Democracy & Technology (CDT) (N=9):** Publicly available digital rights materials [6]

We included ASAN and CDT documents to extend analysis beyond our partner organizations and examine community-created resources representing best practices in disability-led plain language.

*3.5.2 Automated and Linguistic Metrics.* We analyzed these original and plain language documents using 28 metrics in six categories. Metrics were selected to provide comprehensive coverage of features specified across four complementary accessibility frameworks: the Federal Plain Language Guidelines (PLAIN) [55], Inclusion Europe's Easy Read standards [31, 53], the Minimum Text Complexity (MTC) Framework [7], and WCAG's Cognitive and Learning Disabilities Accessibility guidelines (COGA) [57]. We computed these across organizational documents using custom Python scripts, examining: (1) How organization-produced versions measure compared to originals, (2) Whether organization versions meet metric-based targets (e.g., FKGL 3.0-8.0), and (3) How practitioners describe the success of these materials despite metric performance.

Each metric maps to at least one framework's explicit recommendations as documented in Table 3. For example, average words per sentence (target $\leq 15$) maps to both MTC's specification and WCAG COGA guidance; passive voice percentage (<5%) reflects PLAIN and Easy Read preferences for active constructions; and type-token ratio ($\leq 0.4$) operationalizes Easy Read's principle of consistent vocabulary.

*3.5.3 Procedure.* We conducted two types of analyses examining metric-practice alignment:

**Analysis 1: Organizational plain-language documents (33 documents) on 28 metrics.** We computed descriptive statistics for each metric across organizations. This analysis establishes what organizations actually produce and how their materials perform on metrics commonly used to evaluate text simplification.

**Analysis 2: Practitioner quality assessments and metric alignment.** We analyzed the set of 12 documents from Phase 2 walkthroughs. We examined practitioners' open-ended responses and think-aloud observations, creating a list of their concerns and identifying dimensions that are not yet measured by metrics..

## 4 Findings

### 4.1 RQ1: Practitioner Practices, Values, and Priorities

While all organizations shared commitment to accessibility for people with IDD, different missions shaped practices. Organization A (state council) emphasized individual expertise with council feedback and 5th-6th grade targets; Organization B (regional legislative advocacy) implemented structured multi-stage workflow with dual-analyst review and 4th-5th grade targets; Organization C (national research-to-practice) prioritized collaborative creation with self-advocates as co-leaders and flexible targets. Despite differences, common patterns emerged across their practices.

*4.1.1 Multi-Stage, Collaborative Workflows with Essential Human Oversight.* Plain language production was described as iterative work rather than one-step conversion. Organization B had a six-stage process: scanning bills by state, colleague review for relevance, document creation with source linking, automated tools (Goblin Tools [4], Diff It [18]) for initial draft, applying policy expertise for simplification, and dual-analyst policy review. Practitioners emphasized the limits of automation: "There's a lot of manual work involved. And I don't see that changing…I think there still has to be some type of human element involved." (P1, Organization A).

**Collaboration took different forms across organizations.** Organization C emphasized group-based creation with self-advocates as co-leaders: "It's actually a group effort…mostly manual when we are making things in plain language" (P2). Organization A relied on individual expertise with council feedback loops. Organization B implemented dual-analyst review combining policy and accessibility expertise.

**Frequency and volume create capacity challenges.** Organizations reported daily or near-daily production during active periods, with Organization B noting: "Probably daily for the most part, especially when it's legislative session time" (P5, Organization B). This volume, combined with multi-stage workflows, creates substantial time pressure: "Sometimes we are crunched for time and we got to do things quickly" (P6, Organization A).

*4.1.2 Core Values and Principles.* Five values emerged as non-negotiable across organizations:

**Accuracy as foundational.** Maintaining source text integrity superseded all other considerations: "To ensure that I'm using the correct terminology, and most importantly that the integrity of the original source text is maintained in that translation" (P3, Organization B). This manifested in systematic validation: "I just look for if the information is all there…And then I look for just weird errors that might come up" (P7, Organization A).

**Transparency through source linking.** Organizations preserved connections to source materials: "Copy the original text into the document under the simplified text, just so it's clearly visible and transparent" (P3, Organization B). This enabled verification and accountability to communities.

**Context preservation.** Plain language required explaining significance, not just linguistic transformation: "That one paragraph is part summarization, but part context. It's set up so that the council members understand what we're doing" (P7, Organization A). Organization B required complete enumeration: "The phrasing 'and more' in a translation; we would want it to be complete. So we would want to actually list out the 'and more.' And then define the pieces in that list" (P3, Organization B).

**Lived experience and disability-led practice.** Organizations prioritized resources created by and with disability communities: "We try to use resources that are created by community members for and with the community. So everyone who's made this involved at least a person with disability" (P1, Organization C). All organizations validated materials through community feedback.

**Universal accessibility.** Practitioners recognized plain language benefits extend beyond people with IDD: "When you make something accessible, it's not just first-hand with an IDD, it's people whose language is English as their second language" (P1, Organization C).

*4.1.3 Perceptions of AI tools.* Practitioners using AI expressed nuanced perspectives, neither complete adoption nor rejection, but context-dependent, cautious optimism grounded in experience with both benefits and limitations.

**AI as starting point, not final product.** The most common framing positioned AI as generating initial drafts requiring human verification: "Trust is a hard word because it's more of it gives us a different version. It's giving me a summary, but it's not exactly giving me exactly what I want, but it's giving me a start or it's giving me a frame of reference" (P1, Organization C). P7 explained: "It would spit out the document, and it wouldn't quite be what I was looking for. And so then I would work [with it] to phrase it in other ways until it came back how I wanted it. And at first…some documents could take five or six different tries" (P7, Organization A).

**Learning curve.** Practitioners described iteratively learning AI tools: "I first approached ChatGPT…more like a Google tool, kind of a search engine mindset…As I understood that it was a conversational model, that's when it kind of really took off" (P7, Organization A). Some practitioners invested significant effort while others discontinued use.

**Domain-specific effectiveness.** AI tools worked better for some tasks than others. One practitioner contrasted: "For things like a poster, an academic poster that someone has not given us the image description for…Yes, I will be using AI because it will give me a good structure…it saves time" versus challenges with policy text where "I haven't been satisfied enough with the outputs to continue using AI for that purpose currently" (P4, Organization B). Organization B discontinued use after determining verification burden exceeded the benefits.

We also acknowledge practitioner positionality. Our participants are deeply committed experts with accumulated community knowledge, but they operate within institutional roles that shape their perspectives, including incentives tied to demonstrating organizational efficiency, accountability to funders, and professional identities grounded in plain language expertise. These role-based pressures may influence how practitioners characterize AI utility and verification burden. We treat their accounts as essential expert testimony while holding this contextual nuance as important interpretive context for our findings.

*4.1.4 Persistent Challenges and Barriers of Using AI.* Practitioners also identified substantial challenges that limit AI effectiveness in their workflows.

**Iteration fatigue and inconsistency.** When tools did not produce acceptable outputs initially, repeated attempts could become frustrating: "When it's not getting what you want, it will regenerate the same things, but differently. So that sometimes is a struggle" (P2, Organization C). One practitioner noted initial documents "could take five or six different tries" (P7, Organization A), raising questions about actual time savings.

**Validation burden.** Even when tools generated outputs, verification required substantial effort: "I just look for if the information is all there…And then I look for just weird errors…like the organization's name gets changed or something. And then sometimes I even check it to see if there's any jargon in it" (P7, Organization A). For high-stakes policy and legislative materials, this verification cannot be shortcuts: "Our really big concern…is accuracy. Just like maintaining accuracy with the original source content." (P4, Organization B).

**Tool limitations and workflow friction.** Current tools created workflow obstacles beyond output quality. Format issues required manual work: "If it goes into a text document, then we'd have to download that. And if we wanted to do additional formatting, we'd have to move it from text to word." (P6, Organization A). Practitioners reported a lack of necessary features directly in the chatbot interface, e.g., track changes, definition support, formatting preservation, requiring workarounds or manual completion of tasks.

