Foundations and Trends® in Human-Computer Interaction

Readability Research: An Interdisciplinary Approach

Suggested Citation: Sofie Beier, Sam Berlow, Esat Boucaud, Zoya Bylinskii, Tianyuan Cai, Jenae Cohn, Kathy Crowley, Stephanie L. Day, Tilman Dingler, Jonathan Dobres, Jennifer Healey, Rajiv Jain, Marjorie Jordan, Bernard Kerr, Qisheng Li, Dave B. Miller, Susanne Nobles, Alexandra Papoutsaki, Jing Qian, Tina Rezvanian, Shelley Rodrigo, Ben D. Sawyer, Shannon M. Sheppard, Bram Stein, Rick Treitman, Jen Vanek, Shaun Wallace and Benjamin Wolfe (2022), "Readability Research: An Interdisciplinary Approach", Foundations and Trends® in Human-Computer Interaction: Vol. 16, No. 4, pp 214–324. DOI: 10.1561/1100000089.

Sofie Beier
Sam Berlow
Esat Boucaud
Zoya Bylinskii
Tianyuan Cai
Jenae Cohn
Kathy Crowley
Stephanie L. Day
Tilman Dingler
Jonathan Dobres
Jennifer Healey
Rajiv Jain
Marjorie Jordan
Bernard Kerr
Qisheng Li
Dave B. Miller
Susanne Nobles
Alexandra Papoutsaki
Jing Qian
Tina Rezvanian
Shelley Rodrigo
Ben D. Sawyer
Shannon M. Sheppard
Bram Stein
Rick Treitman
Jen Vanek
Shaun Wallace
Benjamin Wolfe

now
the essence of knowledge
Boston — Delft

Contents

Readability Research: An Interdisciplinary Approach

Sofie Beier[1], Sam Berlow[2], Esat Boucaud[3], Zoya Bylinskii[4], Tianyuan Cai[4], Jenae Cohn[5], Kathy Crowley[6], Stephanie L. Day[3], Tilman Dingler[7], Jonathan Dobres[3], Jennifer Healey[4], Rajiv Jain[4], Marjorie Jordan[6], Bernard Kerr[4], Qisheng Li[8], Dave B. Miller[17], Susanne Nobles[9], Alexandra Papoutsaki[10], Jing Qian[12], Tina Rezvanian[4], Shelley Rodrigo[12], Ben D. Sawyer[3], Shannon M. Sheppard[13], Bram Stein[14], Rick Treitman[4], Jen Vanek[15], Shaun Wallace[11] and Benjamin Wolfe[16]

[1]*Centre for Visibility Design, Royal Danish Academy, Denmark*

[2]*Typography for Good, USA*

[3]*The Readability Consortium, University of Central Florida, USA; sawyer@inhumanfactors.com*

[4]*Adobe Inc., San Francisco, USA; bylinski@adobe.com*

[5]*California State University, USA*

[6]*Readability Matters, USA*

[7]*University of Melbourne, Australia*

[8]*Paul G. Allen School of Computer Science and Engineering, University of Washington, USA*

[9]*ReadWorks, USA*

[10]*Department of Computer Science, Pomona College, USA*

[11]*Brown University, USA*

[12]*University of Arizona, USA*

Sofie Beier, Sam Berlow, Esat Boucaud, Zoya Bylinskii, Tianyuan Cai, Jenae Cohn, Kathy Crowley, Stephanie L. Day, Tilman Dingler, Jonathan Dobres, Jennifer Healey, Rajiv Jain, Marjorie Jordan, Bernard Kerr, Qisheng Li, Dave B. Miller, Susanne Nobles, Alexandra Papoutsaki, Jing Qian, Tina Rezvanian, Shelley Rodrigo, Ben D. Sawyer, Shannon M. Sheppard, Bram Stein, Rick Treitman, Jen Vanek, Shaun Wallace and Benjamin Wolfe (2022), "Readability Research: An Interdisciplinary Approach", Foundations and Trends® in Human-Computer Interaction: Vol. 16, No. 4, pp 214–324. DOI: 10.1561/1100000089.

[13] *Chapman University, Department of Communication Sciences and Disorders, USA*
[14] *The Type Founders, USA*
[15] *World Education, Digital Learning and Research, USA*
[16] *University of Toronto Mississauga, Canada*
[17] *Tufts University, USA*

ABSTRACT

The control provided by digital displays over how visual information is presented to readers has the potential to improve reading for each and every reader, regardless of ability or diagnosis. On screens, text is fluid, allowing for individual customization based on reader needs, content, and reading task. This represents a profound shift in how we think about reading, because text is no longer rendered immutable by writers, designers or publishers at a single stage, and human-computer interaction research is key to realizing its potential. Targeted changes to the visual characteristics of text on screens increases the ease with which a reader can process and derive meaning. In this review, we provide a comprehensive introduction to interdisciplinary methodologies, tools, and materials required for readability research focused on the individual reader. We call on the HCI community to contribute to our growing understanding of readers' needs, to study the interactions between text, user, and task, and to build the tools and interfaces needed to improve reading outcomes for all.

Keywords: reading; readability; text; document; information processing; typography; design; reading interfaces.

1

Introduction

From the moment we wake up to the moment we end our day, we use interfaces built out of the written word. Textual information remains now, as it has for centuries, the cornerstone of human information acquisition. The wide adoption of smartphones, tablets, e-readers and personal computers has shifted the bulk of this reading from inflexible paper to digital content. The amount of information we acquire through reading digitally has grown rapidly over the last 15 years, and continues to grow. At the same time, literacy rates in the United States are staggeringly low: 130 million U.S. adults ages 16 to 74 (54% of the population) read below a sixth-grade level (Rothwell, 2020). Alarmingly, as of a 2022 report by the National Center for Education Statistics, young children's reading scores have experienced the largest decline since 1990 (U.S. Department of Education, 2022). Furthermore, dyslexia – the most common language-based learning disability – affects 15–20% of the population and represents 80–90% of all those with learning disabilities (International Dyslexia Association, 2022; The Yale Center for Dyslexia & Creativity, 2022). Readability research, as we describe here, takes a fundamentally individual approach to what each reader needs. Each reader, even readers who may not struggle, have their

own individual needs. Meanwhile, adapting the written word to the individual reader has never been easier, and the goal of maximizing individual reading efficacy is increasingly attainable.

Readability encapsulates the properties of a document which determine the ease and success with which individual readers decipher, process, and determine meaning from the text. These include (1) content, (2) document-level aspects, and (3) format features. These format features, which include all typographic elements, can have profound impacts on individual readers' speed and comprehension. Readability is discrete from legibility, which refers, in print or handwriting, to the property of being clear enough to read. In traditional printing, a single legible aesthetically pleasing layout was all that was possible, but digital displays now allow the potential of individuation, changing how text appears for each reader. Digital flexibility allows readability interventions to increase accessibility and efficacy. Here, we argue that this opportunity can be addressed with interdisciplinary methods spanning Human-Computer Interaction (HCI), design, user research, psychophysics, neuroscience, and data science.

This monograph is about *Format Readability* – the visual and typographic features of the text, which include font choice, size, spacing, and related attributes. We focus on format readability, rather than content and document factors, although we acknowledge and discuss their importance. We begin by discussing reading itself before turning to the readers. We then talk about reading materials for research, how those materials can be shown to readers, the research tools used to study readability, the experimental paradigms used in this research, and how the resulting data can be analyzed. We conclude by inviting researchers to ask their own questions in readability, using our review as a starting point for conducting readability studies and designing reading interfaces.

The time is now. To date, writers, publishers, and designers have been in control of the reading experience. However, digital reading provides a paradigm shift, through the multitude of device types, screen qualities, digital interfaces, and software settings available to readers. Depending on the technology, the readers – literate or nearly literate children or adults – can now control font size, screen polarity, spacing,

font choice, and other formatting choices. Amazon's Kindle, Apple's iBooks, Microsoft's Immersive Reader, Adobe's Liquid Mode, and modern web browsers all provide some of these controls, occasionally branded as accessibility features. Recent studies indicate that it is possible to dramatically improve reading for each individual – to make it much easier for struggling readers to read and for good readers to read even more efficiently by changing and, more significantly, personalizing the appearance of the text. The power of personalization and individuation has been shown with young (Crowley and Jordan, 2019a,b; Day *et al.*, 2022; Sheppard *et al.*, 2022a,b) and adult readers (Ball *et al.*, 2021; Cai *et al.*, 2022; Wallace *et al.*, 2020a, 2022a,b; Watson and Wallace, 2021), and suggests that every reader, at every level, can realize benefits if we can determine what they, individually, need and give it to them.

No one discipline or field has all the tools or answers, and readability work is inherently interdisciplinary. The authors of this monograph include vision scientists, technology experts, educators, designers, typographers, and data scientists; together, we represent voices from academia, the tech industry, and non-profit institutions, driven by common goals to improve the reading interfaces of today. This monograph is intended as a practical foundational resource for anyone interested in pushing readability research forward, including HCI researchers, practitioners, educators, tech companies, type designers, policymakers, and engineers. Our review cannot cover every topic we touch upon in full detail so we extensively reference related literature, to provide a starting point for our reader to build on. Different sections of this review may be individually useful to different readers from different backgrounds. Taken as a whole, if read from front to back, our review should be accessible to the budding HCI researcher, with prior exposure to cognitive science, computer science, or related disciplines, but without assuming specific knowledge about the psychology of reading, typography, or the latest technological advances, all of which we introduce here.

2

Types of Reading

Readability is influenced by content, typographical features, and document-level aspects, among other properties. A change to the document's readability affects the ease with which a reader can succeed at extracting the information they need. Optimal readability entails a fit between document, reader, and context, producing better reading outcomes. Readability research thus focuses on studying the interactions between the reader, the reading material, and the reading interface. In this monograph, the focus is further reduced to the presentation of text.

It is essential for researchers working on readability to understand the core tenets of the process of reading. As readability is a narrower and less studied topic than reading, we will occasionally borrow methodology from reading studies throughout this monograph. Reading includes deciphering, processing, and making meaning of text and may look very different depending on when, how, and why we read. Considering reading, and by extension readability, means considering everything from how text is presented to the physical process of reading to the strategies that readers use while reading and how those strategies change based on readers' task and motivation.

Reading is a complex process involving phonologic, semantic, perceptual, and cognitive inputs. The leading computational models simulating human visual word recognition, have identified a parallel process that involves a bottom-up operation of identifying letter features and whole letters and a top-down operation of lexical access to words and word parts (Reichle, 2021). This explains the well-established word superiority effect (Reicher, 1969; Wheeler, 1970) (words are easier to recognize than letters), as words receive input signals from both the top-down and bottom-up operations while letters mostly receive input signals from the bottom-up operation of letter features. It has further been shown that for sentence reading, letter decoding accounts for 62% of the reading rate, word reading accounts for 16%, while contextual structure of sentences accounts for the remaining 22% (Pelli and Tillman, 2007).

Reading can encompass encountering a single word as we move through the world, like "stop" on a sign, reading an instruction manual while building a bookshelf, or reading several chapters in a novel. All of these reading activities, therefore, occur on two axes of intention: time and purpose (Figure 2.1); that is, they cannot be thought of as only the simple process of deciphering text. As we consider how we maximize readability, we need to be mindful of why we read so that the display decisions we make support particular tasks.

2.1 Reading on the Temporal Axis

Reading is a process that the reader does over time, with scanning or searching a body of text (or an interface) for a given word being the simplest. *Glanceable reading* describes fast information intake that can be performed within one or two eye fixations (see Sawyer *et al.*, 2020 for single-word glanceable reading); a reader's attention is focused on at most a few words, rather than reading or paying attention to the interstitial text, such as when reading alerts on smart watches, or during driving, noticing the text on road signs (Dobres *et al.*, 2017a). Searching a text or interface for a word may include skimming it to gain context (Wolfe, 2021; Wong *et al.*, 2017). Skimming and searching behaviors can be considered a form of *interlude reading* (Wallace *et al.*, 2020b,2022a), where readers typically read multiple sentences or a shorter part of a

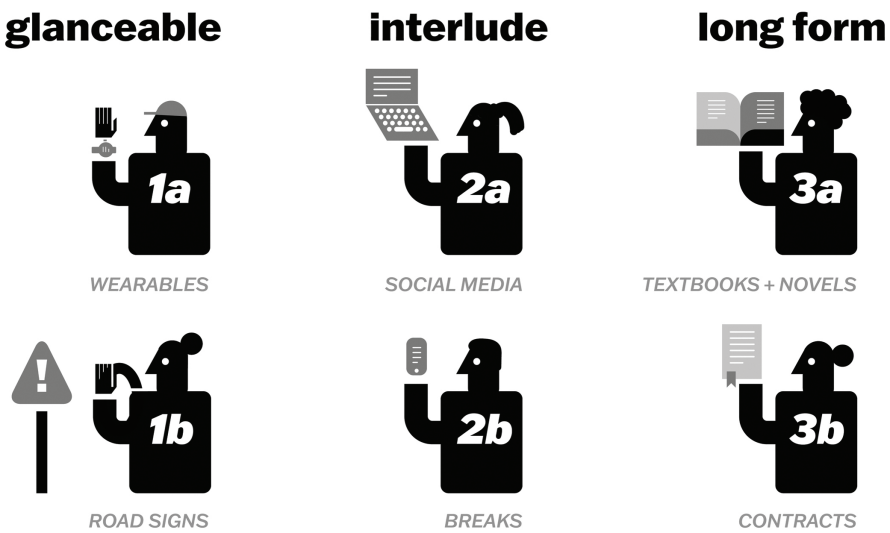

Figure 2.1: Glanceable reading takes place in the context of other tasks, where the text only receives one or two glances, e.g., a notification or a road sign. In long form reading, reading is the primary task, and undue attention on other tasks is detrimental to reading. Interlude reading subtenants the complex space between, in which reading is one of many interleaved tasks.

longer document (e.g., reading social media posts or short portions of news articles, often while engaged in another task).

Readers requiring a more complete understanding will read most, if not all, of the words on the screen and will move progressively through a document, reading for comprehension and, potentially, for deeper meaning. According to Wolf (2018), we can define "deep reading" as reading processes, which:

> *underlie* our abilities to find, reflect, and potentially expand upon *what matters* when we read. They represent the full sum of the cognitive, perceptual, and affective processes that prepare readers to apprehend, grasp, and assimilate the essence of what is read (p. 112).

Deeper reading acts, like Wolf describes, are *long-form reading* (Seaboyer and Barnett, 2019), where paragraphs and pages are focused on without

interruption. Reading academic articles (like this one), textbooks, technical manuals, and legal documents are all examples of this. However, when we read, we do not explicitly choose to engage in only one of these forms of reading; we are likely to switch between them as required.

Reading in a temporal sense is going to vary as a function of the task the researcher is intending to study. While reading a single word in an interface (e.g., an icon label) closely resembles glanceable reading tasks as commonly used in the laboratory (e.g., in lexical decision tasks, or other psychophysical tasks in controlled environments), reading is rarely that simple. Even a task as seemingly simple as reading an icon label likely exists in a larger context of distracting tasks, since the reader is, in this example, likely planning to interact with the interface and may be embedded in a larger environment with its own distractions, as would apply if they are walking and using their smartphone. Interlude reading is even more likely to be embedded in a larger context, and a reader might be moving between email and a document, or quickly reading a text message before returning to their in-person conversation. Long-form reading is more likely to be the user's focus, since it embeds a sense of prolonged time on task, but outside of the laboratory, it is very rare that any of us truly engage in just one task at a time. Our intent in providing this temporal framework is to help HCI researchers think about how their particular tasks unfold over time, but we caution that the larger context is key.

2.2 Reading on the Purpose Axis

Considering reading from the point of view of what the reader *needs* brings us to the question of purpose: reading strategies and motivations. When we think about supporting readers, we think first about basic literacy acquisition: how do we help learners sound out letters and process phonemes? This is important, but an ability to read goes beyond basic literacy acquisition, since the reader must apply their visual knowledge to social and cultural meaning-making.

(1) When we read for *content uptake*, we go beyond decoding text to learn about processes, key terms, concepts, or definitions. We

might read for content uptake when we are trying to learn a new term, skill, or idea, like learning a new recipe (Wolf, 2018).

(2) When we read *rhetorically*, we read for an understanding about the context of the written work itself, its purpose and audience, and what it seeks to communicate. Being a critical consumer of news and social media posts requires this, and schools have been encouraged to teach these skills directly so that students are prepared to be informed consumers (Brandt, 1990; Haas and Flower, 1988; Sweeney, 2018).

(3) When we read for *research*, we read to collect ideas to create and support a larger body of knowledge, aggregating perspectives and extrapolating themes, patterns, and ideas. A financial analyst seeking to make a prediction engages in this type of reading (Bizup, 2008; Downs, 2010; Jamieson, 2013).

(4) When we read for *analysis*, we read to consider broader themes, ideas, or patterns. Often referred to as "close reading", reading for analysis involves examining readings at the word, sentence, and/or paragraph-level to make sense of how a text might fit into broader cultural or historical narratives (e.g., in historical or literary studies) (Fang, 2016; Fisher and Frey, 2014).

These reading purposes depend on a variety of reading strategies (Carillo, 2017; Petrosky *et al.*, 2010). Reading for content uptake, for example, requires information retrieval, summarization, and, sometimes, memorization. Reading for analysis, on the other hand, requires critical thinking and text contextualization or historicization and an ability to think and imagine meanings beyond the literal space of the text itself, and these strategies are themselves a major body of work. (Jo *et al.*, 2015; Schildbach and Rukzio, 2010; Schnell *et al.*, 2009; Wei *et al.*, 2020).

While much can be learned about readability by studying how we read as a temporal process, it is insufficient without the question of reading purpose and strategy. If the goal is to give the user the best experience possible, researchers must consider the visual and temporal

elements as well as the cognitive and motivational processes without which we cannot understand why readers act the way they do. This complexity, which requires researchers to ask questions about the material being read, the amount of time provided, the visual appearance of text, the reader's interpretative abilities and their motivations is required to probe how and why we read. It is certainly possible to design an informative study which focuses on one facet of this, for example, readers' motivations or the impact of time pressure on reading single words, but these designs are inherently limited.