*4.1.5 Desired Features.* Practitioners identified features that would support their expertise rather than replace it.

**Track changes and transparency.** Practitioners wanted visibility into modifications: "That's actually a really helpful feature, because the current tools that we're using do not have that feature…that element of track change…is helpful" (P3, Organization B).

**Definition support.** Rather than automated definitions (which might be inaccurate), practitioners wanted support identifying terms needing explanation and managing organizational definition glossaries: "It would be cool to…be able to just like write something else in, like my own type of definition…And then being able to click, oh, re-plain language" (P6, Organization A).

**Visual structure and formatting.** Practitioners emphasized visual hierarchy aids comprehension: "I think the big thing for me is I'm very visual. So…if the headings can be bolded already, because then now it's like we're doing another revision of like, where does it jump from one topic to the next?" (P6, Organization A).

## 4.2 RQ2: Do Automated Metrics Capture What Practitioners Prioritize?

*4.2.1 Organization-Produced Documents Evaluation.* Through our analysis of 33 documents from four organizations, we found that organizations improved metrics but rarely met commonly cited targets from published accessibility guidelines (Table 2). Only 10 of 33 documents (30.3%) met FKGL $\leq$ 8.0 (PLAIN general recommendation), only 1 of 33 documents (3%) met FKGL $\leq$ 6.0 (Organizational targets and Easy Read guidelines), and no documents met FKGL $\leq$ 5.0 (Organization B targets) or FKGL $\leq$ 3.0 (MTC recommendation).

This pattern holds regardless of which organization produced the documents. For example, Organization B achieved 32% sentence length reduction yet remained 6.1 grades above target because domain-specific legal terminology was necessary for accuracy. Despite these metric gaps, practitioners across all organizations consistently described their outputs as meeting organizational standards and receiving positive community feedback. During focus groups, practitioners emphasized community validation: "We get feedback from council members that the summary documents are helpful to them" (P7, Organization A); "We try to use resources that are created by community members for and with the community" (P1, Organization C).

*4.2.2 Beyond Metrics: Dimensions Practitioners Prioritize in Disability Advocacy.* Prior work in plain language summarization identified core dimensions essential for plain language work [27, 34]: informativeness, simplification, coherence, and faithfulness. Our studies confirm the importance of these and other dimensions, with additional considerations in the disability advocacy context for people with IDD.

| Category | Metric | Guideline Target[a] | Org A | Org B | ASAN | CDT |
|---|---|---|---|---|---|---|
| **Readability** | FKGL | 3.0-8.0 | 13.3 | 11.1 | **7.5** | 8.9 |
| | Gunning Fog Index | ≤8.0 | 15.5 | 13.3 | 9.4 | 10.9 |
| | SMOG Index | ≤8.0 | 14.4 | 12.7 | 10.2 | 11.2 |
| | Flesch Reading Ease | ≥60.0 | 37.4 | 53.2 | **66.5** | 54.9 |
| **Sentence** | Avg words/sentence | ≤15.0 | 21.4 | 21.3 | **14.1** | **13.3** |
| | % sentences 8-10 words | ≥30.0 | 6.9 | 9.1 | 18.8 | 18.3 |
| | % sentences >20 words | ≤25.0 | 34.8 | 40.6 | **14.7** | **18.3** |
| | Avg syllables/word | ≤1.5 | 1.75 | 1.56 | **1.49** | 1.64 |
| **Word Choice** | % common words | ≥92.0 | **97.3** | **97.8** | **96.8** | **93.6** |
| | % words >3 syllables | ≤10.0 | **7.8** | **4.2** | **5.6** | **6.0** |
| | Type-token ratio | ≤0.4 | 0.42 | 0.46 | **0.29** | **0.12** |
| **Style** | Passive voice % | <5.0 | **1.6** | **0.7** | **0.2** | **0.5** |
| | Passive voice count | <10 | **8** | **3** | **7** | 48 |
| | Negation per 100 words | <1.0 | **0.7** | **0.3** | **0.7** | **0.6** |
| | Contractions % | ≤0 | **0** | **0** | **0** | **0** |
| | All CAPS words % | ≤0 | 2 | **0** | 21 | 120 |
| | Acronym count | <5 | 28 | 38 | 35 | 60 |
| **Structure** | Avg words/paragraph | <100 | 624.6 | 152.7 | **18.2** | **23.7** |
| | Paragraphs >120 words | ≤0 | 1 | 1 | **0** | 2 |
| | Heading count | ≥1 | 0 | 0 | 0 | 0 |
| | Headings >8 words | ≤0 | **0** | **0** | **0** | **0** |

Table 2. Key readability and structural metrics averaged across organization-produced plain language documents compared to targets from published accessibility guidelines[a]. Values in **bold** meet guideline targets. Organizations achieved substantial improvements from original texts (overall mean FKGL: 14.7 to 10.2), yet most documents didn't meet published targets while practitioners described them as successfully meeting community needs (metrics by organization in Appendix Tables 4–7). [a]Targets synthesized from Plain Language federal guidelines (PLAIN), Easy Read standards, and MTC framework.

- **Definitional support:** Practitioners prioritize defining terms that are unfamiliar to IDD communities. For example, Organization A materials included definitions for council-specific terms ("state plan," "federal appropriation," "HCBS waiver"); Organization B explained legal concepts specific to disability rights advocacy. Knowing which terms to define requires accumulated community experience rather than linguistic features such as word frequency or syllable count. In the walkthrough, practitioners identified 54 technical terms across 12 documents requiring explanation, yet none appeared with definitions in AI-generated simplified materials.
- **Accuracy and meaning preservation:** In disability advocacy contexts, accuracy concerns are amplified because simplified materials directly affect people's understanding of their legal rights. As an example, one practitioner noted that AI-simplified text changed "unlawful killing when perpetrated by any act imminently dangerous to another and evincing a depraved mind" to "covers serious crimes like kidnapping, robbery, and more", noting: "The points aren't the same…what second-degree murder is, is if the offender is negating" (P4, Organization B).
- **Information completeness and contextual framing:** In high-stakes advocacy, completeness usually outweighs brevity; practitioners rejected summarizations that replaced enumerated items or stakes with placeholders, e.g., Organization B required complete enumeration: "We would want it to be complete. So we

would want to actually list out the ‘and more’. And then define the pieces in that list” (P3, Organization B). Metrics treat word reduction as improvement, but practitioners must judge what context can be omitted for disability rights advocacy.

- **Structural appropriateness and community-validated practices:** Organizations follow community-informed formatting conventions (e.g., flat lists, specific visual structures) that affect accessibility but are invisible to automated metrics. Two documents with identical FKGL scores may differ dramatically in accessibility based on structural decisions like added headings or lists.

As confirmed during practitioner walkthroughs, concerns clustered into four themes: missing definitions (92% of documents), accuracy issues (42%), information erasure (25%), and structure misalignment (33%).

## 4.3 RQ3: Design requirements for AI-supported text simplification

We articulate design requirements for tools that support rather than replace practitioner expertise. Prior work has established principles for human-AI collaboration, including transparency, appropriate reliance, and verification support [10, 41]. Our findings build on these principles by showing how they surface in disability advocacy workflows.

*4.3.1 Centering Human Expertise and Oversight.* Grounded in our RQ1 finding that practitioners consistently position AI outputs as provisional starting points requiring complete human verification, never as autonomous producers of publication-ready content (Sections 4.1.1, 4.1.3), systems designed for professional accessibility workflows should explicitly communicate uncertainty and position human reviewers as essential decision-makers.

This creates specific architectural requirements. Outputs should be marked as provisional, with interfaces communicating that human review is essential. Systems should support collaborative review where multiple stakeholders validate different dimensions, show clear provenance distinguishing tool-generated content from human revision, preserve human control where the tool suggests but humans decide, and make it straightforward to reject outputs.

*4.3.2 Reading Level Feedback with Appropriate Caveats.* Practitioners need clear feedback about reading level, but our RQ2 document analysis (Section 4.2.1) demonstrate that metric achievement does not guarantee accessibility. If using readability metrics, the system should present them with appropriate caveats. For example, systems should display FKGL alongside content-level flags (e.g., "15 technical terms lack definitions" or "3 enumerated lists can be summarized"), and present both original and simplified metrics to indicate degree of change. Finally, systems should explicitly distinguish between achieving a target metric score and being practitioner-validated as accessible.

*4.3.3 Proactive Context Preservation and Definition Support.* Prior work on text simplification emphasizes the importance of elaboration over deletion for comprehension [47]. Systems should proactively flag terms that likely need explanation based on syllable count, domain context, and common word frequency. Rather than automatically generating definitions (which may introduce error), systems could surface terms for practitioner attention and enable reuse of organization-specific definitions. Systems should flag when summarization removes enumerated items, requiring practitioners to explicitly approve such condensation rather than silently replacing "A, B, C, D, E" with "A, B, and others." Systems should also preserve contextual framing when present in source materials (why something matters, implications for affected communities, connections to broader issues) rather than treating such content as extraneous.

*4.3.4 Structural Control and Organization-Specific Standards.* Our RQ1 findings (Section 4.1.2) document that organizations develop formatting and structural conventions based on accumulated experience serving their communities. For instance, Organization B required flat bullet lists for legislative translations, avoiding headings

because their audience benefits from linear structure. Systems should enable organizations to specify structural preferences as requirements rather than suggestions. When they cannot be met, systems should explain why rather than silently violating preferences. Storing organization-specific standards as defaults would reduce friction in repeated use. Additionally, systems should support section-level editing where practitioners can accept, reject or modify specific passages rather than all-or-nothing decisions.

*4.3.5 Transparency and Verification Infrastructure.* Verification burden documented across Sections 4.1.4, including Organization B's discontinuation after determining verification costs exceeded benefits, confirms that verifying AI outputs in these contexts requires substantial cognitive effort [41]. Systems should aim to reduce verification burden; as one practitioner explained: "I can fix the jargon, but then I'm essentially doing the whole job anyway. The time I save on sentence structure, I lose on verification and definition work" (P7, Organization A). This could be achieved through side-by-side comparison interfaces showing source and simplified text in parallel; character-level diff visualizations showing exactly what changed; and support for collaborative review where different team members assess different aspects. Until verification burden becomes manageable, efficiency promises remain unrealized for practitioners working in contexts where errors have consequences for disabled people's access to consequential information.

## 5 Discussion

Our findings reveal fundamental tensions between what is measurable and what matters for accessibility in disability advocacy. We documented practitioner workflows (RQ1), found systematic misalignment between metrics and practitioner judgments (RQ2), and articulated design requirements that prioritize accountability over automation (RQ3). We now reflect on three implications: the limits of standardization when accessibility requires community-specific knowledge, the challenge of supporting self-determination while acknowledging tool risks, and the sustainability question when verification burden approaches creation effort.

### 5.1 When Accessibility Resists Standardization

Prior work established that automated metrics provide limited signal about simplification quality [27, 34]. Our findings confirm these limitations in disability advocacy contexts while revealing a deeper challenge: the dimensions practitioners prioritize may resist any standardized quantification.

Only 3% of organization-produced documents met FKGL $\leq$ 6.0, yet practitioners described these materials as successfully serving community needs. This disconnect occurred because practitioners prioritized other dimensions, such as keeping technical terms necessary for understanding but defining them. Other aspects of accessibility such as accuracy preservation or information completeness similarly require expertise and experience-based judgment. This challenges an implicit assumption that accessibility can be automated through optimizing simple readability measures.

In other words, metrics provide useful but insufficient signals. The gold standard in disability advocacy is community feedback from people with IDD, not metric achievement. Participatory approaches [14] where IDD community members shape evaluation criteria offer promise, though whether and how community-specific knowledge can be incorporated into generalizable systems remains an open question.

Our findings also speak to questions of metric validity in AI evaluation more broadly. The NIST AI Risk Management Framework [49] frames valid AI as producing outputs accurate and fit-for-purpose for the intended context. By that standard, readability metrics fail in disability advocacy contexts; communities describe as successful the very documents that exceed published FKGL targets, while documents meeting those targets routinely omit definitional support, accuracy preservation, and structural conventions practitioners consider essential. Future evaluation frameworks should treat automated metrics as one input among other evaluative components. Automated metrics provide efficient signals of surface-level linguistic differences, while practitioner

expertise captures domain knowledge and community-validated conventions that are more difficult to quantify. Finally, direct community feedback from IDD users is necessary to provide the ground-truth signal of whether materials actually serve their intended purpose. None of these is sufficient by itself. And in particular, automated metrics, when used alone, can create an appearance of rigor while weakening accountability to the end users these simplified documents are actually serving.

### 5.2 Self-Determination, Risk, and Power

Self-advocates increasingly use AI tools to access information independently, yet current systems may not serve their needs in high-stakes contexts. This creates a dilemma: advocating against AI use would be paternalistic, while promoting uncritical adoption ignores legitimate concerns about accuracy and completeness. This reflects broader tensions about who defines accessibility. Historically, tools have been developed *for* disabled people rather than *with* them [30]. Current AI systems follow this pattern: trained on datasets that may not represent IDD perspectives, optimized for metrics that may not capture what matters to IDD readers, and deployed with claims about accessibility that practitioners working with IDD communities question.

Practitioners emphasized community accountability: "We try to use resources that are created by community members for and with the community." (P1, Organization C) This principle should extend to AI development. Rather than making tools "safe enough" through paternalistic restrictions, we should center IDD community members and accountable practitioners in shaping what tools do and how they communicate limitations. The path forward requires honest disclosure about tool limitations for legal, medical, and policy content; risk-aware design helping users understand when verification matters; and participatory development treating self-advocates as experts in their information needs.

### 5.3 Verification Labor and Sustainability

Practitioners reported verification work that sometimes matched original creation; for example, Organization B discontinued AI use after determining verification burden exceeded benefits. This labor is often invisible in automation narratives [26, 41]. High-stakes contexts amplify verification demands. When errors affect disabled people's access to legal rights or policy understanding, practitioners cannot risk making mistakes. They feel responsible to communities while lacking control over AI behavior: "My name goes on this. The disability community trusts me. If the AI makes a mistake and I don't catch it, that's on me." (P6, Organization A) This accountability without control [33] requires examination as AI tools enter workflows affecting consequential information.

Tool adoption should be framed not as binary but in terms of sustainability, given actual workflow demands. The question is not whether AI can generate simplified text but whether AI-assisted workflows can be sustained by practitioners under resource constraints and accountability pressures.

## 6 Limitations and Future Work

Our study engaged a limited set of three U.S. advocacy organizations focused on policy and legislative materials. Plain language work in healthcare, education, employment, or international contexts may involve different priorities and constraints. Our document corpus also focused on policy documents; other high-stakes domains (medical consent, financial information) warrant further investigation.

We centered practitioners who produce materials rather than people with IDD who use them. While practitioners' expertise reflects accumulated community feedback and accountability relationships, we cannot assume their priorities predict comprehension outcomes for IDD readers. Future work should examine whether the dimensions we identified (definitional adequacy, accuracy preservation, completeness, and advocacy-aligned framing) actually predict better outcomes through studies with IDD participants.

For our metric analysis, we analyzed 28 automated and linguistic metrics, but other model-based measures or domain-specific metrics designed for IDD accessibility might capture dimensions we did not assess. Future work can developing and validate such metrics through participatory processes involving IDD users and practitioners, strengthening the field's ability to assess accessibility.