3

Readers

Readability research is first and foremost about *readers* – it asks questions about the ease with which they can successfully decode the document (i.e., decipher, process, and make meaning of the text). Researchers interested in particular populations (e.g., children, older adults, those with visual impairments, financial workers, cyber defenders, readers with dyslexia) should bear in mind each group's needs and abilities, whether in the context of compensation, motivation, fatigue, or their ability to do the task. Here, we discuss the reader themselves with an eye towards designing readability studies, as well as discussing reader-relevant considerations in experiment design, that is, where and how we ask them to read. We also briefly cover the topic of research ethics, as an integral component.

3.1 Who Are the Readers?

Readability affects any literate or nearly literate child (Crowley and Jordan, 2019a,b; Day *et al.*, 2022; Sheppard *et al.*, 2022a,b) or adult (Ball *et al.*, 2021; Cai *et al.*, 2022; Wallace *et al.*, 2020a,2022a,b; Watson and Wallace, 2021). Improved readability may also help bridge gaps in the process of learning to read. Thus, any reader can help us understand

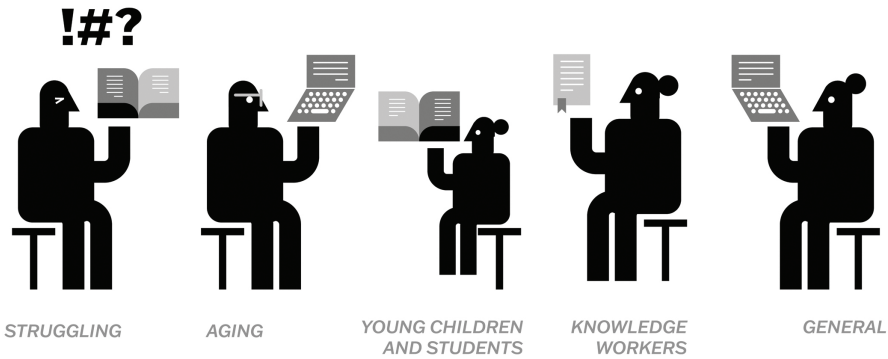

Figure 3.1: There are no outlier populations in readability; rather, this work addresses a continuum of need and a continuum of skill, in which each individual could be provided with affordances. Format readability studies show sizable gains even for readers competent by their own assessment, and within the "normal" range of reading speed and comprehension.

how they read and how they might read more effectively by participating in readability studies (Figure 3.1). Investigating readers who struggle can help us understand the cognitive, linguistic, and environmental factors that influence how we read (Galliussi *et al.*, 2020; Goel *et al.*, 2012; Martelli *et al.*, 2009; Rello and Baeza-Yates, 2016; Rello *et al.*, 2012, 2017; Shaywitz and Shaywitz, 2020; Storer and Branham, 2019; Watson and Wallace, 2021; Yamabe and Takahashi, 2007). In our view, there are no outlier populations; rather, participants should be seen as lying on a continuum of need and on a continuum of skill.

Readers with dyslexia. A population of particular interest to many researchers are readers with developmental *dyslexia*. Given the complex nature of reading, involving the visual, phonological, and semantic processing of the reading materials, it is increasingly clear that dyslexia is a multifactorial condition (Pennington, 2006; Ramus *et al.*, 2003; Shaywitz and Shaywitz, 2020; Vellutino *et al.*, 2006). A popular theory is the phonological impairment theory of dyslexia, although this is debated (Catts *et al.*, 2017; Fostick and Revah, 2018). Also, whether dyslexic readers experience increased visual crowding or reduced motion sensitivity remains not fully understood (e.g., Demb *et al.*, 1997; Joo *et al.*, 2018; Li *et al.*, 2020; Martelli *et al.*, 2009; Olulade

et al., 2013). Prior research also recognizes the coexistence of ADHD and Autism within the dyslexic population, which brings additional individual variability into the picture (Fletcher *et al.*, 2019; Germanò *et al.*, 2010; Pennington, 2006). Readers with dyslexia are only one example of readers who struggle; for example, non-native speakers may have deficits in oral language and reading comprehension skills, despite adequate decoding capabilities (Spencer and Wagner, 2017) and could benefit from readability-focused interventions and help us understand reading.

Aging readers. People with declining reading skills may similarly appreciate improvements to their reading experience to compensate for the loss of visual acuity (e.g., age-related presbyopia), declining cognitive ability (Bokulich *et al.*, 2016; Owsley, 2011), and changes to other sensory capabilities such as crowding and visual span (Legge *et al.*, 2007; Levi, 2008; Rayner *et al.*, 2010). Reading speed slows down with age (after age 40 (Beier and Oderkerk, 2019a)), visual acuity decreases and critical print size increases (Bernard *et al.*, 2001; Calabrese *et al.*, 2016). Like with readers who struggle, aging readers can also benefit from having text customized to their own particular needs (Beier and Oderkerk, 2019a; Cai *et al.*, 2022; Wallace *et al.*, 2022a).

Readers with low vision. Vision impaired people often struggle to read text at low luminance contrast (Rubin and Legge, 1989) and have trouble seeing higher spatial frequencies (Legge *et al.*, 1985). The nature of reading difficulties can vary greatly depending on diagnosis and individual differences, with the biggest variation being attributed to the absence or presence of central vision (Legge *et al.*, 1985). The most widespread low-vision diagnosis among elderly in developed countries, is age-related macular degeneration (AMD) where one of the main symptoms is loss of central vision. AMD patients are known to have a greater instability in eye-movement reading patterns (Kumar and Chung, 2014) and to benefit from magnification as well as excessive added space within and between letters (Beier *et al.*, 2021; Bernard *et al.*, 2001; Xiong *et al.*, 2018).

Readers learning to read. People learning to read can also benefit from individual, visual reading interventions (Hughes and Wilkins, 2002; Powell and Trice, 2020; Reid *et al.*, 2004) including older readers not

fully proficient in early literacy skills such as letter knowledge and print awareness (e.g., adult literacy learners) (Graesser *et al.*, 2019; Sabatini *et al.*, 2011). The learning-to-read population of young children between the ages of 3–7 is not well understood in the context of readability, particularly in regard to the formation of sound-letter understanding. An increased understanding of how readability impacts this critical process has the potential for great impact, since when students are not proficient readers by fourth grade, they are far less likely to complete high school with serious consequences for economic and civic prospects for the remainder of their lives (Cramer *et al.*, 2014; Hernandez and Napierala, 2013).

Knowledge workers. Other points on the continuums of need and skill are knowledge workers who may be completing different reading tasks (Section 2) – this includes medical experts leafing through patient records (Bouaud and Seroussi, 1996; Elson and Connelly, 1995; Henriksen *et al.*, 2020; Nygren *et al.*, 1992; van Engen-Verheul *et al.*, 2016), cybersecurity experts scrutinizing text for potential threats (Ehrlich *et al.*, 2017; Jang-Jaccard and Nepal, 2014; Lee *et al.*, 2019; Lotem *et al.*, 2012), financial analysts integrating information to make predictions (Bradshaw, 2011; Hoitash *et al.*, 2021; Lehavy *et al.*, 2011; Li, 2010; Loughran and McDonald, 2010; Ravula, 2021), and scientists immersing themselves in the literature to stay up to date (Yeung *et al.*, 2018). These populations can also benefit from readability research and individual customization, because it will improve these readers' throughput and quality of reading.

The "general reader". This reader is competent at the reading task, and often reads regularly for pleasure, whether this is in interludes or for longer blocks of time. Among these so-called "normal" readers, we observe a wide range of reading speeds and abilities, with more typical speeds in the 200–300 WPM range (Brysbaert, 2019; Legge, 2007; Taylor, 1965), although it is not uncommon to find less-versed readers reading below 180 WPM. Under some conditions, it is possible to achieve speeds of 600-700 WPM or higher if using skimming strategies (Just and Carpenter, 1987; Rayner, 1998) or RSVP reading (Carver, 1990; Legge, 2007). In fact, the "general reader", like the "average user" does not actually exist, but they are a useful fiction and serve to remind

us of the universal potential of readability and its potential to benefit every reader.

HCI researchers and others interested in developing tools for improving readability need to study reading across many populations, considering diversity in age, reading skill level, deficits, and other characteristics. This will enable the development of individualized recommendations and a better understanding of the wide variety of readability factors which affect some individuals but not others. Of course, different populations of readers require different considerations. Across all populations, we must attend to certain global features that may influence our results. Aside from the commonplace factors of age, education level, and occupation, additional sources of variability can result from whether vision correction is available to participants (and whether it is used during the study), reading proficiency, prior diagnoses of reading or learning disabilities, eye conditions (e.g., cataracts, glaucoma, retinal degradation), possible influences of stimulants or depressants (including common ones like nicotine and caffeine), other languages spoken/read by the participant, lighting conditions, and reading environment at time of study, etc. (see Appendix B for some of the questions we ask participants in our readability studies in order to capture some of these individual differences and possible confounding factors).

3.2 Population Size: Quantitative Versus Qualitative Considerations

Having chosen who to study, you must then consider how many readers to recruit. The degree to which researchers can generalize from their data is a function of the population and the question; there is no magic number of participants.

Sample size matters. This is not just dependent on the raw number of participants, but also on the number of trials per participant, the total amount of data gathered, and the diversity of the participant pool – its ability to act as a representative sample. A good way to think about this is not "how many participants" but how many "experimental units" are available for analysis. It is not uncommon for behavioral/neuroscience studies to derive conclusions from studies of

12–24 participants (Sihoe, 2015). Data from such "low-N" studies needs to be very high quality, with possible confounding factors identified and controlled for, and many trials run per participant to ensure the trends captured are representative of an underlying truth. For these reasons, smaller studies are usually carried out in more controlled, laboratory settings. On the other hand, studies on hundreds of participants are more common in reading experiments for education research. Given the hierarchical structure of education data (schools within districts, classrooms within schools, and students within classrooms), larger sample sizes, particularly at the top levels, are desirable to capture a representative sample of the population and to ensure appropriate statistical power (Lee and Hong, 2021). When running large-scale studies of hundreds or thousands of participants, a large N can help wash out individual participant noise (Bolthausen and Wüthrich, 2013), at the expense of experimental control. The right N depends in part on the goal and design of the study – the metrics used, practical significance level desired, statistical power (Broberg, 2013; Demets and Lan, 1994), number and nature of variables in the experiment, and the overall design. Prior readability studies (Banerjee and Bhattacharyya, 2011; Bernard *et al.*, 2001; Boyarski *et al.*, 1998) have made general recommendations on the basis of dozens of participants' worth of data, but these recommendations can change dramatically as the number of participants is increased by an order of magnitude (Cai *et al.*, 2022; Wallace *et al.*, 2020a,2022a).

Individuals matter. Given our focus on individuation, a sample of fewer participants can also be very valuable. However, focusing on individuals cannot be at the expense of the larger whole, since it is necessary to determine whether individuals are idiosyncratic or if there are clusters within the larger group. While size matters, examining individual participants can tell the researcher what individuals need and why, by providing qualitative insights (Crowley and Jordan, 2019a). In fact, recent work on individuation demonstrates that significant gains can be achieved if we look for clusters and focus on individual differences (see Cai *et al.*, 2022; Wallace *et al.*, 2020a,2022a).

3.3 Recruiting Participants

Recruit respectfully. Recruiting participants is essential to capture an accurate sample of high-quality, real-world data, and it must be done respectfully and ethically. Readers should be able to choose to participate, understanding the risks and benefits of doing so. Failing to consider this can put your participants at risk, damages trust between researchers and participants and makes future research more difficult.

Partner with a domain expert. Domain experts can simplify the recruiting process and provide important insights into the attributes and limitations of the target population. For example, partnering with a school district, individual educator, or third-party reading organization can provide access to student populations, help you navigate the complexities of working with minors, and help you develop the best, most informative study. You might approach organizations that work with adults developing literacy skills or second language learners or organizations that support K-12 students and educators.

Consider crowdsourcing platforms. If it makes sense for the study question to recruit a sample of the general population, crowdsourcing platforms like Amazon's Mechanical Turk, UserTesting.com, Crowdflower, and Prolific are available (Buhrmester *et al.*, 2011; Paolacci and Chandler, 2014; Peer *et al.*, 2017). These platforms provide a wide range of extrinsically-motivated users. Related approaches, like friendsourcing (relying on voluntary participation by friends) can be easy, but may bias your results (Brady *et al.*, 2015). You may also consider recruiting volunteer participants, who are intrinsically motivated (e.g., LabInTheWild), and choose to participate in studies for no compensation, simply because they want to. Motivating crowdsourced participants can be improved by providing your participants insights on how they compare with others in terms of reading speed, preference, and other reading tasks, as this may encourage them to share your study and return in the future (Ikeda and Bernstein, 2016).

Special populations can also be recruited via targeted messages in relevant forums like Reddit (Shatz, 2017) or advertising in social media or markets like Craigslist (Alto *et al.*, 2018). However, these are all prone to self-selection bias, i.e., participants who sign up may not

be representative of the larger population of interest (Ho *et al.*, 2015; Mason and Suri, 2011; Wong *et al.*, 2017). Some of these issues can be addressed by having a diverse research team from different areas (e.g., scientists, educators, technologists, designers) to help with participant recruiting and designing a study with greater awareness of the attributes and limitations of the target population.

3.4 Where to Conduct Studies

Intertwined with the question of *who* is the question of *where*. While your population of interest may constrain where you can run your study, you will still be faced with a number of choices about where you should conduct your study. Each location involves trade-offs between ease of recruiting, data quality, and ecological validity.

3.4.1 Laboratory-Based In-Person Studies

An important consideration for choosing where to conduct a study is data obtainability, accessibility and representativeness. From this perspective, in-person, laboratory studies offer the highest degree of control. Also, you have the option of taking physiological recordings from your participants while they read – via eye trackers, brain imaging technology, or other equipment that can be set up and carefully calibrated for each participant (see Section 5). For fundamental questions in readability, including many questions about the visual mechanics of reading (how readers move their eyes and why, depending on task, goal, experience, and strategy, see Section 2), laboratory studies are extremely revealing. However, there is always tension between laboratory and non-laboratory studies, since the behavior that readers show in the lab may not be the same as what they do at home. While some of these gaps can be overcome by replicating study designs outside the laboratory, or by developing and using more naturalistic tasks in the laboratory, this gap will always exist, and minimizing it is a function of developing fundamentally generalizable tasks and experiments.

3.4.2 Online Remote Studies

The main advantages of conducting remote studies – studies hosted online on crowdsourcing platforms – are the ease, speed, and availability of participants. In particular, the numbers of participants available online can be substantially higher than those that are able to come in for in-person studies. While remote studies can increase the diversity of the participant pool, it is important to realize that these platforms self-select for participants who can participate and choose to do so. Given shifts to remote work in the last few years, remote online studies are becoming an increasingly popular option in readability research.

Beware of unobserved reading behavior. When using measurements captured from online studies, or from situations where the participant is not directly observed, there are likely to be unseen behaviors which can impact your data, including distraction or multitasking (Ophir *et al.*, 2009; Reeves *et al.*, 2020), task switching, or "short-cutting" activities like taking screenshots (Brishtel *et al.*, 2020). While crowdsourced data can be easy to acquire, the data must be handled with care and analyzed for anomalies that should be excluded in a principled fashion from final analyses (Section 7.1).

Improving remote studies. While remote studies lose some internal validity by giving up control of the reading environment, they gain ecological validity by studying participant reading habits in their own environments. Recent work, like the "virtual chinrest" (Li *et al.*, 2020) has increased researchers' ability to control key visual factors like viewing distance (and therefore the *visual angle* of stimuli), allowing for a wider range of research questions to be asked outside of the laboratory. There has also been work on capturing eye movements remotely using webcams (Papoutsaki *et al.*, 2017), which can provide validation of whether participants complete the task honestly (did they move their eyes or did they just click through?). Data quality and motivation can also be improved by gamifying online studies or providing personalized feedback. In particular, readability studies can offer personalized font or reading format recommendations, motivating readers with possible improvements to their reading effectiveness (Ball *et al.*, 2021; Wallace *et al.*, 2020a; Watson and Wallace, 2021).

3.4.3 In-Context Studies

Another option for running studies outside of the laboratory is to partner with organizations or professionals with access to specific reader populations. These collaborations also increase external validity, studying readers in their usual contexts. It can also help ensure high data quality, since external collaborators may be able to help administer the study and observe participants (see also Section 3.3).

For example, partnering with educators to conduct both small- and large-scale studies can be an effective method to evaluate the impact of readability on reading outcomes. Teachers, and students themselves, can provide additional insights which can assist in assessing impacts, and has the benefit of assessing reading habits in their customary environments.

Similar benefits exist when looking to readers in professional contexts, for example, workers in military, healthcare, and financial institutions that engage in reading as part of their job. In these cases, your study may leverage the reading materials familiar to participants, provided there are no privacy or security risks. However, in all of these cases, it is also important to consider potential risks inherent to the study itself or recommendations you may make, since they might interfere with students' ability to learn or professionals' ability to do their jobs.

Another possibility for in-context studies is to consider the question of reading on smartphones, which readers use throughout their day in a wide range of settings. These readers may be able to engage with your study briefly, and any effects you see are likely to be quite generalizable, since the environmental conditions (e.g., lighting, noise levels, multitasking) are so variable and uncontrolled from an experimental design sense.

3.4.4 Replicating Studies Across Environments

Additional considerations for readability research include the degree to which reading behaviors replicate across environments and why they exist or do not in a particular context. Remote studies allow readability researchers to gather large samples of readers from natural settings at the cost of experimental control, rendering the second half of this question difficult to answer. Rerunning studies in controlled environments can

help tease apart these questions. Likewise, re-running laboratory studies when possible with crowdworkers can help you understand whether laboratory-recorded effects are generalizable. Similar considerations exist, of course, for educational work in readability, where classroom or home-based studies may differ from laboratory-based. A question to bear in mind here, particularly with our focus on individual needs and what influences them, is whether the conclusions we reach in any particular setting are broadly or narrowly applicable for readers and why.