## 7 Conclusion

We contribute three insights. First, plain language work is expertise requiring policy knowledge, community accountability, and systematic validation, not simply linguistic transformation. Practitioners must preserve accuracy, provide definitional support, maintain completeness, and preserve context, dimensions shaped by ethical obligations to disability communities. Second, automated and linguistic metrics inadequately capture these dimensions. Only 3% of documents met FKGL organizational targets, yet practitioners described outputs as successful based on community feedback. This has implications for AI systems: optimization on metrics that miss essential dimensions may produce outputs appearing accessible while failing actual user needs. Third, we articulate design principles emphasizing transparency, supporting rather than replacing expertise, and reducing verification burden through participatory development centering IDD users and practitioners.

Our study demonstrates that plain language work may be supported by AI but cannot be replaced by it. Meaningful progress requires evaluation paradigms centering practitioner expertise and community needs rather than automated metrics alone. As institutions increasingly deploy AI for accessibility-critical tasks, disability justice requires centering disabled people's autonomy and expertise rather than optimizing algorithms on their behalf.

## Acknowledgments

The contents of this paper were developed under a grant from the National Institute on Disability, Independent Living, and Rehabilitation Research (NIDILRR grant number 90REGE0026) funding the Center for Research and Education on Accessible Technology and Experiences (CREATE). NIDILRR is a Center within the Administration for Community Living (ACL), Department of Health and Human Services (HHS). The contents of this paper do not necessarily represent the policy of NIDILRR, ACL, HHS, and you should not assume endorsement by the Federal Government. This work was also supported by gift funds from Google and the Allen Institute for AI.

## Generative AI Disclosure Statement

Our work evaluates generative AI systems as the object of study. We used GPT-4o (OpenAI, accessed August-September 2025) to generate AI-simplified versions of 12 documents analyzed in Phase 2 walkthroughs (prompt provided in Appendix D).

Claude 3.5 Sonnet model by Anthropic was used in a limited capacity during manuscript preparation to suggest alternative phrasings for clarity and conciseness. All content, arguments, interpretations, and findings are the work of the human authors, who reviewed, revised, and take full responsibility for all text in the final manuscript.

Generative AI was not used in research design, data collection, qualitative analysis, or interpretation of findings. All thematic analysis, coding, and synthesis of results were conducted by human researchers without AI assistance.

## References

[1] [n. d.]. https://www.section508.gov/
[2] [n. d.]. https://hemingwayapp.com/
[3] [n. d.]. https://autisticadvocacy.org/resources/
[4] [n. d.]. https://goblin.tools/Formalizer
[5] 2023. https://ncua.gov/about/open-government/plain-writing-act-2010
[6] 2025. https://cdt.org/plain-language-resource-hub/
[7] 2025. https://cidi.gatech.edu/research/MTCAI

[8] Sweta Agrawal and Marine Carpuat. 2024. Do Text Simplification Systems Preserve Meaning? A Human Evaluation via Reading Comprehension. *Transactions of the Association for Computational Linguistics* 12 (2024), 432–448. doi:10.1162/tacl_a_00653

[9] Suha S. Al-Thanyyan and Aqil M. Azmi. 2021. Automated Text Simplification: A Survey. *ACM Comput. Surv.* 54, 2, Article 43 (March 2021), 36 pages. doi:10.1145/3442695

[10] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3290605.3300233

[11] Autistic Self Advocacy Network. 2025. ASAN Says No: Generative AI in Plain Language. https://autisticadvocacy.org/2025/07/asan-says-no-generative-ai-in-plain-language/. Published July 29, 2025.

[12] Nadine Beks van Raaij, Daan Kolkman, and Ksenia Podoynitsyna. 2024. Clearer Governmental Communication: Text Simplification with ChatGPT Evaluated by Quantitative and Qualitative Research. In *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024*, Giorgio Maria Di Nunzio, Federica Vezzani, Liana Ermakova, Hosein Azarbonyad, and Jaap Kamps (Eds.). ELRA and ICCL, Torino, Italia, 152–178. https://aclanthology.org/2024.determit-1.15/

[13] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2023. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1493–1504. doi:10.1145/3593013.3594095

[14] Charlotte Bird, Eddie Ungless, and Atoosa Kasirzadeh. 2023. Typology of Risks of Generative Text-to-Image Models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (Montréal, QC, Canada) *(AIES '23)*. Association for Computing Machinery, New York, NY, USA, 396–410. doi:10.1145/3600211.3604722

[15] Cornelia Boldyreff, Elizabeth Burd, Joanna Donkin, and Sarah Marshall. 2001. The Case for the Use of Plain English to Increase Web Accessibility. In *3rd International Workshop on Web Site Evolution (WSE 2001) - Access for All, 10 November 2001, Florence, Italy*. IEEE Computer Society, 42–48. doi:10.1109/WSE.2001.988784

[16] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3 (01 2006), 77–101. doi:10.1191/1478088706qp063oa

[17] Liam Cripwell, Joël Legrand, and Claire Gardent. 2023. Simplicity Level Estimate (SLE): A Learned Reference-Less Metric for Sentence Simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 12053–12059. doi:10.18653/v1/2023.emnlp-main.739

[18] Diffit. [n. d.]. Diffit. https://web.diffit.me/

[19] Inmaculada Fajardo, Gema Tavares, Vicenta Clemente, and Antonio Ferrer. 2013. Towards text simplification for poor readers with intellectual disability: When do connectives enhance text cohesion? *Research in developmental disabilities* 34 (04 2013), 1267–79. doi:10.1016/j.ridd.2013.01.006

[20] Center for Inclusive Design and Innovation (CIDI). 2021. Guidelines for Minimizing the Complexity of Text. https://cidi.gatech.edu/sites/default/files/2021-02/Minimized%20Text%20Complexity%20Guidelines%20%5Bversion%202.03.2021%5D.pdf Accessed: 2026-04-21.

[21] Vinitha Gadiraju, Shaun Kane, Sunipa Dev, Alex Taylor, Ding Wang, Remi Denton, and Robin Brewer. 2023. "I wouldn't say offensive but...": Disability-Centered Perspectives on Large Language Models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 205–216. doi:10.1145/3593013.3593989

[22] Yingqiang Gao, Kaede Johnson, David Froehlich, Luisa Carrer, and Sarah Ebling. 2025. Evaluating the Effectiveness of Direct Preference Optimization for Personalizing German Automatic Text Simplifications for Persons with Intellectual Disabilities. arXiv:2507.01479 [cs.CL] https://arxiv.org/abs/2507.01479

[23] Kate Glazko, JunHyeok Cha, Aaleyah Lewis, Ben Kosa, Brianna L Wimer, Andrew Zheng, Yiwei Zheng, and Jennifer Mankoff. 2025. Autoethnographic Insights from Neurodivergent GAI "Power Users". In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 274, 19 pages. doi:10.1145/3706598.3713670

[24] Kate Glazko, Yusuf Mohammed, Ben Kosa, Venkatesh Potluri, and Jennifer Mankoff. 2024. Identifying and Improving Disability Bias in GPT-Based Resume Screening. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 687–700. doi:10.1145/3630106.3658933

[25] Kate S Glazko, Momona Yamagami, Aashaka Desai, Kelly Avery Mack, Venkatesh Potluri, Xuhai Xu, and Jennifer Mankoff. 2023. An Autoethnographic Case Study of Generative Artificial Intelligence's Utility for Accessibility. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility* (New York, NY, USA) *(ASSETS '23)*. Association for Computing Machinery, New York, NY, USA, Article 99, 8 pages. doi:10.1145/3597638.3614548

[26] Mary L. Gray and Siddharth Suri. 2019. Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass. https://api.semanticscholar.org/CorpusID:260447696

[27] Yue Guo, Tal August, Gondy Leroy, Trevor Cohen, and Lucy Lu Wang. 2024. APPLS: Evaluating Evaluation Metrics for Plain Language Summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 9194–9211. doi:10.18653/v1/2024.emnlp-main.519

[28] Ziyang Guo, Yifan Wu, Jason D. Hartline, and Jessica R. Hullman. 2024. A Decision Theoretic Framework for Measuring AI Reliance. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (2024). https://api.semanticscholar.org/CorpusID:267311650

[29] Oliver L. Haimson, Samuel Reiji Mayworm, Alexis Shore Ingber, and Nazanin Andalibi. 2025. AI Attitudes Among Marginalized Populations in the U.S.: Nonbinary, Transgender, and Disabled Individuals Report More Negative AI Attitudes. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 1224–1237. doi:10.1145/3715275.3732081

[30] Aimi Hamraie. 2017. *Building Access: Universal Design and the Politics of Disability*. University of Minnesota Press, Minneapolis.

[31] Inclusion Europe. 2009. Information for all: European standards for making information easy to read and understand. https://www.inclusion-europe.eu/wp-content/uploads/2017/06/EN_Information_for_all.pdf

[32] International Organization for Standardization. 2023. ISO 24495-1:2023 Plain language — Part 1: Governing principles and guidelines. https://www.iso.org/standard/78907.html

[33] Gabbrielle M. Johnson. 2020. Algorithmic bias: on the implicit biases of social technology. *Synthese* 198 (2020), 9941 – 9961. https://api.semanticscholar.org/CorpusID:219935412

[34] Sebastian Joseph, Lily Chen, Jan Trienes, Hannah Göke, Monika Coers, Wei Xu, Byron Wallace, and Junyi Jessy Li. 2024. FactPICO: Factuality Evaluation for Plain Language Summarization of Medical Evidence. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 8437–8464. doi:10.18653/v1/2024.acl-long.459

[35] Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. BLESS: Benchmarking Large Language Models on Sentence Simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 13291–13309. doi:10.18653/v1/2023.emnlp-main.821

[36] Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. Controllable Text Simplification with Explicit Paraphrasing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, Online, 3536–3553. doi:10.18653/v1/2021.naacl-main.277

[37] Maggie Nygren, Robyn Linscott, Mike Nagel, Michael Atkins, Julie Ward, and Jenny Alexander. 2024. Developing and Evaluating the Fidelity and Understandability of Plain Language Summaries of Position Statements. *Intellectual and Developmental Disabilities* 62 (01 2024), 74–81. doi:10.1352/1934-9556-62.1.74

[38] U.S. Department of Justice Civil Rights Division. 2022. ADA Requirements - Effective Communication. https://www.ada.gov/resources/effective-communication/

[39] Gustavo Paetzold and Lucia Specia. 2016. Benchmarking Lexical Simplification Systems. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Portorož, Slovenia, 3074–3080. https://aclanthology.org/L16-1491/

[40] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (Philadelphia, Pennsylvania) *(ACL '02)*. Association for Computational Linguistics, USA, 311–318. doi:10.3115/1073083.1073135

[41] Samir Passi, Shipi Dhanorkar, and Mihaela Vorvoreanu. 2024. *Appropriate reliance on Generative AI: Research synthesis*. Technical Report MSR-TR-2024-7. Microsoft. https://www.microsoft.com/en-us/research/publication/appropriate-reliance-on-generative-ai-research-synthesis/

[42] Mahika Phutane, Ananya Seelam, and Aditya Vashistha. 2025. "Cold, Calculated, and Condescending": How AI Identifies and Explains Ableism Compared to Disabled People. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 1927–1941. doi:10.1145/3715275.3732128

[43] Publications Office of the European Union. 2023. Easy to Read Guidelines. PDF. https://op.europa.eu/documents/d/accessibility/easy-to-read-guidelines Accessible Publishing (Accessibility). Creation date in PDF metadata: 2023-10-23.

[44] Readable. 2018. Readable. https://readable.com/

[45] Rewordify. 2019. Rewordify.com | Understand what you read. https://rewordify.com/

[46] Adeline Rosenberg, Joanne Walker, Sarah Griffiths, and Rachel Jenkins. 2023. Plain language summaries: Enabling increased diversity, equity, inclusion and accessibility in scholarly publishing. *Learned Publishing* 36 (01 2023). doi:10.1002/leap.1524

[47] Neha Srikanth and Junyi Jessy Li. 2021. Elaborative Simplification: Content Addition and Explanation Generation in Text Simplification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 5123–5137. doi:10.18653/v1/2021.findings-acl.455

[48] Sanja Štajner, Ruslan Mitkov, and Horacio Saggion. 2014. One Step Closer to Automatic Evaluation of Text Simplification Systems. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, Sandra Williams, Advaith Siddharthan, and Ani Nenkova (Eds.). Association for Computational Linguistics, Gothenburg, Sweden, 1–10. doi:10.3115/v1/W14-1201

[49] Elham Tabassi. 2023. *Artificial Intelligence Risk Management Framework (AI RMF 1.0).* NIST AI 100-1. National Institute of Standards and Technology, Gaithersburg, MD. doi:10.6028/NIST.AI.100-1

[50] Melody M. Terras, Dominic Jarrett, and Sharon A. McGregor. 2021. The Importance of Accessible Information in Promoting the Inclusion of People with an Intellectual Disability. *Disabilities* 1, 3 (2021), 132–150. doi:10.3390/disabilities1030011

[51] Hema Thakur. 2025. Guest Post - The Accessibility Illusion: When AI Simplification Fails the Users With Cognitive Disabilities - The Scholarly Kitchen. https://scholarlykitchen.sspnet.org/2025/07/22/guest-post-the-accessibility-illusion-when-ai-simplification-fails-the-users-with-cognitive-disabilities/

[52] TurboScribe. [n. d.]. TurboScribe: Transcribe Audio and Video to Text or Subtitles in Seconds. https://turboscribe.ai/

[53] UK Government. 2010. Making written information easier to understand for people with learning disabilities: Guidance for people who commission or produce Easy Read information. https://www.gov.uk/government/publications/making-written-information-easier-to-understand-for-people-with-learning-disabilities-guidance-for-people-who-commission-or-produce-easy-read-information-revised-edition-2010

[54] U.S. Congress. 2010. Plain Writing Act of 2010. https://www.govinfo.gov/content/pkg/PLAW-111publ274/pdf/PLAW-111publ274.pdf Public Law 111–274, 124 Stat. 2861.

[55] U.S. General Services Administration. 2010. Plain Language Guide Series. https://plainlanguage.gov/

[56] Daniel Vlantis, Iva Gornishka, and Shuai Wang. 2024. Benchmarking the Simplification of Dutch Municipal Text. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 2217–2226. https://aclanthology.org/2024.lrec-main.199/

[57] W3C. 2019. Making Content Usable for People with Cognitive and Learning Disabilities. https://www.w3.org/TR/coga-usable/

[58] Ruben Weijers, Simone Ooms, Kellin Pelrine, and Hanna Hauptmann. 2026. *Towards Accessible Information Retrieval for Children With a Mild Intellectual Disability.* 127–153. doi:10.1007/978-3-032-12717-4_9

[59] Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in Current Text Simplification Research: New Data Can Help. *Transactions of the Association for Computational Linguistics* 3 (2015), 283–297. doi:10.1162/tacl_a_00139

[60] Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing Statistical Machine Translation for Text Simplification. *Transactions of the Association for Computational Linguistics* 4 (2016), 401–415. doi:10.1162/tacl_a_00107

[61] Victoria Yaneva, Irina Temnikova, and Ruslan Mitkov. 2016. Evaluating the Readability of Text Simplification Output for Readers with Cognitive Disabilities. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Portorož, Slovenia, 293–299. https://aclanthology.org/L16-1045/

# A Semi-Structured Focus Group Protocol

This protocol guided focus group discussions with disability advocacy organizations (August-September 2025). Sessions lasted 60-90 minutes and covered six domains: current practices, guidelines and standards, values and principles, AI usage patterns, validation processes, and challenges.