3.5 Ethics

Research ethics, or how we balance our desires as readability researchers with the rights of our participants, is a vital consideration. Considering participants' rights as you design your study is as important as what texts you use and who you recruit. Doing ethical research can take many forms and varies by setting and country, so we will discuss universal fundamentals; that is, what any readability researcher must consider.

Studying readers must respect their choice to participate, both at the onset of research and throughout the process. It is important to consider the question of anonymity and potential harms if participants' data can be linked to their identity. If, for example, an experiment involves a screening procedure for a disorder, and that data were made publicly accessible, it could unintentionally harm the participant and their privacy. This is a particular concern with data that are difficult to truly anonymize, like video or audio of participants, and readers should be aware of what data is recorded and how it will be used.

There should also be some benefit to participating, whether that is simply helping increase knowledge of readability, understanding what helps a particular reader, financial compensation, or some combination of these. Returns can be made at the individual level and more broadly. For example, if you are studying young readers, you should convey what you learn back to the community of educators teaching them to read so that your work has an impact.

Research with school-age minors brings its own issues, including the need for parental consent, administrative approval and awareness of your

impact on other students. Data anonymity is particularly important here, as children are considered a particularly vulnerable population. Another consideration with this kind of work, but one that applies more generally to readability research, is that withholding an intervention is not ethical if it adversely impacts the control group's educational experience. Even readability studies with users need this kind of consideration; how are you changing their experience? What does it mean to do so respectfully and in ways that help you learn what you want while avoiding harm?

Overall, these questions should be part of the planning process for readability studies. If every reader is a potential participant in our studies, we should think about what it means to respect them, whether they are children learning to read, adults who struggle with reading or a doctor reading an electronic medical record.

4

Reading Materials

Inextricable from when, how, and why readers read is the question of *what* a reader is reading: the content and how it is designed (i.e., typographic and visual properties, Figure 4.1). In this monograph, while we mention content in a number of places, we turn our focus to the typographic considerations that can impact reading performance at both the group and individual level.

4.1 Content Curation and Leveling

Reading material must be curated to be appropriate in topic, length, and level to the target study population (for a list of some freely available materials for reading experiments, see Appendix C). Familiarity with the topic and interest in it are possible confounding factors (Spyridakis and Wenger, 1991; Wallace *et al.*, 2020a), as both can affect the ease and speed with which the readers consume the content and whether they switch to skimming (see Section 2). Length must also match the skills and abilities of the target population. Study fatigue or, worse, the inability to complete the study task can significantly affect a study's conclusions. Genre is relevant for speed and comprehension, as narrative

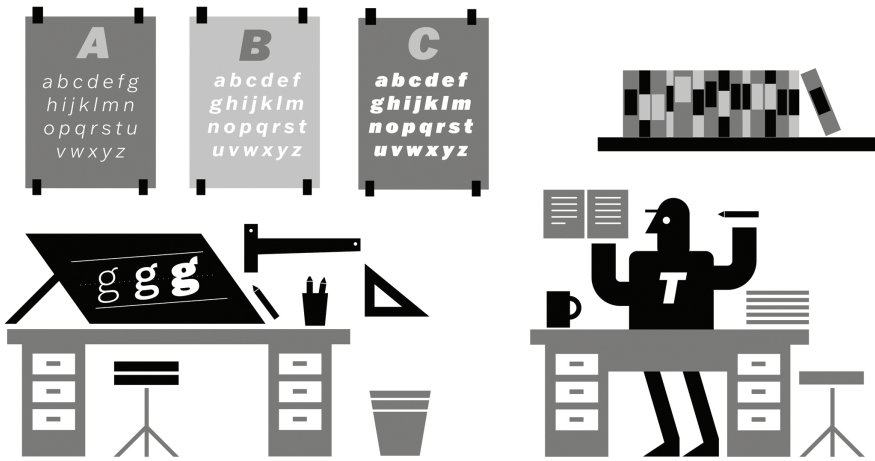

Figure 4.1: Typographic properties refer to the design of the content: how it is presented, and how it is laid out. Format readability research investigates impact to the reader that derives from the visual and typographic features of the text, which include font choice, size, spacing, and related attributes. This includes group-level questions such as "what typographic choices can help populations", as well as individuation questions "what typographic choices have impact when tailored to the individual".

texts tend to be easier for readers as they have a more familiar chronological organization (Hiebert *et al.*, 2010). Texts must also meet target populations' expectations to not distract and should not reflect bias or stereotypes about any group. For example, contentious topics may elicit reactance (Brehm, 1966), which may prompt significant cognitive engagement with the material, affecting speed, comprehension, and emotional affect. Finally, the topic of the reading material also directly affects the level of the material, particularly in vocabulary, and must be tailored to the population.

For assessing the level of a reading passage, the standard is computer-based readability indexes. A *readability index* is a way of measuring the ease of comprehension of a piece of text (McCallum and Peterson, 1982), e.g., Flesch Kincaid Grade Level, Flesch Kincaid Reading Ease, Gunning Fog Index, SMOG Index, Coleman Liau Index, and the Automated Reading Index (Brigo *et al.*, 2015; McLaughlin, 1969; Zhou *et al.*, 2017). These use algorithms based on measures of word

difficulty like average word length and word frequency, sentence length and syntactic consistency, and passage length to make predictions about the reading level of the passage. All these features of the text interact and cannot be viewed in isolation. These calculators are generally reliable for ordering text into levels and predicting the rough difficulty of passages. Nevertheless, because they are not always consistent, leveling usually involves considering multiple indices simultaneously, and these automatically-computed indices may stop providing meaningful results at tenths of grade estimates (Zhou *et al.*, 2017).

Despite the frequent use of readability indices for reporting text difficulty levels in studies across a broad range of domains (Agarwal *et al.*, 2013; Loughran and McDonald, 2014), there is little guidance on precisely matching readers to reading index levels (Olson, 2010), so we recommend consulting with an education expert on questions of appropriate content levels for a given population. For example, for both K-12 students and adult literacy learners, the text they are reading should feel relevant to them, hold their interest, and be a good fit for their literacy proficiency. For that matter, your readers need to know why they are engaged in the activity as they will be more engaged if they do (Knowles, 1970).

4.2 Typographic and Visual Considerations

The following visual attributes have been manipulated since the beginning of personal computing: script, language, category, (classification) typeface, font, glyph, size, color, column width, line spacing, and letter spacing. Each of these choices has an effect on reading, and may additionally vary with the hardware and software used to present the text. Experts, like typographers and designers, can be extremely helpful when thinking about how to visualize the text you want to show readers; they not only understand the theoretical foundations of what makes a typeface function in a given reading situation, but can draw on their skills and training to help make your study better. While we recommend consulting with experts on the specific typographic choices for a given study, this may not always be possible, so we provide a primer on

typographic features and terminology, along with references to further reading.

4.2.1 Understanding Font Classifications

There are several thousand written languages represented by close to a 100 modern scripts, each of which have implications for reading (Kessler and Treiman, 2015). In the Latin script taxonomy, categories like serif, sans-serif, handwriting, blackletter, etc. describe fonts based on their anatomical characteristics. Classifications are then used to identify more specific anatomical features like the shape of the serif, or the angle of stress, or angle of terminal. Fonts also vary on many parameters, such as stroke modulation, letter skeleton, and letter proportions (for a deeper analysis of typeface classification see Bringhurst, 2004; Tracy, 1986).

Uniquely identifying fonts. Referring to fonts solely as "Garamond", "Caslon" or "Bodoni" is not enough information to identify them. Digital fonts based on historical sources exist in multiple versions; for example, the group of Garamond typefaces (Figure 4.2) is a revival of an Old Style serif design by Claude Garamond (16th century), and many versions exist. For the benefit of future researchers, we recommend accurately noting the specific fonts used in a study, referencing the Family – Style and any attributes applied (e.g., CSS property setting, like "ITC Garamond – Regular"), and where possible linking to both files and code used.

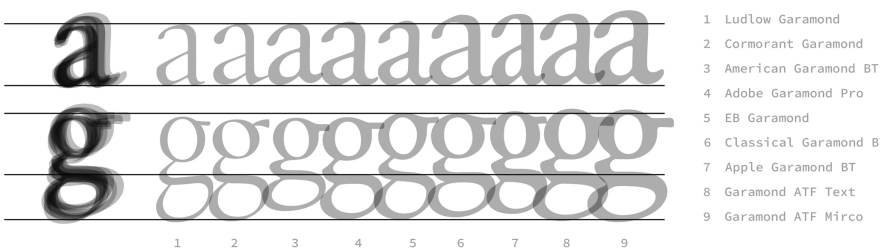

Figure 4.2: *Specificity* in fonts used is extremely important for understanding results and for future replication. For example, a study reporting that the font Garamond was used would be ambiguous, as the examples of Garamond fonts from nine different foundries clearly show. Note how much they vary in the design of the letters. The letters are superimposed at the left to make the amount of variation more evident.

Fonts versus typeface. While the words fonts and typefaces are sometimes used interchangeably, more formally a typeface is the designed set of glyphs, such as "Garamond". A font is a particular incarnation of a typeface, such as "14 point Garamond bold".

4.2.2 Selecting Fonts based on Availability

There are over 600,000+ publicly available fonts. Only a few hundred have been optimized for screen, and only a few dozen of that subset are ubiquitous. For many readability researchers, the availability of test fonts is a practical consideration. Times, Arial, Georgia and Verdana are the most common typefaces, and are often used in studies (Bernard *et al.*, 2001, 2003; Boyarski *et al.*, 1998; Cai *et al.*, 2022; Pušnik *et al.*, 2016; Rello *et al.*, 2016; Sheedy *et al.*, 2005; Wallace *et al.*, 2020a). Sometimes called the *web safe fonts*, they appear on all Apple and Microsoft products and are available to all web browsers. In addition, Google, Adobe, and IBM have also made high-quality typefaces available for free distribution and many can be found at Google Fonts (https://fonts.google.com) and GitHub (https://github.com).

Intentional and unintentional font replacement. All operating systems have lists of font aliases that are used when a typeface is not available. For example, while Helvetica comes preinstalled on Apple's macOS, it is not available by default on Microsoft Windows and is substituted with Arial. To further complicate matters, fonts can be named whatever the user desires, and their Helvetica may not actually be Helvetica. Even more insidious, they could install a different version of the same font with small, but significant, design changes. To ensure that your readers see what you want them to see, we recommend bundling your typefaces with your experiment, or at least testing extensively on different platforms to find out what your readers are being shown.

4.2.3 Controlling Font Properties

No matter the experimental paradigm, the visual appearance of stimuli will always affect the final results. A common approach in readability research is to compare reading performances using fonts of different typeface families. Different font categories (Figure 4.3) vary on multiple

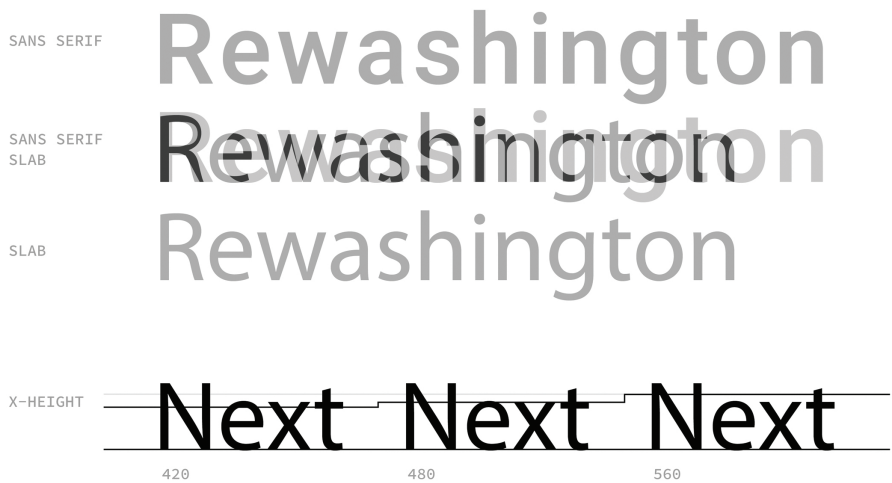

Figure 4.3: (top) Sans Serif and slab fonts from the Roboto family. The two fonts are superimposed to visualize the differences in letter weight, width and serifs. (bottom) Variable font Roboto-Flex with x-height being adjusted to demonstrate the effect on the visual appearance of the font.

properties (Figure 4.4), anatomy, and attributes, each of which may affect reading.

For example, the presence of serifs has been shown to lower reading speed at small sizes compared to the same font without serifs (Morris *et al.*, 2002), yet in other reading situations, serifs can improve

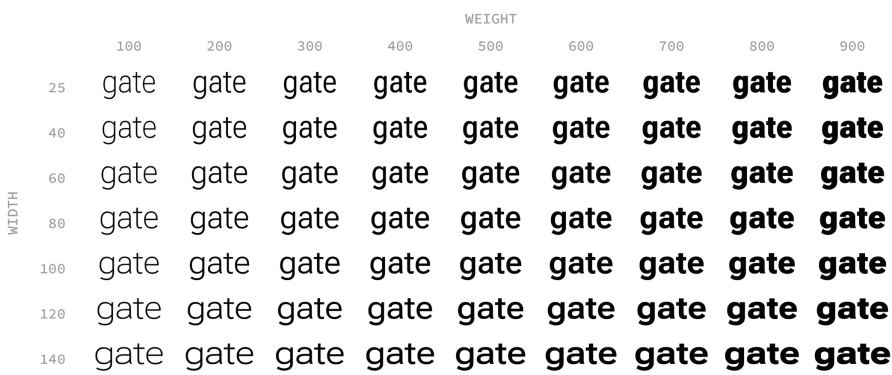

Figure 4.4: A typeface can include many variations of weights and widths, such as this example made with variations of Roboto-Flex, a variable font.

recognition of single letters on vertical extremes at a distance (Beier and Dyson, 2014). Low stroke contrast improves word recognition (Minakata *et al.*, 2020). Simple letter skeletons result in greater letter recognition (Beier *et al.*, 2018; Beier and Larson, 2010). Condensed fonts impair letter recognition (Oderkerk and Beier, 2022), and so do heavy and light letter weight fonts (Beier and Oderkerk, 2019b), which also slow down reading speed (Chung and Bernard, 2018).

Perceptual size matters. Traditionally, many studies have focused on comparing different typefaces such as Arial and Times New Roman and comparing these in the same fixed point size per condition (Bernard *et al.*, 2001; Wallace *et al.*, 2022a). This approach may introduce confounds, as the perceptual size of a font is a function of a multitude of font properties, including its x-height (distance between the baseline and the mean line in a font), glyph width, and ascender/descender lengths (Figure 4.5), rather than point size (Beier, 2012). This has led to efforts to present stimuli fonts at a perceived font size by comparing fonts set at similar x-height (Beier and Oderkerk, 2019b; Wallace *et al.*, 2022b; Yamabe and Takahashi, 2007).

The problem of interacting variables. To identify the effects of specific font properties that can be transferred to other fonts, we need to isolate significant properties. This can be done by comparing fonts belonging to the same typeface family (e.g., width variation between Univers Condensed and Univers Expanded), or designing fonts for experiments where all other possible variables are controlled for

Figure 4.5: A comparison of two fonts of different x-height set at identical font sizes (Helvetica Regular and Adobe Garamond). The different x-heights result in Garamond having longer ascenders and descenders, as well as appearing to have greater leading between the lines of text and having a smaller font size.

Figure 4.6: Decovar, a Google Font, designed by David Berlow, Font Bureau, 2017, demonstrates how different serif structures can co-exist in a single variable font file.

(Beier, 2013; Beier and Oderkerk, 2019b; Chung and Bernard, 2018; Gürtler and Mengelt, 1985).

Variable fonts. The introduction of OpenType 1.8 in 2016 made it possible to have many fonts in a single file (Figure 4.6; *variable fonts* (Hudson, 2016)). This format allows the different properties of the font to vary on single or multiple axes. For example, the weight of a bold font, the width of an expanded font, or the stroke contrast, are not predefined. The user can choose the exact coordinates on many axes. This flexibility can enable researchers to be more in control of the magnitude of each font variable (Figures 4.3–4.5).

4.2.4 Controlling Typographic Settings and Environments

Control typographic settings. In addition to controlling font selection and properties, typographic settings should be controlled. Some of the typographic settings that have shown to influence reading are letter spacing (Perea and Gomez, 2012), word spacing (Slattery and Rayner, 2013), contrast polarity (Dobres *et al.*, 2017b), background complexity (Sawyer *et al.*, 2020), and font color (Ko, 2017). Much of the control of letter and word spacing can be done with cascading style sheets (CSS) (Wallace *et al.*, 2020b) which can be used to manipulate the page layout of the text, and further dimensions can be manipulated through the use of variable fonts.

Consider how fonts may be perceived. In addition, text and presentation familiarity are important (Beier and Larson, 2013). Unfamiliar fonts (Beier and Larson, 2010; Zineddin *et al.*, 2003) or unfamiliar script styles (Ngiam *et al.*, 2018; Pelli *et al.*, 2006) can negatively influence reading. Also, given apparent agreement on perceived font personalities (Grohmann *et al.*, 2013; O'Donovan *et al.*, 2014), text stimuli need to be controlled for semantics, vocabulary and context (see Section 4.1).

Consider the purpose fonts were designed for. It should not be assumed that all fonts are equally appropriate for testing on all reading platforms or all reading situations. Many large-size typeface families include fonts of different optical sizes, where the fonts for smaller sizes typically have larger x-height, low stroke contrast, and greater width and spacing (Ahrens and Mugikura, 2014). Many typefaces were designed and engineered for specific rendering systems (e.g., Microsoft's ClearType fonts (Berry, 2004)). Typefaces can also be designed for how they will be used by content developers (e.g., graphic designers, web designers, app developers, UX/UI developers). Most fonts designed for use on-screen will work well in print, while not all fonts designed for print will work well on screen (Bernard *et al.*, 2001), and some fonts designed for large sizes will not work well in small sizes (Ahrens, 2008). Moreover, new reading formats are constantly emerging. It is difficult to predict if findings from studies using traditional screens will transfer to reading in AR and VR environments (Section 5.1.1).