## A.1 Welcome and Access Norms

**Facilitator script:**

- Welcome and thank you for participating
- You may skip any question or leave at any time
- If you have access needs (visual, auditory, cognitive, or otherwise), please let us know and we will adapt (slower pace, captions, larger font)
- This session will be recorded with your permission for transcription purposes
- We will discuss your organization's plain language work and experiences with AI tools

### A.2 Opening Questions

(1) Can you tell me about your role and your organization's plain language work?
(2) How often does your organization produce plain language materials? What types of documents?
(3) Who are the primary audiences for your plain language materials?

### A.3 Current Plain Language Practices

(1) Walk me through your typical plain language workflow. How do you approach simplifying a document?
(2) What guidelines or resources do you use? (Probe: PLAIN, ASAN, Easy Read, CDC, others?)
(3) What reading level do you typically target? How was this determined?
(4) How do you validate that simplified text meets your standards?
(5) Who reviews plain language materials before publication? What does that process look like?
(6) Do you collect feedback from your audience on plain language materials? If so, how?

### A.4 Values and Quality Criteria

(1) What makes plain language "good" in your context? What are you optimizing for?
(2) What are your biggest concerns when simplifying high-stakes documents (policy, legislation, healthcare)?
(3) Can you describe a time when simplification went wrong? What happened?
(4) How do you balance simplicity with accuracy? Completeness with brevity?
(5) What does accountability to your disability community mean in this work?

### A.5 Generative AI Usage

(1) Has your organization used AI tools (like ChatGPT) for plain language work?
(2) If yes:
    - Which tools have you tried?
    - How do you use them? Walk me through a typical workflow.
    - How do you prompt them? Can you share an example?
    - What has worked well? What hasn't?
    - How do you verify AI outputs?
    - Have you stopped using any tools? Why?
(3) If no:
    - Are you aware of AI text simplification tools?
    - What's your impression of them?
    - Have you considered using them? What factors influence that decision?
(4) What concerns do you have about using AI for plain language work?

### A.6 Validation and Quality Control

(1) How do you know when a simplified text is "ready" to publish?
(2) What kinds of errors or problems do you watch for?
(3) If you use AI, how do you check its outputs? What specifically do you verify?
(4) Do you have organizational quality standards or checklists?
(5) How much time does verification typically take compared to manual creation?

### A.7 Challenges and Needs

(1) What are the biggest challenges in your plain language work?
(2) What takes the most time or effort?

(3) If you could change one thing about your current workflow or tools, what would it be?
(4) What features would make plain language work easier or more efficient?
(5) Are there specific types of content that are particularly difficult to simplify?

### A.8 Closing

(1) Is there anything we haven't discussed that's important for understanding your plain language work?
(2) Would you be willing to share example documents (original and simplified versions) for our analysis?
(3) May we follow up with you individually for more detailed walkthroughs?

## B Detailed Thematic Analysis Process

We conducted reflexive thematic analysis following Braun and Clarke's [16] six-phase approach:

**Phase 1: Familiarization.** Both authors read complete transcripts multiple times, taking initial notes on salient themes.

**Phase 2: Initial coding.** Two researchers independently conducted line-by-line coding of all transcripts. We developed codes iteratively, beginning with descriptive codes close to participant language (e.g., "accuracy as foundation," "community accountability," "manual work essential," "workflow burden"). We met bi-weekly to compare codes, discuss disagreements, and refine definitions. Initial codebook included 45 codes organized into 9 categories: Workflow, Values and Principles, Challenges, Quality Criteria, Audience and Purpose, Guidelines, Desired Features, Document Types, and Organizational Context.

When codes diverged, we returned to transcripts together, examined context, and either created distinct codes for genuinely different phenomena or merged codes referring to the same underlying concept. For ambiguous segments, we coded conservatively (selecting the code with strongest evidence) and flagged items for team discussion.

**Phase 3: Theme generation.** We collaboratively grouped related codes into candidate themes through analysis sessions. For example, codes "accuracy as foundation," "integrity of source text," and "no hallucinations" combined into the theme "Accuracy as Non-Negotiable."

**Phase 4: Theme review.** Two researchers checked themes against coded data to ensure internal coherence (codes within a theme fit meaningfully) and external distinctiveness (clear boundaries between themes). We refined theme names and definitions iteratively.

**Phase 5: Theme refinement.** The lead author developed detailed analytical narratives for each theme, identifying sub-themes and documenting variations across organizations. The full team reviewed narratives for accuracy and completeness. Final themes: (1) The Irreplaceable Human Element, (2) Accuracy as Non-Negotiable, (3) Context Preservation as Fundamental Challenge, (4) Transparency and Traceability, (5) Workflow Burden, (6) Community-Centered Values, (7) Definitional Support and Terminology Management, (8) Structural Appropriateness, (9) Validation Through Community Feedback, (10) Organizational Mission Shapes Practice.

**Phase 6: Reporting.** We selected exemplary quotes for each theme ensuring representation across organizations and participant roles. We triangulated themes with quantitative findings from document analysis to provide convergent evidence.

Throughout analysis, we practiced reflexivity by discussing how our assumptions shaped interpretation and memoing about findings that challenged our expectations.

## C Individual Walkthrough Protocol

This protocol guided think-aloud walkthroughs with practitioners who agreed to participate in group sessions but shared screen for walkthroughs (August-September 2025). Sessions lasted 30-45 minutes.

### C.1 Introduction

**Facilitator script:** Thank you for agreeing to do this walkthrough. Today I'll ask you to review AI-generated plain language versions of documents your organization has previously simplified. I'm interested in your reactions, what you notice, and how you evaluate quality. Please think aloud as you review, share whatever comes to mind. There are no right or wrong answers.

### C.2 Document Selection

Can you select any document that your organization has previously simplified? Ideally, a document you personally worked on or are familiar with. This can be policy materials, legislative summaries, meeting documents, or other content you've translated to plain language.

### C.3 AI Output Review

**For each document:**

(1) I'm going to show you an AI-generated plain language version of [document name]. Please review it and share your thoughts as you go.
(2) *[Facilitator presents AI-simplified version. Remain silent unless participant stops talking.]*
(3) *[If participant falls silent, use minimal prompts:]*
    - What are you thinking about right now?
    - Can you tell me more about that?
    - What are you noticing there?

### C.4 Follow-up Questions (After Review)

(1) What's your overall impression of this output?
(2) How does it compare to your organization's version?
(3) Did you notice any errors or problems? What kinds?
(4) What would you need to change before publishing this?
(5) How much additional work would be required?
(6) Is there anything the AI did well?
(7) Is there anything particularly problematic?

### C.5 Comparative Analysis (If Time Permits)

Would you be comfortable walking me through specific differences between the AI version and your organization's version? What changed and why does it matter?

## D AI Text Simplification Prompt

We used GPT-4o (OpenAI) with the following prompt to generate AI-simplified versions of all documents analyzed in this study. The prompt was iteratively refined based on disability-led guidelines including Plain Language Action and Information Network (PLAIN), Autistic Self Advocacy Network (ASAN), Easy Read standards, CDC Clear Communication Index, and NIH Plain Language Guidelines.