4.2.5 Considering Resolution and Rendering

Screen optimization of fonts usually includes: large x-height, open *apertures*, large *counter forms*, generous letter spacing, limited stroke contrast and *delta hinting* (Figure 4.7, and Larson, 2007). Other central variables to consider are rendering (Ahrens, 2012), font size, resolution, browser, and operating systems (Boyaci *et al.*, 2009).

Nearly all fonts store their outlines as *Bézier curves* so they can scale to any size without losing fidelity. A notable exception is bitmap fonts, which use images instead of resolution-independent vectors (e.g., emoji).

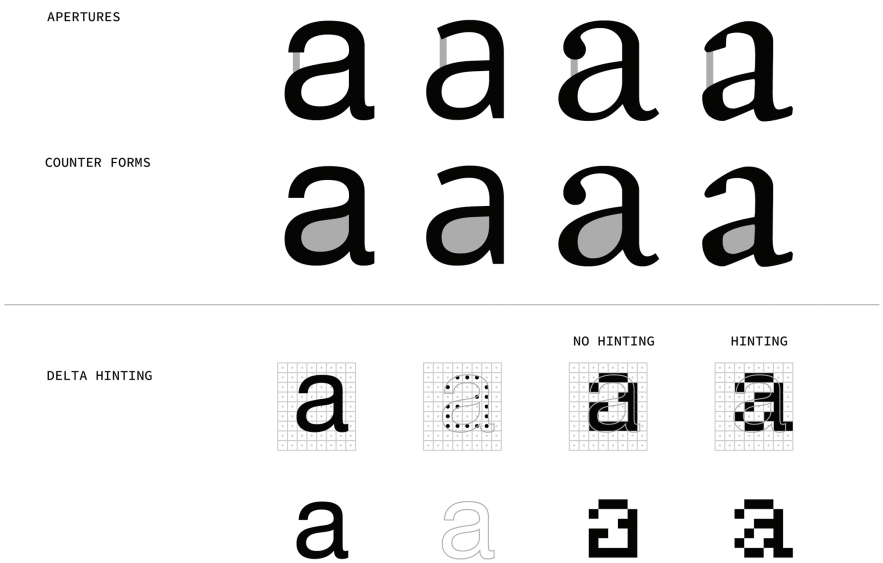

Figure 4.7: Apertures, and counter forms are critical features of a font that can significantly affect their appearance on digital screens.

Rasterizers turn vector fonts into pixels for display on the screen. Most rasterizers are part of the operating system, but there are also standalone rasterizers that can be used on multiple operating systems. There are four rasterizers in common use: GDI and DirectWrite on Windows, Core Graphics on macOS and iOS, and FreeType on Android, Linux, and ChromeOS. Rasterizers turn Bézier curves into pixels by sampling them at the desired resolution. At its most basic, the rasterizer checks whether each pixel is inside or outside of the curve. If this sampling is done at a high enough resolution or a large enough font size, the result is a near-perfect approximation of the curve.

Resolution and font size are linked. High-resolution screens produce good results at low and large font sizes (Gugerty *et al.*, 2004), but legibility suffers when small font sizes are used on a low-resolution screen (Figure 4.8). To address these issues, fonts include hinting instructions, which tell the rasterizer how to behave at low resolutions. Typefaces that include extensive hinting are often advertised as especially geared towards legibility at small sizes. Not all rasterizers support hinting, for

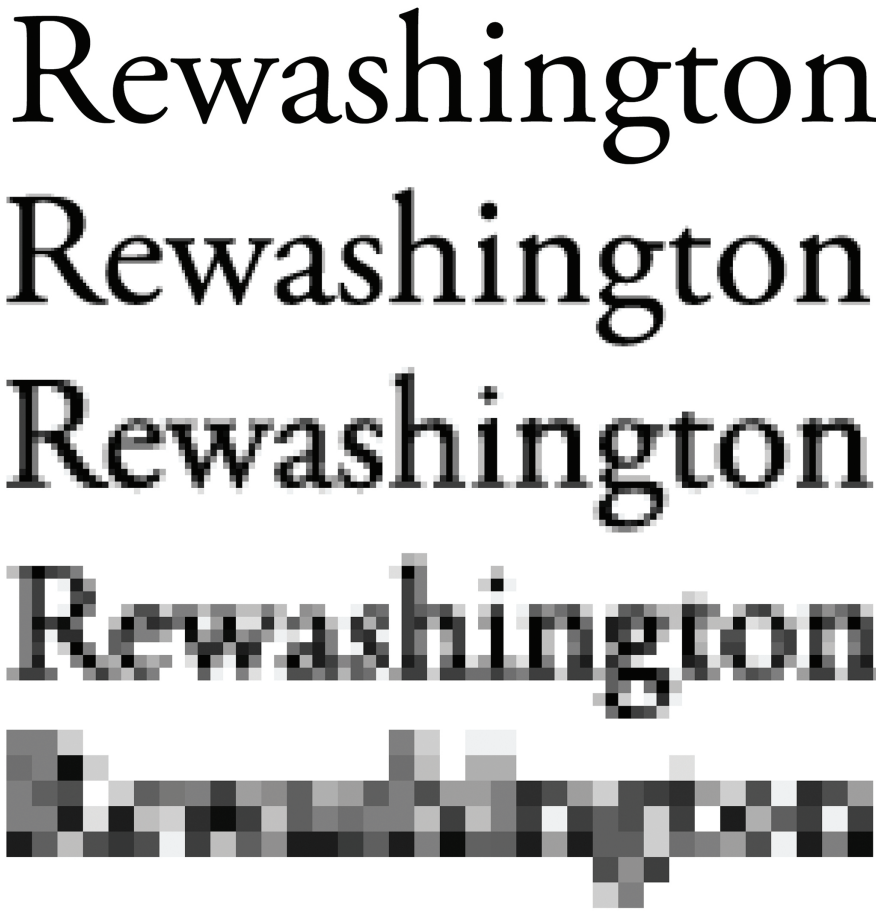

Figure 4.8: Rasterizing at different resolutions and font sizes.

example, Apple's Core Graphics rasterizer ignores them because Apple's devices generally have high-resolution screens which reduce the need for extensive hinting. Further, the impact of size and resolution upon legibility has been shown to be minimal, except at very low values of each (Hancock *et al.*, 2015).

If possible, we suggest using delta hinted typefaces for readability studies. While some rasterizers may ignore the hints, others will benefit from having high-quality hints. There are two types of antialiasing: grayscale and sub-pixel. Grayscale uses grayscale values to approximate

partially filled pixels, while subpixel antialiasing uses a screen's red, green, and blue sub-pixels to achieve the same. While sub-pixel rendering may sound superior, it is often disabled because subpixel antialiased text cannot be rotated, since the RGB subpixels are fixed, and can produce distracting color fringing on low-resolution screens.

4.2.6 Other Typography Related Matters

When presenting text, the experimenter must make decisions about the fonts, sizes, weights, colors, spacings, and other visual features as well as whether they are normalized across different fonts and type settings. As readers' ability to make a correct return-sweep over the text with the eyes is dictated by how easy it is to identify the beginning of the following line of text (Parker *et al.*, 2019), some additional inter-line spacing may ease the return-sweep and allow for a wider column width (Ling and Schaik, 2007; Tinker, 1963). It has further been shown that low luminance contrast between text and background color can impair reading, and lead to greater difficulties in searching for a target word (Ko, 2017), while bolder font weights under low luminance contrast conditions result in faster search time compared to lighter font weights (Burmistrov *et al.*, 2016). If there are complex background textures, reading benefits from fonts with more simple letter skeletons (Pelli *et al.*, 2006). Dark text on a light background is easier to read than light text on a dark background (Dobres *et al.*, 2017a), optimal letter spacing depends on the balance between distance and font sizes (visual angle) (Tejero *et al.*, 2018), and fluent reading is possible between 4 point and 40 points sizes, yet the size threshold for which people are able to read varies greatly between participants of varying abilities (Beier and Oderkerk, 2019a) (refer also to Section 3). If such factors are not part of the research questions, it is essential to also keep these variables constant between test conditions (Tejero *et al.*, 2018).

4.2.7 Perceptual Considerations

The human visual system has its own limitations which impact readability, interacting with the typographic and visual properties discussed above. Most notably, our sensitivity to information at specific spatial

frequencies (Legge *et al.*, 1987) and contrasts (Majaj *et al.*, 2002) is described by the contrast sensitivity function, and making changes outside of human perceptual space will have no impact on readability. It follows that different reading situations set different demands on the choice of font and layout. For example, due to effects of letter crowding (where neighboring letters tend to merge at small visual angles in central vision (Coates *et al.*, 2018)), text in small font sizes and text at great reading distances will profit from greater inter-letter spacing than text read up close at larger sizes (Highsmith, 2020). Further, paragraph reading where a large visual span is beneficial, might demand a different set of considerations than text of few words, where the peripheral vision is of less importance (Beier, 2017). An extensive review of the fundamental perceptual and psychological processes underlying reading can be found in Legge's *Psychophysics of Reading in Normal and Low Vision* (Legge, 2007), recommended as an introduction to the perceptual side of readability.

Another key element of the perceptual mechanics of reading is the question of saccades, fixations and return sweeps – the physical processes by which we move our eyes through a text. Key concepts here include visual span (Rayner, 1975), the number of characters the reader can recognize and read on a given fixation, and parafoveal preview (Blanchard *et al.*, 1989), the reader's ability to glean information from the next point in the text that they will fixate. As with the perceptual topics just mentioned, these are large topics and researchers looking to study gaze behavior in reading and readability are likely to find Rayner's *Psychology of Reading* (Rayner, 2012) a useful entry-point.

The psychophysical literature emphasizes just how much individuals vary, and while the sample sizes in many studies are small by HCI and applied research standards, they should not be discounted, as what they lack in participants they make up in trials (Section 3.2). Since all of readability, particularly our very visual focus in this monograph, is based on this perceptual foundation, you should assume that it interacts profoundly with the typographic considerations described earlier.

4.3 Licensing

To be able to use content in reading studies, the source material needs to be appropriately licensed. This is particularly relevant to researchers in industry who may use the results of reading studies to inform commercial applications and future product development. To reuse content for research, it is advisable to consult with your university's or organization's legal counsel to determine if your research meets the standards of the fair use exemption, or if you need to license the text. Some useful texts may also be in the public domain or available under Creative Commons licenses (see Appendix C).

Content creators may also be open to having their content, whether full texts, excerpts, or fonts, used for research purposes. Getting permissions that allow for research while protecting intellectual property and creating a mutual understanding of how the results will be shared is critical. Similar constraints exist for fonts, as their End User License Agreements vary significantly in how you are permitted to use and alter them. If you are interested in altering an existing font, it is likely easier to use open-source fonts which permit modification.

We advocate that all tools and results are shared with the larger community. In particular, we recommend that raw data is made publicly available to enable new analyses and investigations in the future. We would go so far as to argue that the lack of both data from prior studies and, critically, the materials required to run readability studies, is a significant impediment to research at the moment. In particular, developing and sharing properly-leveled reading materials (passages and support materials, see Appendix C) for different populations of readers will facilitate readability research in the future. As a researcher, it is vital to understand your own ability to enable future researchers by permissively licensing and making publicly available your tools, content, and datasets.

5

Equipment, Devices, and Software Tools

People read using many different interfaces and in many different contexts, from glancing at notifications to scrutinizing news articles (Macfadyen, 2011; Margolin *et al.*, 2013). In this section, we discuss experimental set-ups for studying readability, from brain imaging and eye tracking devices in the lab, to web-based experiments (Figure 5.1). The appropriate hardware and software for a reading study depend on the context and environment, the target reader populations, the specific research questions, and the availability of resources, and this section is meant as an introduction to the possibilities that exist.

5.1 Digital Displays

Reading has changed the widespread adoption of the digital display in the 1970s. Instead of text only being read on the printed page, reading is now done on a wide variety of display types: large desktop monitors, smartphones, purpose-built displays, e-ink devices, smart watches, and (less frequently, for now) in immersive displays. Early studies of digital text legibility suggested that it was inferior to traditional printed text (Mills and Weldon, 1987). More recent work suggests that as the resolution and fidelity of displays has improved, they have achieved parity

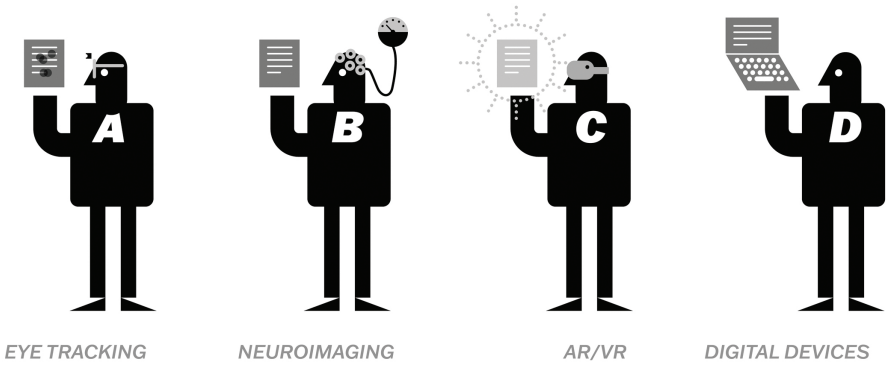

Figure 5.1: Fundamental tools for understanding readability have changed dramatically, with the promise of more change to come.

with print in terms of pure legibility (Margolin *et al.*, 2013), though readers may be able to maintain better awareness of their performance with print (Clinton, 2019).

LCD displays versus e-ink. One key difference between print and digital displays is that print (and e-ink) reflect light, while digital displays emit light. E-ink displays (e.g., like Amazon's Kindle) have no backlight. Research comparing them has been mixed, and suggests that print/e-ink and digital LCD displays have equivalent practical legibility (Lee *et al.*, 2008; Siegenthaler *et al.*, 2011, 2012). Differences in legibility between display types may in fact have more to do with the amount of illumination, both in the environment (Dobres *et al.*, 2017a; Lee *et al.*, 2008) and in the amount of light being emitted by the screen (Dobres *et al.*, 2016). Research suggests that lower illumination settings cause the pupil to dilate over the imperfect surface of the eye, exacerbating the effects of astigmatism and smaller flaws in the lens, hindering legibility (Piepenbrock *et al.*, 2014; Taptagaporn and Saito, 1990). These findings present a particular problem for popular "dark mode" designs, which are self-reported to reduce eye strain (Eisfeld and Kristallovich, 2020), but may have reduced legibility.

Capturing reader behaviors on digital displays. General-purpose digital displays (Yeykelis *et al.*, 2014) can also reveal how readers are moving through a piece of text through incidental movements. Such movements can be captured without impinging on natural reading

behavior. These can include a reader's click-stream as they move through a document, where they position their mouse (Cooke, 2006; Huang and Liang, 2015), and how they scroll through a document (Fitchett and Cockburn, 2009), including multi-touch behaviors on phones and tablets (Gooding *et al.*, 2021; Srivastava *et al.*, 2021). Screenshot software can be employed to determine what participants have on screen (Brinberg *et al.*, 2022; Reeves *et al.*, 2019) but incidental data can also include gyroscope data from modern smartphones (Pires *et al.*, 2018) which can reveal orientation and device movement (e.g., walking) (Barnard *et al.*, 2007; Mustonen *et al.*, 2004). Finally, audio recordings, which can be supported by any device with audio input, can be used to approximate reading activity through read-aloud protocols.

Taking viewing distance into account. For readability, text display parameters, like angular size (visual angle) are important design considerations as they interact with perceived size, crowding, and visual span. Viewing distance and viewing angle also differ by the type of digital display technology. In the case of VR displays, Google introduced a unit for perceived size called "distance-independent millimeter" (dmm), where 1 dmm equals 1 mm height at a 1 m viewing distance, enabling distance-independent designs in VR.

5.1.1 Immersive Displays

Virtual Reality (VR) and Augmented Reality (AR) – often referred to as Mixed Reality (MR or XR) – are another tool and setting for readability studies. In VR, readers are fully immersed into an environment which can include all sensory modalities (Brooks *et al.*, 2020). VR simulations are often used for training and educational purposes as users adopt similar behaviors in VR as they do in the real world. Studies in VR allow researchers to put participants in different real-world scenarios and investigate different in-field environments at scale (Mäkelä *et al.*, 2020). Platforms such as the HoloLens or Vive Pro headsets (Microsoft HoloLens, 2019; VIVE, 2018) have higher fidelity head and eye tracking than mobile devices and allow for 3D interaction with content.

Opportunities for immersive reading. AR applications are designed to enrich users' physical activities with visually overlaid information. This paradigm makes digital text powerful by tying it to context of on-going activities—for example, remote assistance and responsive instructions (Wisotzky *et al.*, 2019) in industrial training or gaming experiences (Kim *et al.*, 2019; Niantic, 2021; Ružický *et al.*, 2020). Other work has started to explore text renderings on 3D objects where text is warped across concave or convex surfaces (Wei *et al.*, 2020) and text interaction in virtual environments (Dingler *et al.*, 2020). Virtual environments have the potential to immerse the reader in multimodal reading experiences where the visual, audio, and haptic environment adjusts to the content. A challenge is that reading happens as users engage in other activities. It can be difficult to ensure users see the text when mental load is high (Lindlbauer *et al.*, 2019), and continuously changing surroundings such as background textures can cause legibility issues (Gabbard *et al.*, 2006). Laboratory studies offer a controlled way to guide participants through different reading conditions in order to determine readability parameters for mixed reality text rendering.

Challenges for rendering text in AR/VR. Reading in mixed-reality environments is becoming more prevalent with advances in display technology that allow high-quality text rendering, although fundamental limitations mean these capabilities lag behind many other display types. Beyond resolution and rendering, VR and AR platforms introduce readability challenges when presenting text in simulated 3D environments or when superimposed over the ambient environment in AR settings. The placement of text with AR can be a safety consideration, and early work showed that users preferred consistent placement (Orlosky *et al.*, 2013), although this depends on the task. When focused on reading, central positions were preferred, and when walking, a bottom-center position was preferred (Rzayev *et al.*, 2018). In immersive displays, resolution can limit readability and magnification and floating text lead to favorable experiences for users (Knaack *et al.*, 2019). Other research has explored optimal readability settings for font and distance (Büttner *et al.*, 2020), as well as text size and positioning (Dingler *et al.*, 2018).

5.2 Research Equipment

5.2.1 Eye Tracking

Since reading requires a reader to move their eyes from word to word along a line of text (e.g., to make saccades from one word to the next), *eye tracking* (Table 5.1) has been a key technique in reading research since it was first developed more than a century ago (see Huey, 1908 which provides the first English translation of Javal's 1879 work; Rayner, 1998; Tinker, 1946). Tracking where a reader looks whilst they read can reveal what words in a sentence they skip, whether they backtrack, and how they move through a passage–and, potentially, show what visual strategies they are adopting based on the type of reading (see Section 2). That being said, there are significant limits on what gaze information can tell researchers, since fixating a word is no guarantee it was read or understood (Drew *et al.*, 2013), and deducing what a reader was doing based only on where they looked is difficult since it requires knowing how to classify different types of reading based on gaze behavior, what the reader's task was, and whether that task is appropriate to the text they were reading (Ahlström *et al.*, 2021; Wolfe *et al.*, 2020).

Hardware-based eye tracking. A range of eye-tracking equipment exists, typically in the form of non-intrusive hardware that uses near-infrared light to create reflections on the eye and using a camera pointed at the participant to infer eye position, orientation, and movement from these reflections (Hammoud, 2008). This specialized equipment comes in two main forms: head-mounted systems (Cognolato *et al.*, 2018) and remote systems (Niehorster *et al.*, 2018). Head-mounted systems structurally resemble eye glasses and are preferred in naturalistic studies that involve a lot of movement, but have limited spatial and temporal resolution, and many are not adequate for readability studies (Hendrickson and Ailawadi, 2014). Remote eyetracking systems have a stationary base, with the eye tracker mounted near or integrated in a display. These eye trackers are capable of higher speed and accuracy compared to head-mounted systems. These characteristics can be further augmented when the eye tracker is coupled with head stabilization (chin rests) which keep the participant at a constant distance from the

Table 5.1: Advantages and disadvantages of eye-tracking system types

Eye Tracking System	Typical Environments	Advantages	Disadvantages
Head-mounted	Out of the lab studies which can include a lot of movement (e.g., sports, driving, marketing).	Allow for free movement of the participant.	Lower accuracy, precision, and volume of data compared to remote eye tracking. More difficult to map the gaze locations on text and perform statistical analysis.
Remote	Lab studies that aim for highly-controlled experiments.	High volume of data compared to head-mounted eye tracking. Data is directly mapped to the screen containing the reading task which simplifies subsequent analysis.	Little to no movement is allowed to acquire data which can lead to unnaturalistic behavior.
Web Cam	Experiments deployed with remote crowdworkers.	Can be run on many participants, in parallel, to produce a large volume of data,from potentially diverse participants in their naturalistic environments.	Lack of control over data quality, noise, and confounding environmental factors.

screen. Beyond desktops, laptops, and mobile devices, eye tracking has also been embedded in virtual and augmented reality devices. Some popular eye tracking manufacturers include SR-Research, Pupil Labs, and Tobii which offer a range of research- and consumer-grade systems.

Camera-based eye tracking. Recently, software solutions that use standard webcams have begun to be developed as an answer to the high cost of more conventional eye-tracking equipment. However, there are compromises to this approach, as they are severely limited in spatial and temporal precision, and may not be suitable, as of this writing, for readability studies. However, they represent an area with high potential for future studies, pending the development of better cameras and algorithms. There are browser-based (e.g., Papoutsaki

et al., 2017), desktop (e.g., Wisiecka *et al.*, 2022; Zhang *et al.*, 2019), or mobile applications (e.g., Krafka *et al.*, 2016; Valliappan *et al.*, 2020).

5.2.2 Neuroimaging

Neuroimaging research can inform our understanding of which brain regions and networks are active during reading, as well as underlying processes. Since reading is a visuo-cognitive process, non-invasive neuroimaging techniques like electroencephalography (EEG) and functional Magnetic Resonance Imaging (fMRI) have the potential to reveal internal cognitive and linguistic processes that are otherwise inaccessible to researchers.

EEG systems. Electroencephalography is used to measure electrical activity in the brain using non-invasive electrodes placed on participant's scalp while the participant is performing a cognitive or linguistic task of interest. Choosing an appropriate EEG system depends on the population being studied and the goals of the experiment. For instance, the number of electrodes in an EEG system varies greatly from a handful of electrodes up to 256. Systems with more electrodes will naturally require longer and more extensive setup, but will provide better localization of where activity is occurring in the brain. Some experiments may benefit from using a mobile EEG system, which allows participants greater freedom of movement when compared to a traditional EEG system, at the cost of a coarser-grained and noisier signal. A challenge when using EEG for reading studies is the noise introduced by eye movements. Methods such as independent component analysis (ICA) allow for researchers to identify and remove eye blinks from the signal (Jung *et al.*, 2000). Additionally, some researchers are combining EEG and ICA with eye tracking to better identify the relevant signal (Dimigen *et al.*, 2011; Plöchl *et al.*, 2012).

ERP components. Event-related brain potentials (ERPs) are waveforms extracted from EEG, and are believed to be generated from the summed activity of specific cortical neurons (Peterson *et al.*, 1995). ERPs have excellent temporal resolution and are prime candidates for investigating the time course of multiple rapid processes underlying

reading comprehension. Specific ERP components are indicative of different types of processing. Some ERP components frequently examined in reading research include the N250, which likely reflects form-based processing (Holcomb and Grainger, 2007), the N400 which reflects semantic processing (Kutas and Federmeier, 2011; Kutas and Hillyard, 1980), and the P600 which reflects syntactic processing (Osterhout and Holcomb, 1992). Researchers can compare the effect of various manipulations on the latency, amplitude, and scalp distribution of the ERPs of interest. They can also compare ERPs across different populations.

MRI and fMRI. Structural Magnetic Resonance Imaging (MRI) generates high resolution images of the brain, with the ability to distinguish between different types of tissues and brain structures. In addition to acquiring structural information, MRI can be used to investigate functional activity in the brain using methods like functional Magnetic Resonance Imaging (fMRI). When a particular area of the brain is engaged by a task of interest the blood becomes more oxygenated in that region as neural activity increases. fMRI is sensitive to the blood oxygen level dependent (BOLD) signal as a marker of brain regions that are more activated during a task.

fMRI has been used to investigate the processes underlying reading comprehension within specific brain regions, such as the visual word form area (VWFA), important for decoding written words (Cohen *et al.*, 2002; Dehaene and Cohen, 2011). Researchers can also investigate brain networks involved in tasks like reading. This is accomplished by examining functional connectivity, which is defined as the coactivation of multiple brain regions during a task. For example, research suggests that children with developmental dyslexia have disrupted functional connectivity between left occipitotemporal, left inferior frontal, and left inferior parietal, regions that are important for reading comprehension (van der Mark *et al.*, 2011). fMRI can also be used to investigate the contribution of specific regions and brain networks in specific populations of interest such as young developing readers, or dual language learners. For instance, Gaillard and colleagues (Gaillard *et al.*, 2003) found that the reading network in young developing readers is very similar to the reading network in adults.

Currently, fMRI displays use low resolution displays compared to other methods (e.g., eye-tracking, EEG), because traditional display methods are disrupted by the magnet. The low resolution displays can limit the typography features that can be compared in an fMRI experiment. It is difficult to compare subtle typographic differences, but may be used to compare more obvious differences, like font size, or to compare reading across different groups of participants.

Comparing EEG and MRI. Electrophysiological methods like EEG have excellent temporal resolution, which is an advantage when studying reading, where many processes occur in quick succession. However, EEG lacks sufficient spatial resolution which means the specific brain areas involved in various processing steps cannot be inferred from EEG alone. Methods like fMRI have relatively poor temporal resolution, but have excellent spatial resolution. Recent work in cognitive neuroscience shows promise for "fusing" the temporal resolution of EEG with the spatial resolution of fMRI via analysis techniques, for a deeper understanding of brain processes (Cichy and Oliva, 2020). Unfortunately, high quality neuroimaging systems and data analysis are expensive and require specialized training to use, calling out the need to collaborate with neuroimaging specialists for studies that require them.

5.3 Software Tools

Readability studies rely on the ability to manipulate the visual appearance of text to readers to evaluate the impact of these changes on readers' ability to decode the document. Here we discuss platforms that allow for manipulating text formats in this way, including existing commercial tools and new research platforms.

5.3.1 Commercial Tools

Reading on digital devices, whether those devices are laptops, smartphones, tablets, or dedicated e-readers, brings with it a new set of possibilities for manipulating the visual appearance of text. Readers using these devices for consuming documents, webpages, or e-books, are increasingly able to change the font, text size, character and line

spacing, background color, and more to suit their individual needs and preferences. Adobe Acrobat Reader with Liquid Mode, Amazon Kindle Reader, Apple iBooks, and Microsoft Immersive Reader are all examples of reading applications with a subset of these features (available at the time of this writing) summarized in Table 5.2. These applications can serve as platforms for research on the effects of different formatting interventions on reading performance.

Instead of being limited by existing tools, researchers can create customized reading materials by varying font features (like type, size, character and line spacing). Options include using Microsoft's Office Suite or similar document editors, working with variable fonts on support platforms (https://v-fonts.com/support), and using design software like InDesign.

5.3.2 Research Platforms

The Virtual Readability Lab, or VRL (https://readabilitylab.xyz/), is a new web platform containing several essential building blocks to engage users interested in self-paced reading studies. The VRL contains smaller 5-minute versions of *interlude reading* tests (Section 2.1) measuring reading speed and font preference (Wallace *et al.*, 2022a,b) and tests for users to find their optimal character spacing. The VRL allows other researchers to develop additional tests by using a unified database and building on current and future modules. The VRL also contains functionality to allow for teachers to enroll their students and download their progress as each student takes various tests on the VRL to find out which font optimizations can improve their reading. The VRL relies on the voluntary participation of users by providing them insights about different ways to improve their reading behaviors, and it allows for users to compare themselves to the general population.

Readability Matters has developed and made available the open-source Readability Sandbox (https://readabilitymatters.org/readabilitysandbox/). The Sandbox uses variable fonts to allow users to explore standard readability features such as font, font size, character spacing, character width, font-weight, line spacing, column width, and background

color. Researchers can use https://github.com/ReadabilityMatters/TuneYourText for testing purposes.

Table 5.2: Typographical manipulations available by reading application

Reading Application		Character									
		Fonts	Size	Spacing	Weight	Width	Line Spacing	Color	Justify	Page Width/ Columns	Notes*
Adobe Acrobat Reader	1		✓	✓			✓				✓
Amazon Kindle App	2	✓	✓				✓	✓	✓	✓	✓
Apple Books		✓	✓					✓			✓
Google Play Books Reader	3	✓	✓				✓	✓	✓		✓
Microsoft Immersive Reader Office/OneNote	4	✓	✓	✓			✓				✓
Browser Reading											
Apple Reader Mode, Safari		✓	✓					✓			
Google Play Books		✓	✓				✓		✓	✓	
Google Reader Mode	5	✓	✓					✓			
Microsoft Immersive Reader Extension	6	✓	✓	✓			✓				✓
Mozilla Firefox Reader View	7	✓	✓				✓		✓	✓	
Overdrive Reader	8	✓	✓		✓		✓	✓	✓	✓	
Reader Mode	9	✓	✓	✓			✓				✓

6

Experimental Methodologies

This section provides broad methodological guidance for readability research, borrowing from user experience, human factors, and psychophysics. This section is not meant to be exhaustive, but represents our perception of "core" methodologies in readability research. However, these are only options, and future work can and should build on them (Figure 6.1).

6.1 Metrics

Readability interventions must be measured through metrics which gauge reader efficacy. Efficacy metrics should not be confused with *readability formulas* (see Crossley *et al.*, 2011), predictive tools which exist to predict content readability before reading, relative to the level of the reader. Such predictive metrics are helpful in attaching material to a grade level but are inherently unable to relate experimental manipulations to changes in reading efficacy. The metrics we present here all focus on efficacy, and in various ways measure readability by measuring factors which indicate the reader's success.

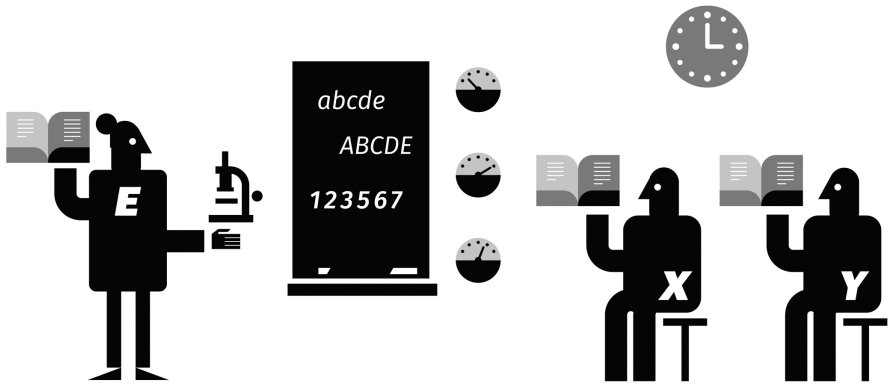

Figure 6.1: Readability literature contains great diversity in terms of experimental metrics and methods.

6.1.1 Reading Speed

Reading speed, often measured in words per minute (WPM) is calculated as the number of words read divided by the number of minutes taken to read them, and is a standard metric of readability assessment. Different researchers have found different ways to identify "words", including absolute word count, number of characters per "standard word", other aggregate numbers across paragraphs and pages, and still more schemes intended to smooth out the relative differences between written passages (Legge, 2007). WPM is a standardized measure in English, and while differences between languages and forms of writing make comparing across languages with WPM difficult, it is certainly possible. For example, symbolic and alphabetic languages have been successfully compared in the literature (Fraser, 2007; Gooding *et al.*, 2021).

Mechanically, the speed of reading is a function of moving the eyes across the page in a series of jerking movements (*saccades*) and longer motions from the end of each line to the beginning of the next (*return sweeps*). As a reader makes saccades, the distance between stopping points is called *saccade length*. The *visual span* or *perceptual span* further refers to the angular span within which a reader has sharp enough vision to perceive words and letters accurately. A larger span is associated

with faster reading as more letters or words can be recognized during a fixation. Of course, when the information on the page is not received by the reader, the reader must either forgo that information or read again. We have all experienced stopping mid-page to discover we have no memory of the content, even though our eyes have mechanically swept across every line. Indeed, speed in reading must be considered together with the accuracy of acquiring information, or reading comprehension.

6.1.2 Reading Comprehension

Reading comprehension is another standard metric, measured by probing participants' understanding of what they have read. The most common method for assessing reading comprehension is comprehension questions testing knowledge of what has been read. Most studies favor questions delivered shortly after reading, although naturalistic scenarios would seem to favor assessing comprehension further from the time of reading. Computing comprehension often takes the form of a percentage, where the number of correct comprehension questions are divided by the total number of comprehension questions.

The present lack of common and consistently used collections of passages and questions in readability research means that comparing between studies can be challenging as it is difficult to know whether the measurement instruments are comparable. Reading comprehension is presently a standard metric which lacks a standard instrument of measurement.

Comprehension questions can be designed to tap a variety of comprehension strategies. Recall questions require readers to directly recall specific information from the text. Inference questions require readers to connect the information to fully understand the text. For example, questions can probe within-text inferencing abilities by requiring readers to connect information from multiple parts of the text. Summarizing questions require readers to combine main ideas presented in the text. Questions regarding the main idea or purpose of the passage can rely primarily on recall and recognition or, in more complicated texts, will require readers to synthesize information across the text and infer the main idea(s). Readers with strong inferencing skills are better able to

fully conceptualize the text's purpose and meaning. A related concern is the background knowledge of the participant, which can be addressed through surveys asking for reader familiarity or by removing questions that pilot participants indicate can be answered without reading the passage (Johnston, 1984).

6.1.3 Speed and Comprehension Together

Speed in reading is joined by comprehension, defined as accuracy in acquiring information and understanding. Indeed, to certain populations comprehension is the more meaning bearing metric. For example, young readers are far more likely to be assessed for reading comprehension than speed. Notwithstanding, these metrics are linked. The upper bound of speed at which readers can move their eyes from word to word will certainly have a negative impact on the ability of that reader to comprehend the material. Therefore, the aggregate effectiveness of a reader depends upon a speed-comprehension trade-off, likely with some similarities to classic speed-accuracy tradeoffs (McElree *et al.*, 2006; Reed, 1973).

Speed-comprehension trade-offs in naturalistic reading are not presently well understood and appear not to be a simple exchange but one contextually sensitive to, at the very least, reading purpose, material, and reader skill (see Section 2). A less skilled reader may move more slowly through a passage than a skilled reader. However, comparing reading speed between participants across passages can be difficult, since the participant and the passage may both be sources of variance. To complicate matters further, slower reading can signal deeper engagement by a skilled reader. A reader's adjustments to their reading speed to compensate for the difficulty of the comprehension questions also represents this speed-comprehension tradeoff. When designing a study to measure reading speed, it is essential to counterbalance the order of passages and participants so that all passages are read the same number of times by all participants, and if the study involves varying typographical settings of stimuli, counterbalance this as well. That way, any possible difference found will relate to the difference of stimuli and not differences in the passages or order.