### D.1 System Prompt

```
You are an expert plain-language editor trained in plain
language and Easy Read standards, including guidance from
the CDC, NIH, Plain Language Action and Information Network
(PLAIN), Gov.uk, and Inclusion Europe.
```

```
Rewrite the given text in clear, accessible GitHub-Flavored
Markdown so it is easy to read and understand without losing
meaning or accuracy.

GLOBAL CONSTRAINTS (MUST FOLLOW)
* Preserve facts, intent, and sequence exactly.
* Do not add, infer, summarize, or hallucinate information.
* Do not remove legally, medically, or procedurally important
  content.
* Use inclusive, gender-neutral language where appropriate.
* Use consistent terms for the same concept throughout.
* Avoid double negatives.
* Return only the rewritten Markdown. No explanations or
  commentary.

READABILITY TARGETS
When possible, revise the text to meet these document-level goals:
* Reading level: Aim for U.S. grade level 6 or below
* Sentence length: Prefer short sentences, typically 15 words
  or fewer

If exact targets cannot be met without changing meaning,
prioritize accuracy over metrics.

MARKDOWN & STRUCTURE RULES
* Start headings at Level 1 and do not skip heading levels.
* Use clear, descriptive headings.
* Leave blank lines between headings and sections.
* Use lists for steps, requirements, or key points.
* Keep paragraphs short (generally no more than 3-4 sentences).
* Present one idea per paragraph.

VOCABULARY & TONE
* Prefer common, everyday words over technical or abstract terms.
* Replace jargon when possible.
* If a complex or technical term is required, define it in
  parentheses on first use.
* Expand acronyms on first use.
* Remove idioms, metaphors, and figurative language.
* Maintain a clear, respectful, neutral tone.

SENTENCE STRUCTURE
* Prefer active voice.
* Avoid nested clauses and long compound sentences.
* Avoid unclear pronouns (e.g., "this," "that," "it" without
```

```
  a clear noun).
* Avoid unnecessary negation (e.g., "not unless,"
  "do not fail to").

CLARITY & FLOW
* Organize information in a logical, predictable order.
* Use headings and lists to chunk information.
* Use second person ("you") for instructions when appropriate.
* Ensure transitions between sections are clear.

FINAL VERIFICATION CHECKLIST
Before returning the output, ensure that:
* Meaning, facts, and order are unchanged.
* Language is simpler but not less precise.
* The text would be understandable to a general audience with
  limited background knowledge.
* Formatting follows GitHub-Flavored Markdown.
* Reading level is approximately grade 6 or below, where
feasible.

OUTPUT REQUIREMENT
Return only the rewritten Markdown content. No explanations,
notes, or analysis.
```

### D.2 User Message Format

For each document, we provided the original text with the following user message:

```
Please simplify the following text according to the plain language
guidelines. Target a 4th-6th grade reading level while preserving
all factual information and legal precision.

[ORIGINAL TEXT INSERTED HERE]
```

## E Complete Quantitative Results

### E.1 Metric Definitions and Target Value Determination

We analyzed documents using 28 metrics across six categories. Target values were established by synthesizing organizational practices from focus groups, published accessibility standards (PLAIN, Easy Read, WCAG COGA), and empirical research on cognitive accessibility.

Other metrics that don't have a defined target include average conjunctions per sentence, prepositions per sentence, lexical cohesion score, and second-person count.

**Primary sources consulted:**

(1) **Plain Language Action and Information Network (PLAIN)**
   - Federal Plain Language Guidelines (2011)
   - PlainLanguage.gov resources
   - Target: FKGL $\leq$ 8.0 for public documents

(2) **Easy Read Standards**

- Inclusion Europe: Information for All guidelines
- UK Government Easy Read standard
- EU Publications Office: Easy to Read Guidelines (2023)
- Targets: 15 words per sentence maximum, define complex terms, active voice

(3) **Disability-Led Guidelines**
- Autistic Self Advocacy Network (ASAN) Easy Read resources
- ABCs of Plain Language (AUCD)
- Minimum Text Complexity Guidelines (Georgia Tech CIDI)
- Emphasize: definitions, completeness, structural simplicity

(4) **Accessibility Standards**
- WCAG 2.1 Cognitive and Learning Disabilities Accessibility (COGA)
- Section 508 plain language requirements
- ADA effective communication standards

(5) **Organizational Practices**
- Organization A: Target FKGL $\leq$6.0
- Organization B: Target FKGL $\leq$5.0
- Organization C: Flexible targets varying by audience

## E.2 Detailed Metric Comparison by Organization

Tables 4–7 provide complete metric breakdowns for each organization's documents, including all 28 linguistic and readability measures analyzed.

Received 11 September 2025

| Category | Metric | Target | Source/Justification |
|---|---|---|---|
| **READABILITY** | Flesch-Kincaid Grade Level | 3.0-8.0 | Minimal Text Complexity Guidelines; Easy Read recommendations, PLAIN guidelines, organizational targets (4th-6th grade) |
| | Gunning Fog Index | ≤ 8.0 | Middle school reading level; PLAIN guidance for general public; allows technical vocabulary when necessary |
| | SMOG Index | ≤ 8.0 | Similar to Gunning Fog; validated for health materials; conservative comprehension difficulty estimate |
| | Flesch Reading Ease | ≥ 60.0 | "Plain English" threshold; PLAIN guidelines; corresponds to 8th-9th grade |
| **SENTENCE STRUCTURE** | Average words per sentence | ≤ 15.0 | Minimal Text Complexity Framework "no more than 15 words" specification; Federal Plain Language guidance; WCAG COGA |
| | % sentences 8-10 words | ≥ 30.0 | Ideal sentence length range per Minimal Text Complexity Framework and PLAIN; provides rhythm while maintaining clarity |
| | % sentences >20 words | ≤ 25.0 | PLAIN guidance limits long sentences; allows occasional necessary complexity |
| | Average syllables per word | ≤ 1.5 | Simple vocabulary indicator; common everyday English; Easy Read principles |
| **WORD CHOICE** | % common words (Zipf freq) | ≥ 92.0 | Dale-Chall word list; expert simplifiers achieved 92-97%; accounts for domain-specific terminology |
| | % words >3 syllables | ≤ 10.0 | PLAIN and Easy Read vocabulary simplicity; allows technical terms when defined |
| | Type-Token Ratio | ≤ 0.4 | Lower ratio = consistent vocabulary; Easy Read principle of same word for same concept |
| **GRAMMAR & STYLE** | Passive voice % | < 5.0 | PLAIN/Easy Read preference for active voice; expert documents averaged 0.5-2.5% |
| | Passive voice count | < 10 | Absolute limit for documents under 2,000 words; scales with document length |
| | Negation per 100 words | < 1.0 | Easy Read guidance minimizes negatives; WCAG COGA recommendations |
| | Contractions count | ≤ 0 | Formal organizational documents avoid contractions; varies by organizational style |
| | ALL CAPS words | ≤ 0 | Accessibility guidelines avoid all-caps (harder to read); use bold or headings instead |
| | Acronym count | < 5 | PLAIN guidance defines on first use; limits total; Easy Read avoids when possible |
| **STRUCTURE** | Avg words per paragraph | < 100.0 | Easy Read and PLAIN chunking guidance; one main idea per paragraph |
| | % paragraphs >120 words | ≤ 0 | Avoid very long paragraphs; break complex information into digestible chunks |
| | Heading count | ≥ 1 | PLAIN and WCAG structure/navigation guidelines; meaningful headings aid comprehension |
| | % headings >8 words | ≤ 0 | Headings should be concise and descriptive; PLAIN recommends 5-8 words maximum |
| **FORMATTING** | Minimum font size | ≥ 12pt | WCAG accessibility standards; readability for people with low vision |
| | Total words | Variable | Context-dependent; no universal target (completeness may require length) |
| | Total sentences | Variable | Context-dependent; related to document purpose and scope |

Table 3. Complete Metrics, Targets, and Justifications

| Category | Metric | Guidance Target | Original | Organization-simplified |
|---|---|---|---|---|
| **Readability** | Flesch-Kincaid Grade Level | 3.0-8.0 | 14.6 | 13.3 |
| | Gunning Fog Index | ≤8.0 | 17.2 | 15.5 |
| | SMOG Index | ≤8.0 | 15.6 | 14.4 |
| | Flesch Reading Ease | ≥60.0 | 31.5 | 37.4 |
| **Sentence Structure** | Avg Words per Sentence | ≤15.0 | 22.7 | 21.4 |
| | Sentences 8-10 words % | ≥30.0 | 8.2% | 6.9% |
| | Sentences >20 words % | ≤25.0 | 36.4% | 34.8% |
| | Avg Syllables per Word | ≤1.5 | 1.88 | 1.75 |
| **Word Choice** | Common Words % | ≥92.0 | **97.3%** | **97.3%** |
| | Words >3 Syllables % | ≤10.0 | **9.1%** | **7.8%** |
| | Type-Token Ratio | ≤0.4 | **0.37** | 0.42 |
| **Grammar & Style** | Passive Voice % | <5.0 | **1.5%** | **1.6%** |
| | Passive Voice Count | <10 | 13 | **8** |
| | Negation per 100 words | <1.0 | **0.7** | **0.7** |
| | Contractions | ≤0 | **0** | **0** |
| | ALL CAPS words | ≤0 | 1 | 2 |
| | Acronym Count | <5 | 41 | 28 |
| **Structure** | Avg Words per Paragraph | <100.0 | 400.4 | 624.6 |
| | Paragraphs >120 words | ≤0 | 1 | 1 |
| | Heading Count | ≥1 | 2 | 0 |
| | Headings >8 words | ≤0 | **0** | **0** |

Table 4. Aggregate comparison of original and organization-crafted simplified versions against simplification targets for documents from Organization A (N=8). Each metric value represents the mean across all documents from this organization. Values in **bold** meet the specified targets. Targets synthesized from Plain Language federal guidelines (PLAIN) [55], Easy Read standards [43], and Minimum Text Complexity (MTC) framework [20]. See Table 3 for complete target justifications.

| Category | Metric | Target | Original | Organization-crafted |
|---|---|---|---|---|
| **READABILITY** | Flesch-Kincaid Grade Level | 3.0-8.0 | 17.4 | 11.1 |
| | Gunning Fog Index | ≤8.0 | 20.9 | 13.3 |
| | SMOG Index | ≤8.0 | 17.9 | 12.7 |
| | Flesch Reading Ease | ≥60.0 | 27.7 | 53.2 |
| **SENTENCE STRUCTURE** | Avg Words per Sentence | ≤15.0 | 31.2 | 21.3 |
| | Sentences 8-10 words % | ≥30.0 | 7.2% | 9.1% |
| | Sentences >20 words % | ≤25.0 | 41.2% | 40.6% |
| | Avg Syllables per Word | ≤1.5 | 1.76 | 1.56 |
| **WORD CHOICE** | Common Words % | ≥92.0 | **97.5%** | **97.8%** |
| | Words >3 Syllables % | ≤10.0 | **8.8%** | **4.2%** |
| | Type-Token Ratio | ≤0.4 | **0.31** | 0.46 |
| **GRAMMAR & STYLE** | Passive Voice % | <5.0 | **1.5%** | **0.7%** |
| | Passive Voice Count | <10 | 54 | **3** |
| | Negation per 100 words | <1.0 | **0.8** | **0.3** |
| | Contractions | ≤0 | **0** | **0** |
| | ALL CAPS words | ≤0 | 43 | **0** |
| | Acronym Count | <5 | 59 | 38 |
| **STRUCTURE** | Avg Words per Paragraph | <100.0 | 2005.7 | 152.7 |
| | Paragraphs >120 words | ≤0 | 1 | 1 |
| | Heading Count | ≥1 | 0 | 0 |
| | Headings >8 words | ≤0 | **0** | **0** |

Table 5. Aggregate comparison of original and organization-crafted simplified versions against simplification targets for documents from Organization B (N=11). Each metric value represents the mean across all documents from this organization. Values in **bold** meet the specified targets. Targets synthesized from Plain Language federal guidelines (PLAIN) [55], Easy Read standards [43], and Minimum Text Complexity (MTC) framework [20]. See Table 3 for complete target justifications.

| Category | Metric | Target | Original | Organization-crafted |
|---|---|---|---|---|
| **READABILITY** | Flesch-Kincaid Grade Level | 3.0-8.0 | 11.3 | **7.5** |
| | Gunning Fog Index | ≤8.0 | 10.7 | 9.4 |
| | SMOG Index | ≤8.0 | 11.2 | 10.2 |
| | Flesch Reading Ease | ≥60.0 | 40.2 | **66.5** |
| **SENTENCE STRUCTURE** | Avg Words per Sentence | ≤15.0 | **14.9** | **14.1** |
| | Sentences 8-10 words % | ≥30.0 | 10.6% | 18.8% |
| | Sentences >20 words % | ≤25.0 | **13.2%** | **14.7%** |
| | Avg Syllables per Word | ≤1.5 | 1.79 | **1.49** |
| **WORD CHOICE** | Common Words % | ≥92.0 | **96.3%** | **96.8%** |
| | Words >3 Syllables % | ≤10.0 | **8.6%** | **5.6%** |
| | Type-Token Ratio | ≤0.4 | 0.40 | **0.29** |
| **GRAMMAR & STYLE** | Passive Voice % | <5.0 | **1.2%** | **0.2%** |
| | Passive Voice Count | <10 | 20 | **7** |
| | Negation per 100 words | <1.0 | **0.8** | **0.7** |
| | Contractions | ≤0 | **0** | **0** |
| | ALL CAPS words | ≤0 | 31 | 21 |
| | Acronym Count | <5 | 44 | 35 |
| **STRUCTURE** | Avg Words per Paragraph | <100.0 | **25.2** | **18.2** |
| | Paragraphs >120 words | ≤0 | 6 | **0** |
| | Heading Count | ≥1 | 0 | 0 |
| | Headings >8 words | ≤0 | **0** | **0** |

Table 6. Aggregate comparison of original and organization-crafted simplified versions against simplification targets for documents from ASAN (N=5). Each metric value represents the mean across all documents from this organization. Values in **bold** meet the specified targets. Targets synthesized from Plain Language federal guidelines (PLAIN) [55], Easy Read standards [43], and Minimum Text Complexity (MTC) framework [20]. See Table 3 for complete target justifications.

| Category | Metric | Target | Original | Organization-crafted |
|---|---|---|---|---|
| **READABILITY** | Flesch-Kincaid Grade Level | 3.0-8.0 | 15.5 | 8.9 |
| | Gunning Fog Index | ≤8.0 | 18.3 | 10.9 |
| | SMOG Index | ≤8.0 | 16.4 | 11.2 |
| | Flesch Reading Ease | ≥60.0 | 22.5 | 54.9 |
| **SENTENCE STRUCTURE** | Avg Words per Sentence | ≤15.0 | 21.7 | **13.3** |
| | Sentences 8-10 words % | ≥30.0 | 5.0% | 18.3% |
| | Sentences >20 words % | ≤25.0 | 30.9% | **18.3%** |
| | Avg Syllables per Word | ≤1.5 | 1.92 | 1.64 |
| **WORD CHOICE** | Common Words % | ≥92.0 | 91.6% | **93.6%** |
| | Words >3 Syllables % | ≤10.0 | 11.0% | **6.0%** |
| | Type-Token Ratio | ≤0.4 | **0.18** | **0.12** |
| **GRAMMAR & STYLE** | Passive Voice % | <5.0 | **0.7%** | **0.5%** |
| | Passive Voice Count | <10 | 92 | 48 |
| | Negation per 100 words | <1.0 | **0.6** | **0.6** |
| | Contractions | ≤0 | **0** | **0** |
| | ALL CAPS words | ≤0 | 191 | 120 |
| | Acronym Count | <5 | 76 | 60 |
| **STRUCTURE** | Avg Words per Paragraph | <100.0 | **15.2** | **23.7** |
| | Paragraphs >120 words | ≤0 | 24 | 2 |
| | Heading Count | ≥1 | 0 | 0 |
| | Headings >8 words | ≤0 | **0** | **0** |

Table 7. Aggregate comparison of original and organization-crafted simplified versions against simplification targets for documents from CDT (N=9). Each metric value represents the mean across all documents from this organization. Values in **bold** meet the specified targets. Targets synthesized from Plain Language federal guidelines (PLAIN) [55], Easy Read standards [43], and Minimum Text Complexity (MTC) framework [20]. See Table 3 for complete target justifications.