Fundamentally, reading comprehension is a more complex and subjective measurement than speed, requiring consideration of the reader's ability to retrieve, use, and integrate the phonological, morphosynthetic, semantic, and orthographic aspects of reading, as well as considering their ability to recall background knowledge and synthesize it with the text (Alexander *et al.*, 1994; Elbro and Buch-Iversen, 2013). All of this complexity is filtered through necessary consideration of the cognitive processes involved in reading, starting with the transformation of the visual information presented onscreen, passing through poorly understood intermediary processes, and ending in equally poorly understood mental representations. The challenge in clearly elucidating the measurement of reading comprehension, and its role in any trade-off, is considerable.

6.1.4 Oral Fluency

For younger students, oral reading, or reading aloud, is commonly used by elementary school teachers to assess reading behaviors (Fuchs *et al.*, 2001). Measurement tools, such as Running Record and Qualitative Reading Inventory (QRI), evaluate oral reading fluency, including speed, accuracy, and prosody. Oral reading fluency, the speed at which accurate reading occurs, is expressed in Words Correct per Minute (WCPM), the number of words spoken correctly relative to their written form divided by the number of minutes taken. Prosody is a more subjective measurement of expressive reading that measures appropriate timing, phrasing, emphasis, and intonation (Idsardi, 1992).

Reading aloud can also provide valuable metrics for work with teen and adult readers. Oral reading can reveal reading format sensitivities (Fuchs *et al.*, 2001; Rasinski *et al.*, 2005, 2017) and difficulties in reading aloud can be a result of deficits in the visuo-cognitive linkages necessary for fluent reading. Comparing this with readers reading silently can also reveal where gaps exist in a reader's skillset.

6.1.5 Phrase, Word, and Letter Identification

Some research focuses upon very short phrases or single words or individual letters. While reading at-a-glance is something naturalistically performed on electronic devices, some of these tasks have no applied

equivalent. These methods instead probe sentence- and word-level processing, allowing researchers to carefully control stimuli in terms of the words themselves, varying factors such as word length, age of acquisition, or number of syllables. These methods lend themselves to manipulations involving the presentation of each word with regards to orthographic and typographical characteristics of a text, as well as syntactic structure in the case of sentences. In word-level *semantic categorization tasks*, participants are asked to view single words and make a semantic decision about each word (e.g., "is it alive or not?"). *Lexical decision tasks* may also be used (i.e., "is this a real word?"), but semantic categorization tasks ensure participants comprehend stimuli to successfully complete the task. At the sentence level researchers often study sentence structure to see how syntax affects comprehension (Brothers and Traxler, 2016; Brown *et al.*, 2012; Sorenson Duncan *et al.*, 2020).

Comprehension of individual letters and words via orthographic processing is complex, and must be understood in concert with integrating that information with syntactic and contextual information. Scientists have debated whether letter identification occurs primarily via a template-matching versus a feature-based paradigm, but most researchers now support a feature-based approach (Grainger *et al.*, 2008). Thus, letter identification occurs primarily through the identification of individual features, such as horizontal lines, curves (e.g., open up versus open down), and terminations. The set of features that are most important differ depending on the specific letter (Fiset *et al.*, 2009). One measure of letter identification involves presenting participants with single letters or letters flanked by one of two other letters to the left and right. Often, the aim of such experiments is to investigate limitations of the perceptual system relating to visual acuity, visual angle, or physical size of the stimuli (Hancock *et al.*, 2015) and visual crowding, a phenomenon of neighboring letters seeming to merge perceptually, resulting in misidentification (Beier *et al.*, 2018; Bouma, 1970; Chung and Bernard, 2018).

Different models exist to explain the process of word recognition (for example, see Davis, 2010; Davis and Bowers, 2004; McClelland and Rumelhart, 1981; Whitney, 2001). In general, word recognition involves the activation of relatively flexible letter position coding. For

example, some models propose that a letter in a specific position (e.g., "o" is in the second position of the word "goat") will activate the node representing a letter in that specific position as well as other nearby positions (e.g., also the third position and to a weaker degree the fourth position, etc.) (Davis and Bowers, 2004), whereas some other models propose that within-word letter pairs are activated (e.g., the letter pairs "go", "oa", "gt" will be activated for the word "goat") (Grainger *et al.*, 2004; Snell *et al.*, 2018). This will in turn activate lexical representations of other words with similar letters. Next, whole word representations are mapped onto semantic information in the lexicon (Holcomb and Grainger, 2007).

Word recognition research can use a variety of methods to understand the cognitive processes that subserve word recognition. For example, in lexical decision tasks, participants are presented with real words and either pseudowords or nonwords one at a time and are asked to indicate whether the stimulus in each trial is a word or not. Pseudowords are strings of letters that do not form a word but follow orthographic and phonological rules of the language so they are pronounceable (e.g., "pable"). Nonwords are strings of letters that do not follow orthographic and phonological rules of a language and are unpronounceable (e.g., "pbtlk"). Through the use of single, isolated words and pseudowords/nonwords researchers can probe specific questions that are easier to examine in a more tightly bound context compared to a task using word recognition in a sentence context.

Researchers may also use *masked priming* where a prime is presented for a short period of time and is masked by either a forward or a backward mask (often a row of hashtags "####") to ensure the prime isn't consciously perceived. A target is then presented and they are asked to make a decision about the target. Manipulating features of the prime and target allows researchers to investigate the influence of various orthographic and phonological factors on word recognition. These types of experimental tasks can be conducted using behavioral methods where longer reaction times, and potentially lower accuracy, are indicative of more effortful processing.

6.1.6 Visual Search

Reading is not always a linear, sequential task (e.g., reading through a paragraph in order); readers often have to find a given word or phrase or concept in a text, and while this is reading, it represents a very different task than reading a paragraph from start to finish. Drawing from the cognitive psychology literature, this would be considered a ***visual search*** task; that is, looking for a target (for example, a particular word, phrase or even a concept) among many distractors. This question has been the focus of extensive basic research in the study of visual attention (c.f., Treisman and Gelade, 1980), and can be broadly thought of as "how do we find what we are looking for?"

While the breadth of this literature is outside the scope of this work, Guided Search (Wolfe *et al.*, 1989, 2021), which frames our question in terms of the similarities and differences between the target and the distractors, and uses the similarities to guide where the observer attends, is a promising place to start. It is essential to think of search as less "reading" in a more conventional sense, but more of an object identification problem, and it can be influenced by a range of visual factors in presentation (e.g., font, spacing, density, visual crowding), and by cognitive and linguistic factors. However, readers are likely to transition between searching for something specific in a larger text and reading in more depth, and understanding this initial search behavior is key for guiding and helping readers.

6.1.7 Pleasure and Preference

Reading for pleasure is a neglected measure of readability, in a literature more likely to focus on speed and accuracy, and we speculate pleasure may be a principal reason for quite a lot of reading. Reading for pleasure is a primary reason for purchasing e-readers, as opposed to school or work (Pew Internet Center, 2013), although this may be because readers do not find work-related reading straightforward using them (Massimi *et al.*, 2013). Leisure reading requires a simpler set of functionalities, and may be more sensitive to pleasure and preference than reading for work (Hancock *et al.*, 2005). Few evaluations of readability to date explore this dimension (see Agarwal and Meyer, 2009).

Font preference is inherently subjective (Scaltritti *et al.*, 2019), and deriving a user's preference is not easy. There are over 600,000 digital fonts available, and time and attention constraints make the evaluation of even 100 fonts challenging. O'Donovan *et al.* (2014) identified the struggles graphic designers have when selecting their preferred fonts during real-world tasks. Prior reading studies have most commonly used Likert scales to determine participant font preference (Banerjee and Bhattacharyya, 2011; Bernard *et al.*, 2001; Bhatia *et al.*, 2011; Rello *et al.*, 2016; Wang *et al.*, 2020). While Likert-type scales are straightforward, and can be easily averaged across users, when averaging these results they lose their subjective nature (Stevens, 1946). The results can be noisy and inconsistent (Negahban and Chung, 2014) due to a number of factors that are difficult to control for such as visual discomfort (Li *et al.*, 2018a,b).

A promising alternative to Likert scales are pairwise comparisons (Li *et al.*, 2018c), which are more stable because they are not affected by irrelevant alternatives (Ailon, 2008). Pairwise comparisons for a large number of text formats can take longer given the total number of comparisons a participant must complete. This method can suffer from the transitive property where a participant could prefer text format A > B > C > A. Another disadvantage of pairwise comparison is there is currently no accepted hypothesis test available. Recent work by Wallace *et al.* used a double-elimination tournament to eliminate the transitive property and limit the number of comparisons between 16 different font pairings (Wallace *et al.*, 2022a,b). Other algorithmic approaches to this problem often focus on synthetically completing a pairwise matrix (Kou *et al.*, 2016) or other adaptive approaches (Qian *et al.*, 2015).

6.2 Other Methodological Considerations

Readability studies have a number of specific considerations which set them apart from other studies. Here, we attempt to capture some of the most common issues that we feel are particularly applicable to help HCI researchers in planning their own studies.

6.2.1 The Method of Constant Stimuli Versus Thresholding

An enduring feature of large-scale readability research are the large individual differences seen between participants (Wallace *et al.*, 2022a,b). Readability researchers should keep such differences in mind when choosing between two broad methods to measure responses: the method of constant stimuli, or thresholding. The method of constant stimuli dates to the beginnings of experimental psychology (Sanford, 1888; Spearman, 1908). The researcher chooses levels of stimulus parameters based on predefined assumptions. Data from such techniques allow for the estimation of psychophysical functions that map the relationship between stimulus levels and performance. However, data collection is limited by the number of trials that can be tolerably collected in a session (the more stimulus levels tested, the more trials required). Stimulus levels must be well chosen for the intended audience; e.g., a text contrast that is reasonably challenging for a younger participant may be too difficult for an older participant (refer to Sections 3.1 and 4.1).

Researchers may instead choose to employ thresholding or ***staircase methods***. With these methodologies, parameters of the stimulus are adjusted in real-time based on participants' responses, with the goal of converging on a preselected response accuracy level. Staircasing rules (Leek, 2001; Levitt, 1971) can be employed to converge on several different accuracy levels. For example, if the task is made more difficult immediately after a participant's correct response, and made easier by the same amount after an incorrect response, the experiment will eventually converge on a stimulus level representing the participant's 50% accuracy threshold. A threshold performance value can be determined for every participant without "wasting" trials with parameters that are too difficult or trivially easy. Techniques such as QUEST have updated this thresholding procedure with more advanced statistical assumptions, allowing for faster convergence (Watson and Pelli, 1983). However, if the "step" of the staircase (the amount by which stimulus difficulty is adjusted) is poorly chosen or if the staircase is initialized far from threshold values, it may fail to converge on a good threshold estimate. It can also be more difficult to estimate a full psychometric function from threshold data (Treutwein and Strasburger, 1999).

The method of constant stimuli and staircasing are two sides of the same coin. The former holds stimulus parameters constant while measuring changes in performance accuracy; the latter changes stimulus parameters in real-time while holding accuracy constant. Both have their place in the toolkit of legibility research. For an excellent detailed review of such methods, see Klein (2001).

6.2.2 Time on Task, Fatigue, And Vigilance Decrements

We do not read equally well all of the time, and so studies of readability must be sensitive to fluctuations of individual or aggregate ability. Alertness varies throughout the day, and over the course of a task. Fluctuations in alertness affect cognitive performance and impact higher level cognitive capacities, including perception, memory, and executive functions (Kleitman, 1923) and the ability to concentrate over the course of a study will vary. A lack of alertness can manifest itself in repeatedly re-reading sentences, difficulty with comprehension, and visual fatigue. Some tasks are highly demanding and induce fatigue, while other "vigilance tasks" create specific problems for information processing which grow over time. Vigilance effects are characterized by simplicity of stimuli, high rates of evaluation, and low target-present rates; that is, the observer looks at many items trying to find a rare target. These types of tasks are intensely stressful for participants, and can be created inadvertently, so the researcher should be mindful in readability studies to avoid them (Warm *et al.*, 2008). Vigilance effects can be detected with eye tracking (e.g., by recording blink rate, as it increases with fatigue and time on task). Conducting reading sessions at "reasonable hours" during the day, i.e., avoiding the early morning hours and the "post-lunch" dip, is advisable. Even caffeine consumption (much as it fuels a great deal of research) can affect alertness levels as well and should be considered, and can be asked about in pre-study survey questions (see Appendix C).

6.2.3 The Value of Pilot Studies

It is essential to spend time imagining how specific variables and decisions might affect your study and to test your assumptions before

committing to large-scale data collection. Time spent reading, reading positions, and reading passages can affect participants across different study environments. How long does it take for participants to learn the interface and become comfortable? At what point in the study do readers naturally slow down or speed up? In a lab study with an eye-tracker and chinrest, participants are likely to be less comfortable than reading on their phone on the couch.

Study pacing can also be refined through piloting, since time to complete will vary significantly between participants. One participant might read at 100 WPM while another reads at 700 WPM, and if the task asks them to read 7000 words, the first might take over an hour and the second merely ten minutes. Also, readers slow down when reading more difficult passages (e.g., those normed to a higher level). Understanding how and why readers change the speed with which they read is an integral part of building your study and gathering the data you want.

Another vital contribution of pilot studies is to help to determine appropriate compensation. In many studies, participants are paid for their time, and to compensate participants fairly, you need to consider whether this is appropriate for your design, or if you should compensate based on task completion. Compensating for time may disincentivize speed, encouraging readers to go slowly. While an extensive literature on this question exists, standards vary widely between fields and study environments, and should be determined both by piloting and consultation with collaborators.

It can also be worthwhile to probe your participant's subjective experience when piloting new studies. Asking how they felt, whether (or when) they were confused and whether the task and instructions were clear can be very useful in the study development and data analysis process. For that matter, it is hard to underestimate the value, if possible, in running yourself through your own study, as it is an excellent way to find pain points even before you recruit pilot participants. However, you should bear in mind that just because you can do your own task does not mean your participants can.

7

Data Analysis for Readability Studies

Here, we provide an overview of best practices and approaches for handling and analyzing data generated by readability studies, with a focus on data quality management, as well as statistical and machine learning modeling approaches. Examples have been gathered from existing readability research, but as the area is growing, are not an exhaustive list of the approaches which might be useful.

7.1 Data Quality Management

Since many reported effects of format readability on performance have small to medium effect sizes, it is necessary to repeat many trials within individuals, or collect data from many individuals (Section 3.2). Both situations provide plenty of opportunities for data quality issues. Not all participants perform the task with the same level of dedication, and there are many individual differences in reading ability and strategy, which can result in anomalies or outliers. Defining and detecting them is something of an art, and must be tailored depending upon the study design and population. In studies which involve many trials within an individual, manipulation checks for effects of time on task or training effects are key. In studies with a large number of participants, especially

remote studies, you will likely need to screen for participant dishonesty (Peer *et al.*, 2017) and consider uncontrolled settings (Schneegass and Draxler, 2021). Your data quality issues can be mitigated through careful planning, piloting, and in studies with significant data collection time, ongoing data quality assurance.

A common statistical approach to handling outliers is to assume normal distribution of the data and isolate points that fall further than three standard deviations from the mean (Stevens, 2012). These points can then either be filtered out or analyzed separately. You can also leverage anomaly detection methods during data pre-processing. For instance, in the case of reading speed data, outlier removal can be done based on distributions of reading speed. Typical reading speeds for participants over the age of 18 range from 138 to 600 WPM (Carver, 1990) with an average speed for native English speakers at 240 WPM. Participants whose speed falls outside of this range might be distracted or disengaged from the material, and may be removed from the analysis, although reader behavior will be strongly influenced by task demands.

Variability within a particular participant's data poses a significant challenge for analysis, and should create concern for similar patterns across participants in the entire study. When a similar task is repeated by an individual multiple times, the random error associated with the repeated measurement of independent performance factors, such as attention (Buckner *et al.*, 2008; Christoff *et al.*, 2009; Killingsworth and Gilbert, 2010; Raichle *et al.*, 2001), can attenuate the association between independent and dependent variables and result in poor statistical inference (Barnett, 2004), a bias known as regression dilution (Berglund, 2012; Hutcheon *et al.*, 2010). In general, unusually high intra-participant variability may be a sign of design problems, and confounds like unmet training requirements, excessive time on task, uneven population reading ability, and technical failures should be investigated.

7.2 Exploration and Visualization

Readability data is best initially analyzed through Exploratory Data Analysis, which can help determine data quality, assist in numerical analysis, or build hypotheses for further investigation. Exploration must

take into account the nature of readability data, and must respect any *a priori* plan for analysis. While inspection of raw readability data in a tabular format can be revealing, visualizations are helpful for revealing interesting patterns (e.g., see Figures 14 and 15 in Wallace *et al.*, 2022a). Because readability data is often collected over time, across environments, and between devices, we see opportunities to use spatio-temporal illustrations to explain the complex emerging patterns, heatmaps to visualize one-time movement patterns shared across users (Klemm *et al.*, 2014), or Sankey diagrams (Riehmann *et al.*, 2005) to incorporate higher levels of complexity in participants' shared journeys across different stages of the reading process. Other visualizations can also help represent eye and mouse movement data from readers. For example, areas where the reader looked more or clicked can be highlighted (Blignaut, 2010; Burch *et al.*, 2019).

7.3 Statistical Modeling

Provided tools are used appropriately relative to the data collected, we do not argue for the primacy of any statistical approach. The data analyses needed for readability experiments examining the effect of visual manipulations on outcomes such as speed, comprehension, and preference are similar to those used throughout the social sciences. The standard practice for statistical analysis is to start with numerical and graphical techniques for estimating the distribution of the data and determining the best mechanism accordingly. A simple Kolmogorov-Smirnoff test can determine whether readability scores such as reading speed and comprehension are normally distributed. Parametric tests, in the case of normality, and non-parametric tests, otherwise, are often used in the readability research studies (Soleimani *et al.*, 2008; Soleimani and Mohammadi, 2012), and indeed the non-normality of distributions of many metrics may not be cause for concern, so much as cause for use of the appropriate tools.

Individual patterns in readability are of great interest, and have shown promise in delivering insight beyond that provided through group-level analyses. It is worth acknowledging that statistical tools and approaches for individual differences are less agreed upon, especially

those that do not rely upon central tendency and the general linear model (GLM). For researchers asking questions regarding the impact of readability manipulations on individuals, therefore, it is important to support the methodology used somewhat more than might be otherwise necessary. Likewise, reviewers should keep an open mind about methods that may be new to them, and nonetheless appropriate.

Fundamentally, there is no prescriptive analysis approach or statistical tool for readability research. Readability papers we have introduced in this work are most commonly analyzed with multiple generalizations of the GLM. Many studies rely upon multiple analysis of variance (MANOVA) to isolate the impact of independent variables (IVs) upon multiple dependent variables (DVs), often including both reading speed and comprehension (Ball *et al.*, 2021; Gao *et al.*, 2019; Nam *et al.*, 2020; Rello *et al.*, 2016; Sawyer *et al.*, 2020; Wallace *et al.*, 2020a, 2022a). It's not uncommon to augment these larger analyses with smaller "manipulation checks" which rely upon t-tests or simple analysis of variance (ANOVA) to test out assumptions. In determining how manipulations affect populations across continuous variables, it is certainly appropriate to use regression analysis. Indeed, in usability studies where a simple question of "A or B" is of interest, and where multiple DVs are not used, a simple t-test may suffice. Readability, as an inherently multidisciplinary area of inquiry, should ultimately be modelled using the tools most appropriate to your specific study.

7.4 Machine Learning

For research that aims to assist participants in improving readability, it can be useful to evaluate the performance of statistical and machine learning (ML) models that can predict reading outcomes. Consider one question in the literature: given a font style, can we predict a reader's WPM (Cai *et al.*, 2022)? Here, regression models which predict the relationship between input and output variables might be used to predict participants' reading level from their reading speed. This regression question would be valid with statistical and ML approaches alike, and indeed the outcomes of these two approaches might be notably similar. Classification ML allows the prediction of a label for a given set of

input variables, and so in the context of readability might be useful for predicting the "bin" into which such an input set might fall. A simple binary classification might detect whether a participant is skimming, or reading deeply, given their reading speed. Similarly, using ranking ML, a given set of input can predict an ordered set of labels. In more sophisticated learning models, ranking ML can predict the relative ordering of labels by either comparing pairs of inputs at a time, or by comparing the entire set of labels associated to a criterion (Liu, 2007). As an example, consider ranking the fonts for each reader such that their most readable font is ranked first, and least readable is ranked last. Clustering ML groups similar items together, perhaps providing groups of similar readers and identifying populations in need. A full survey of traditional approaches can be found in Xu and Wunsch (2005) with authors often using the classical approach of K-means. More recently, clustering research has focused on metric learning to learn a feature representation where neighboring items are closer together in feature space. Of course, as with statistical tools, ML approaches are best used together to achieve complex goals of prediction.

ML tools in the family of Deep Learning approaches, multi-layer and often convolutional ML which advance the state-of-the-art for each approach named above, have special considerations (LeCun *et al.*, 2015). They are data hungry and building models using these approaches is challenging for small and medium datasets. When properly attached to truly big data, these methods do allow very large parameter models to optimize a loss function, thus maximizing prediction accuracy, but they are challenging in terms of transparency. Indeed, what these models give in prediction they take away in terms of understanding the causal reason for their explanation, and specifically in terms of understanding which features are important (Samek *et al.*, 2017).

8

Looking to the Future of Readability Research

Reading efficiently and easily has a direct and dramatic impact on education, health, and career outcomes. At the same time, digital devices provide an opportunity to create new personalized reading interfaces to build capacity for all readers, based on their individual needs and the demands of their specific reading tasks. Digital devices and their connected nature are rapidly changing availability and access to information, and only through research can we maximize everyone's reading potential. The HCI community is in a unique position to design and investigate reading interfaces that promote readability and facilitate the effective processing of text for all. While it is widely accepted that certain fonts or reading tools can help specific subpopulations (e.g., specialized fonts and rulers for readers with dyslexia, magnified text or contrast adjustment for ageing eyes, etc. (Duranovic *et al.*, 2018; Rello *et al.*, 2012, 2013, 2016; Scaltritti *et al.*, 2019)), a growing body of research in Human Computer Interaction and related disciplines is showing the benefits of individuating text formats to each reader (Ball *et al.*, 2021; Beier and Oderkerk, 2019a; Cai *et al.*, 2022; Crowley and Jordan, 2019a; Day *et al.*, 2022; Sheppard *et al.*, 2022a,b; Wallace *et al.*, 2020a,2022a; Watson and Wallace, 2021).

Influential early surveys on the typographical factors influencing reading performance include Tinker's *Legibility of Print* (1963) and Legge's work on the *Psychophysics of Reading in Normal and Low Vision* (in a series of articles published between 1985–2002; also in a book, by the same name, published in 2007). Between then and now, reading and information exchange more broadly has predominantly shifted to digital devices, requiring a re-examination of the text formatting, display, and customization opportunities. Nearly every combination of reader × device × task can prompt a separate research investigation, opening many future doors for researchers. Possible research questions include:

- How is reading performance affected by typographical factors in different languages and alphabets?
- How do the recommended text formats vary between populations who are "learning to read" (with a focus on letter and word recognition) and those that are "reading to learn" (with a focus on information extraction)? How do readability recommendations change over the lifespan (Section 3)?
- How do readability recommendations change over the course of a day, or between reading modes (Section 2)?
- How does the format or layout of a document affect the readability of the individual text components?
- What happens to reading and readability in the context of complex backgrounds and design elements (e.g., in the case of graphic designs, immersive displays, etc.; Section 5.1.1)?
- How does format readability interact with content readability (Section 4)? Can format adjustments reduce the cognitive load of, or increase comfort with, higher level content?
- How do the various reading metrics – speed, comprehension, preference, endurance, prosody, confidence, accuracy, etc. – relate to each other, and what are the trade-offs (Section 6.1)?

Who should care? An increased understanding about the effects of various typographic factors on reading performance stands to benefit a wide range of players: typographers, designers, and publishers can improve the general legibility of their outputs, or tailor designs to target particular audiences; technology companies and educational institutions can improve the accessibility and availability of their content and tools, to reach broader audiences; vision, cognitive, and neuro scientists can gain additional insights about the development and operation of the human brain; and every reader can discover what is best for them.

8.1 Take-Aways

Individuation benefits the reader. We advocate moving beyond one-format-fits-all approaches. Prior format readability studies have contributed significantly to our understanding of how typographical variables affect reader efficacy. However, these have focused on the idea that big changes, like font, can benefit all readers. We suggest a shift: readability research should ask how small changes to text format for each reader can create significant outcomes for them, noting that these variations will be different for each reader and will likely depend on content, device and context.

Personalization and accessibility go hand-in-hand. When we design for the individual and their unique needs, we can simultaneously improve the experiences of diverse groups of users. Branding text format adjustments as "accessibility tools" may be contributing to a low uptake of some of these tools by individuals who do not resonate with the label. Given the demonstrated benefits of customizing format readability for children and adults with varying needs, both the options themselves and their potential benefits should be made known more broadly.

Longitudinal research is required. Even if tuned to the individual's needs at a particular point in time, a reading format may not be optimal indefinitely. Readers change over their lifespans; they may gain reading proficiency or lose visual acuity, develop or overcome reading or learning conditions, change their reading behaviors or frequency, pick up new languages, and undergo a wide range of other changes, big and small, to their visual, cognitive, and professional abilities. Reading

adjustments will need to adapt to readers over time, and longitudinal research is required to tease out which properties of the reader and text, if any, are stable over time, and which necessitate regular updating. This will require developing personalized predictive models, the capture and analysis of the relevant behavioral and environmental characteristics, and consideration for the privacy and security of the data over time.

Multidisciplinary research is the key. Readability is a vast field requiring a multidisciplinary approach. Here, we provided a taste of the elements that must come together to form a readability study for HCI researchers, including: choosing reading materials and how they are displayed; selecting and recruiting participants; selecting the relevant equipment and tools; designing the study and collecting the relevant measurements. Readability studies in the future will range from simple studies of the readability of single characters and words all the way to advanced neuroimaging studies and long-term studies in schools or workplaces. These studies will focus on individuals, subpopulations or mainstream populations, reading on desktop, mobile, wearable, or immersive devices. Therefore, an understanding of perceptual science, human factors, reading subject matter expertise, design, neuroimaging, statistics, software engineering, sensors, and systems as well as machine learning must come together to craft meaningful experiments and analyze and interpret the results. The great benefit of the multidisciplinary research discussed here is that no one researcher needs to have all of this expertise. Experts each come with their own viewpoint, their own tools, and their own approaches for research design and data analysis. The most valuable revelations in this topic area are at intersections between fields.

Publicly available data and tools facilitate reproducible readability research. We urge communities of researchers, engineers, and designers to release reading content, typography, experimental designs, software platforms, analysis tools, and computational models. This will allow other groups to benefit from their expertise and work, to build directly on it, and compare results across populations, context, and settings. This may mean adopting approaches from experimental psychology, where many studies are preregistered and stimuli, data and analysis code are made available, or it might mean following practices

from computer science, releasing code and datasets at publication; there is no one perfect approach, but we advocate for openness when possible. Readability research has so much potential to help each and every reader, and we believe that achieving this goal requires making as much as we can available to the larger research community.

Appendices

A

Glossary

Apertures: The opening of the counters in letters such as “s”, “a”, “c” and “e”.

Augmented reality: A type of mixed reality that provides a user a composite of fabricated visual information and visual information originating from the real world.

Bézier curves: A mathematical way to describe curves used in computer graphics and animations.

Counter forms: The spatial area enclosed within a letter.

CSS: Cascading Style Sheets. It is a computer language used to manipulate the layout and visual presentation of HTML and XML documents within a web browser.

Delta hinting: Delta hinting instructions tell the rasterizer how best to render a font at given point sizes.

Dyslexia: A specific learning disability that is neurobiological in origin. It is characterized by difficulties with accurate and/or fluent word recognition and by poor spelling and decoding abilities.

Electroencephalography (EEG): EEG is used to measure and record small voltages generated by electrical brain activity using non-invasive electrodes placed on a participant's scalp while the participant is performing a cognitive or linguistic task of interest.

Event-related brain potentials (ERPs): ERPs are generated from EEG data. EEG is time-locked to brain activity impacted by a stimulus event of interest. A large number of trials of the same condition are averaged together to generate an ERP. ERP amplitudes, latencies, and scalp distributions can be compared from different experimental conditions or participant groups.

Exploratory Data Analysis (EDA): Initial investigations performed on data to help determine data quality, assist in numerical analysis, or build hypotheses for further investigation.

Eye tracking: The process of measuring eye activity, either in the form of eye movements or as the point of gaze. Measuring the movement of the eyes over time, usually separated into fixations, keeping the point of gaze at one position for a period of time, and saccades, movements between two locations (e.g., two words).

Fixation: Maintaining gaze position on a specific location in the world (e.g., on a word) for a period of time, often lasting a few hundred milliseconds.

Functional Magnetic Resonance Imaging (fMRI): An MRI scanner can be used to measure functional activity in the brain to investigate what brain regions (or brain networks) are activated by a particular task of interest. They use the Blood Oxygen-Level Dependent (BOLD) signal to measure changes to oxygenated blood as a marker of brain activity. Specific brain areas that are more active during a task require more oxygen to support the additional brain activity.

Glanceable reading: Reading a single word or multiple words: a label on an icon, notification on a wearable device, a road sign, or when looking for a target word on a page.

Interlude reading: Reading for a brief period of time; reading a few sentences, but not reading as the reader's primary task (e.g., reading a text message and then going back to another task).

IRB: Institutional Review Boards are committees that review research protocols to ensure that any proposed research is ethical.

Lexical decision task: A reading task where participants classify a string of letters as a word or non-word.

Long-form reading: Reading as a primary task, reading where the focus is on extracting information from the text.

Magnetic Resonance Imaging (MRI): MRI scanners use strong magnetic fields to generate high resolution anatomical images. Reading research uses structural brain MRI, which has the ability to distinguish between different types of tissues (e.g., gray matter and white matter) and brain structures.

Masked priming: A prime is presented for a short period of time and is visually masked (by some visual stimulus) at the same position, either before or after the prime.

Oral reading fluency: The process includes an evaluator documenting any errors made (words read or pronounced incorrectly, omitted, read out of order, or words pronounced for the student by the evaluator after a 3-second pause) and then calculating the total words read correctly per minute (WCPM). Fluency is the speed of accurate reading.

Orthographic processing: The spatial area enclosed within a letter.

Prosody: Prosody is the defining feature of expressive reading, comprising all of the variables of timing, phrasing, emphasis, and intonation that speakers use to help convey aspects of meaning.

Rasterizers: A software program that takes a font, a point size, and text as input and creates a bitmap rendering of the text in the given font and point size.

Readability: The ease with which a reader can recognize words, sentences, and paragraphs. The choice of typeface, characteristics of the type, and page layout can create a better (or worse) reading experience.

Readability formulas: Predictive tools which measure the level of the reading material. Examples include Flesch-Kincaid, Gunning Fog, SMOG, among others.

Readability index: An estimate of how difficult a text is to read, often measured using a readability formula.

Return sweep: The large eye movements from the end of a line to the beginning of the next.

Running record: An assessment of a child's fluency and behavior when reading out loud.

Saccade: A rapid shift of gaze between one location and another; for example, looking from one word to another word.

Semantic categorization task: A reading task where participants classify the categories of words (e.g., "is it alive or not?").

Staircase method: A psychophysics procedure to find a threshold for the task by adjusting parameters of the stimulus in real-time based on participants' responses, with the goal of converging on a preselected response accuracy level.

Qualitative Reading Inventory (QRI): An informal assessment of reading ability, used for schoolchildren.

Variable fonts: This font specification allows for different typeface variations to be contained in a single file, allowing CSS to manipulate their continuous range of design variants (i.e., width, weight, or style) contained in a single variable font.

Visual angle: The measurement, in degrees, that an object subtends when viewed by the eye, based on the size of the object and the distance to it from the eye.

Visual crowding: The inability to recognize objects in clutter. When applied to text, it impairs the ability read because of how cluttered the letters appear together.

Virtual Reality (VR): A type of mixed reality that provides a user fabricated visual information while obscuring visual information from the real world.

Visual search: The task of looking for a target (for example, a particular word, phrase or even a concept) among many distractors.

Visual span: The distance between stopping points when the eyes move across a line of text during reading; the amount of text that can be recognized in a single glance.

Web safe fonts: Fonts that are available to all web browsers.

B

Sample Survey Questions

(1) What is your age? (in years)

(2) What is your gender?

(3) What is/are your native language(s)?

(4) If you are a non-native English speaker, how many years have you lived in an English-speaking country?

(5) What other languages do you speak?

(6) What is your highest attained education level?

(7) Please describe your current occupation.

(8) Do you feel comfortable with reading articles written in English?

(9) How would you rate your speed as a reader?

(10) How would you rate your proficiency as a reader?

(11) Do you read to young children under the age of 6?

(12) Have you ever been diagnosed with a reading or learning disability (e.g., dyslexia)? If yes, which one and how long ago?

(13) Have you ever been diagnosed with any medical and neurological conditions (macular degeneration, diabetes, ADD, memory disorders, LPD, dyspraxia, other speech/pronunciation difficulties, etc.)? If yes, which one/s and how long ago?

(14) Are you currently under the influence of any drugs, medications, alcohol, or other stimulants (e.g., caffeine, nicotine) that may affect reading/attention? If yes, which?

(15) Do you have normal or corrected vision?

(16) If your vision is corrected, how was it corrected (glasses, lenses, surgery, etc.)?

(17) What device/s do you read on for leisure or personal interest?

(18) What device/s do you read on for work or study?

(19) What do you read for leisure or personal interest?

(20) What do you read for work or study?

(21) How often do you read English written articles for leisure or personal interest?

(22) How often do you read English written articles for work or study?

(23) Which device are you using right now to participate in this study?

(24) Please describe your current surroundings. For example, are you indoors/outside, by a window, under natural or artificial light, is the room light/dark, is the room small/large?

C

Openly Available Reading Corpora

A number of the recent readability works referenced in this manuscript (Cai *et al.*, 2022; Wallace *et al.*, 2020a; 2022a,b; Watson and Wallace, 2021) used a repository of 15 reading passages (300–500 word length, and a reduced 100–200 word set) at an 8th grade level with multiple-choice comprehension questions, provided under a research license: https://github.com/virtual-readability-lab/tochi-paper-materials-towards-individuated-reading.

The above reading passages were curated from Project Gutenberg, a library of over 60,000 free eBooks for which the U.S. copyright has expired: https://www.gutenberg.org/.

The CommonLit Library, provided by a nonprofit education technology organization, provides access to 2,000 free reading passages for grades 3–12 with assessments in English and Spanish under a CC BY-NC-SA 4.01 license: https://www.commonlit.org/en.

The OneStopQA corpus contains 30 Guardian articles in three difficulty levels (Elementary, Intermediate, Advanced), composed of a total of 162 paragraphs, with each paragraph corresponding to three multiple-choice comprehension questions, under a CC BY-SA 4.0 license: https://github.com/berzak/onestop-qa.

Asian and Pacific Speed Readings for ESL Learners (Millett, 2007) includes 20 reading passages of 550 words each with ten comprehension questions based on topics related to Asia and the Pacific written in the 1,000 more common words of the English language for teaching and research purposes.

Newsela provides hundreds of articles with corresponding activities (questions and writing prompts) at the elementary, middle, and high school levels across a variety of subjects available for academic research only: https://newsela.com/data/.

ReadWorks has a corpus of thousands of high-quality professionally written passages that are available for academic research into reading comprehension via a request form: https://about.readworks.org/corpus.html.

Acknowledgements

We acknowledge the following individuals for their valuable contributions to this manuscript: Betsy Laxton, Sarah Barrientos, Jose Echevarria, Xander Koo, Max Rose, Xi Wang, and Nardos Gebriye. We acknowledge the Readability Research Community. This monthly meeting of scientists from whom our authorship was drawn, provides so much inspiration in monthly discussions that the group's influence must surely be contained within these pages. Authors presented in alphabetical author for this collaborative work.

References

Agarwal, N., A. Chaudhari, D. R. Hansberry, K. L. Tomei, and C. J. Prestigiacomo (2013). "A comparative analysis of neurosurgical online education materials to assess patient comprehension". *Journal of Clinical Neuroscience*. 20(10): 1357–1361.

Agarwal, A. and A. Meyer (2009). "Beyond usability: Evaluating emotional response as an integral part of the user experience". *Extended Abstracts on Human Factors in Computing Systems*: 2919–2930. DOI: 10.1145/1520340.1520420.

Ahlström, C., K. Kircher, M. Nyström, and B. Wolfe (2021). "Eye tracking in driver attention research—How gaze data interpretations influence what we learn". *Frontiers in Neuroergonomics*. 2. DOI: 10.3389/fnrgo.2021.778043.

Ahrens, T. (2008). *Size-Specific Adjustments to Type Designs. An Investigation of the Principles Guiding the Design of Optical Sizes*. Mark Batty Publisher Academic.

Ahrens, T. (2012). "A closer look at font rendering". *Smashing Magazine*. URL: https://www.smashingmagazine.com/2012/04/a-closer-look-at-font-rendering/.

Ahrens, T. and S. Mugikura (2014). *Size-Specific Adjustments to Type Designs: An Investigation of the Principles Guiding the Design of Optical Sizes*. Just Another Foundry.

Ailon, N. (2008). "Reconciling real scores with binary comparisons: A unified logistic model for ranking". *Advances in Neural Information Processing Systems*. 21: 34–38.

Alexander, P. A., J. M. Kulikowich, and S. K. Schulze (1994). "How subject-matter knowledge affects recall and interest". *American Educational Research Journal*. 31(2): 313–337.

Alto, K. M., K. M. McCullough, and R. F. Levant (2018). "Who is on craigslist? A novel approach to participant recruitment for masculinities scholarship". *Psychology of Men and Masculinity*. 19(2): 319–324.

Ball, R. V., D. B. Miller, S. Wallace, K. C. Macias, M. Ibrahim, E. R. Gonzaga, O. Karasik, D. R. Rohlsen-Neal, S. Barrientos, E. A. Ross, A. Asmar, A. M. Hughes, P. A. Hancock, and B. D. Sawyer (2021). "Optimizing electronic health records through readability". *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care*. 10(1): 65–70.

Banerjee, J. and M. Bhattacharyya (2011). "Selection of the optimum font type and size interface for on screen continuous reading by young adults: An ergonomic approach". *Journal of Human Ergology*. 40(1_2): 47–62.

Barnard, L., J. S. Yi, J. A. Jacko, and A. Sears (2007). "Capturing the effects of context on human performance in mobile computing systems". *Personal and Ubiquitous Computing*. 11(2): 81–96.

Barnett, A. G. (2004). "Regression to the mean: What it is and how to deal with it". *International Journal of Epidemiology*. 34(1): 215–220.

Beier, S. (2012). *Reading Letters: Designing for Legibility*. BIS Publishers.

Beier, S. (2013). *Legibility Investigations: Controlling Typeface Variables*. APA.

Beier, S. (2017). *Type Tricks: Your Personal Guide to Type Design*. BIS Publishers.

Beier, S., J.-B. Bernard, and E. Castet (2018). *Numeral Legibility and Visual Complexity*. DRS Design Research Society. DOI: 10.21606/drs.2018.246.

Beier, S. and M. C. Dyson (2014). "The influence of serifs on '*h*' and '*i*': Useful knowledge from design-led scientific research". *Visible Language*. 47(3): 74–95.

Beier, S. and K. Larson (2010). "Design improvements for frequently misrecognized letters". *Information Design Journal*. 18(2): 118–137.

Beier, S. and K. Larson (2013). "How does typeface familiarity affect reading performance and reader preference?" *Information Design Journal*. 20(1): 16–31.

Beier, S. and C. A. T. Oderkerk (2019a). "The effect of age and font on reading ability". *Visible Language*. 53(3): 51–69.

Beier, S. and C. A. T. Oderkerk (2019b). "Smaller visual angles show greater benefit of letter boldness than larger visual angles". *Acta Psychologica*. 199: 102904.

Beier, S., C. A. T. Oderkerk, B. Bay, and M. Larsen (2021). "Increased letter spacing and greater letter width improve reading acuity in low vision readers". *Information Design Journal*. 26(1): 73–88.

Berglund, L. (2012). "Regression dilution bias: Tools for correction methods and sample size calculation". *Upsala Journal of Medical Sciences*. 117(3): 279–283.

Bernard, M. L., B. S. Chaparro, M. M. Mills, and C. G. Halcomb (2003). "Comparing the effects of text size and format on the readability of computer-displayed Times New Roman and Arial text". *International Journal of Human-Computer Studies*. 59(6): 823–835.

Bernard, M., C. H. Liao, and M. Mills (2001). "The effects of font type and size on the legibility and reading time of online text by older adults". In: *CHI'01 Extended Abstracts on Human Factors in Computing Systems*. 175–176.

Berry, J. D. (2004). *Now Read This: The Microsoft ClearType Font Collection; www.microsoft.com/typography/ctfonts*. Microsoft Corporation.

Bhatia, S. K., A. Samal, N. Rajan, and M. T. Kiviniemi (2011). "Effect of font size, italics, and colour count on web usability". *International Journal of Computational Vision and Robotics*. 2(2): 156–179.

Bizup, J. (2008). "BEAM: A rhetorical vocabulary for teaching research-based writing". *Rhetoric Review*. 27(1): 72–86.

Blanchard, H. E., A. Pollatsek, and K. Rayner (1989). "The acquisition of parafoveal word information in reading". *Perception and Psychophysics*. 46(1): 85–94.

Blignaut, P. (2010). "Visual span and other parameters for the generation of heatmaps". In: *Proceedings of the 2010 Symposium on Eye-Tracking Research and Applications*. 125–128. DOI: 10.1145/1743666.1743697.

Bokulich, N. A., J. R. Rideout, W. G. Mercurio, A. Shiffer, B. Wolfe, C. F. Maurice, R. J. Dutton, P. J. Turnbaugh, R. Knight, and J. G. Caporaso (2016). "Mockrobiota: A public resource for microbiome bioinformatics benchmarking". *MSystems*. 1(5). DOI: 10.1128/mSystems.00062-16.

Bolthausen, E. and M. V. Wüthrich (2013). "Bernoulli's law of large numbers". *ASTIN Bulletin: The Journal of the IAA*. 43(2): 73–79.

Bouaud, J. and B. Seroussi (1996). "Navigating through a document-centered electronic medical record: A mock-up based on www technology". In: *Proceedings of the AMIA Annual Fall Symposium*. 488.

Bouma, H. (1970). "Interaction effects in parafoveal letter recognition". *Nature*. 226(5241): 177–178.

Boyaci, O., A. Forte, S. A. Baset, and H. Schulzrinne (2009). "vDelay: A tool to measure capture-to-display latency and frame rate". In: *2009 11th IEEE International Symposium on Multimedia*. 194–200. DOI: 10.1109/ISM.2009.46.

Boyarski, D., C. Neuwirth, J. Forlizzi, and S. H. Regli (1998). "A study of fonts designed for screen display". *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems—CHI'98*: 87–94.

Bradshaw, M. T. (2011). "Analysts' forecasts: What do we know after decades of work?" Available at *SSRN 1880339*.

Brady, E., M. R. Morris, and J. P. Bigham (2015). "Gauging receptiveness to social microvolunteering". In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1055–1064. DOI: 10.1145/2702123.2702329.

Brandt, D. (1990). *Literacy as Involvement: The Acts of Writers, Readers, and Texts*. Southern Illinois University Press.

Brehm, J. W. (1966). *A Theory of Psychological Reactance*. Academic Press.

Brigo, F., W. M. Otte, S. C. Igwe, F. Tezzon, and R. Nardone (2015). "Clearly written, easily comprehended? The readability of websites providing information on epilepsy". *Epilepsy and Behavior*. 44: 35–39.

Brinberg, M., N. Ram, D. E. Conroy, A. L. Pincus, and D. Gerstorf (2022). "Dyadic analysis and the reciprocal one-with-many model: Extending the study of interpersonal processes with intensive longitudinal data". *Psychological Methods*. DOI: 10.1037/met0000380.

Bringhurst, R. (2004). *The Elements of Typographic Style*. WA: Hartley & Marks Point Roberts.

Brishtel, I., A. A. Khan, T. Schmidt, T. Dingler, S. Ishimaru, and A. Dengel (2020). "Mind wandering in a multimodal reading setting: Behavior analysis and automatic detection using eye-tracking and an EDA sensor". *Sensors*. 20(9): Article 9.

Broberg, P. (2013). "Sample size re-assessment leading to a raised sample size does not inflate type I error rate under mild conditions". *BMC Medical Research Methodology*. 13(1): 94.

Brooks, J., S. Nagels, and P. Lopes (2020). "Trigeminal-based temperature illusions". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

Brothers, T. and M. J. Traxler (2016). "Anticipating syntax during reading: Evidence from the boundary change paradigm". *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 42(12): 1894.

Brown, M., V. Savova, and E. Gibson (2012). "Syntax encodes information structure: Evidence from on-line reading comprehension". *Journal of Memory and Language*. 66(1): 194–209.

Brysbaert, M. (2019). "How many words do we read per minute? A review and meta-analysis of reading rate". *Journal of Memory and Language*. 109: 104047. DOI: 10.1016/j.jml.2019.104047.

Buckner, R. L., J. R. Andrews-Hanna, and D. L. Schacter (2008). "The brain's default network: Anatomy, function, and relevance to disease". In: *The Year in Cognitive Neuroscience 2008*. Blackwell Publishing. 1–38.

Buhrmester, M., T. Kwang, and S. D. Gosling (2011). "Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data". *Perspectives on Psychological Science*. 6(1): 3–5.

Burch, M., A. Veneri, and B. Sun (2019). "EyeClouds: A visualization and analysis tool for exploring eye movement data". In: *Proceedings of the 12th International Symposium on Visual Information Communication and Interaction*. 1–8. DOI: 10.1145/3356422.3356423.

Burmistrov, I., T. Zlokazova, I. Ishmuratova, and M. Semenova (2016). "Legibility of light and ultra-light fonts: Eyetracking study". In: *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*. 1–6. DOI: 10.1145/2971485.2996745.

Büttner, A., S. M. Grünvogel, and A. Fuhrmann (2020). "The influence of text rotation, font and distance on legibility in VR". In: *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. 662–663. DOI: 10.1109/VRW50115.2020.00182.

Cai, T., S. Wallace, T. Rezvanian, J. Dobres, B. Kerr, S. Berlow, J. Huang, B. D. Sawyer, and Z. Bylinskii (2022). "Personalized font recommendations: Combining ML and typographic guidelines to optimize readability". In: *Designing Interactive Systems Conference*. 1–25. DOI: 10.1145/3532106.3533457.

Calabrese, A., A. M. Cheong, S.-H. Cheung, Y. He, M. Kwon, J. S. Mansfield, A. Subramanian, D. Yu, and G. E. Legge (2016). "Baseline MNREAD measures for normally sighted subjects from childhood to old age". *Investigative Ophthalmology and Visual Science*. 57(8): 3836–3843.

Carillo, E. C. (2017). *A Writer's Guide to Mindful Reading: Practices and Possibilities*. The WAC Clearinghouse and University Press of Colorado. URL: https://wac.colostate.edu/books/practice/mindful/.

Carver, R. P. (1990). *Reading Rate: A Review of Research and Theory*. Academic Press.

Catts, H. W., A. McIlraith, M. S. Bridges, and D. C. Nielsen (2017). "Viewing a phonological deficit within a multifactorial model of dyslexia". *Reading and Writing*. 30(3): 613–629.

Christoff, K., A. M. Gordon, J. Smallwood, R. Smith, and J. W. Schooler (2009). "Experience sampling during fMRI reveals default network and executive system contributions to mind wandering". *Proceedings of the National Academy of Sciences*. 106(21): 8719–8724.

Chung, S. T. and J.-B. Bernard (2018). "Bolder print does not increase reading speed in people with central vision loss". *Vision Research*. 153: 98–104. DOI: 10.1016/j.visres.2018.10.012.

Cichy, R. M. and A. Oliva (2020). "A M/EEG-fMRI fusion primer: Resolving human brain responses in space and time". *Neuron*. 107(5): 772–781.

Clinton, V. (2019). "Reading from paper compared to screens: A systematic review and meta-analysis". *Journal of Research in Reading*. 42(2): 288–325.

Coates, D. R., D. M. Levi, P. Touch, and R. Sabesan (2018). "Foveal crowding resolved". *Scientific Reports*. 8(1): Article 1.

Cognolato, M., M. Atzori, and H. Müller (2018). "Head-mounted eye gaze tracking devices: An overview of modern devices and recent advances". *Journal of Rehabilitation and Assistive Technologies Engineering*. 5: 2055668318773991.

Cohen, L., S. Lehéricy, F. Chochon, C. Lemer, S. Rivaud, and S. Dehaene (2002). "Language-specific tuning of visual cortex? Functional properties of the visual word form area". *Brain*. 125(5): 1054–1069.

Cooke, L. (2006). "Is the mouse a 'poor man's eye tracker?'" In: *Proceedings of the 53rd Annual Conference of the Society for Technical Communication*. 252–255.

Cramer, E. D., L. Gonzalez, and C. Pellegrini-Lafont (2014). "From classmates to inmates: An integrated approach to break the school-to-prison pipeline". *Equity and Excellence in Education*. 47(4): 461–475.

Crossley, S. A., D. B. Allen, and D. S. McNamara (2011). "Text readability and intuitive simplification: A comparison of readability formulas". *Reading in a Foreign Language*. 23(1): 84–101.

Crowley, K. and M. Jordan (2019a). *Base Font Effect on Reading Performance*. URL: https://readabilitymatters.org/articles/font-effect.

Crowley, K. and M. Jordan (2019b). *Tech Proof of Concept Results Summary*. URL: https://readabilitymatters.org/results-summary.

Davis, C. J. (2010). "The spatial coding model of visual word identification". *Psychological Review*. 117(3): 713.

Davis, C. J. and J. S. Bowers (2004). "What do letter migration errors reveal about letter position coding in visual word recognition?" *Journal of Experimental Psychology: Human Perception and Performance*. 30(5): 923.

Day, S. L., A. Giroux, S. Wallace, R. Treitman, K. Crowley, M. Jordan, and B. D. Sawyer (2022). "The effect of font formats on reading speed and comprehension in grades 3–5". In: *Society for the Scientific Study of Reading (SSSR) Annual Conference*.

Dehaene, S. and L. Cohen (2011). "The unique role of the visual word form area in reading". *Trends in Cognitive Sciences*. 15(6): 254–262.

Demb, J. B., G. M. Boynton, and D. J. Heeger (1997). "Brain activity in visual cortex predicts individual differences in reading performance". *Proceedings of the National Academy of Sciences*. 94: 13363–13366.

Demets, D. L. and K. K. G. Lan (1994). "Interim analysis: The alpha spending function approach". *Statistics in Medicine*. 13(13–14): 1341–1352.

Dimigen, O., W. Sommer, A. Hohlfeld, A. M. Jacobs, and R. Kliegl (2011). "Coregistration of eye movements and EEG in natural reading: Analyses and review". *Journal of Experimental Psychology: General*. 140(4): 552.

Dingler, T., K. S. Kunze, and B. Outram (2018). "VR reading UIs: Assessing text parameters for reading in VR". In: *CHI, 2018—Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems: Engage with CHI, LBW094*. DOI: 10.1145/3170427.3188695.

Dingler, T., S. Li, N. van Berkel, and V. Kostakos (2020). "Page-turning techniques for reading interfaces in virtual environments". In: *32nd Australian Conference on Human-Computer Interaction*. 454–461.

Dobres, J., N. Chahine, and B. Reimer (2017a). "Effects of ambient illumination, contrast polarity, and letter size on text legibility under glance-like reading". *Applied Ergonomics*. 60: 68–73.

Dobres, J., S. T. Chrysler, B. Wolfe, N. Chahine, and B. Reimer (2017b). "Empirical assessment of the legibility of the highway gothic and clearview signage fonts". *Transportation Research Record*. 2624(1): 1–8.

Dobres, J., B. Reimer, and N. Chahine (2016). "The effect of font weight and rendering system on glance-based text legibility". In: *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. 91–96.

Downs, D. (2010). "Teaching first-year writers to use texts: Scholarly readings in writing-about-writing in first-year comp". *Reader*. 60(1): 19–50.

Drew, T., M. L.-H. Võ, and J. M. Wolfe (2013). "The invisible gorilla strikes again: Sustained inattentional blindness in expert observers". *Psychological Science*. 24(9): 1848–1853.

Duranovic, M., S. Senka, and B. Babic-Gavric (2018). "Influence of increased letter spacing and font type on the reading ability of dyslexic children". *Annals of Dyslexia*. 68(3): 218–228.

Ehrlich, M., L. Wisniewski, H. Trsek, D. Mahrenholz, and J. Jasperneite (2017). "Automatic mapping of cyber security requirements to support network slicing in software-defined networks". In: *22nd IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*. 1–4.

Eisfeld, H. and F. Kristallovich (2020). *The Rise of Dark Mode: A Qualitative Study of an Emerging User Interface Design Trend*. URL: http://urn.kb.se/resolve?urn=urn:nbn:se:hj:diva-50563.

Elbro, C. and I. Buch-Iversen (2013). "Activation of background knowledge for inference making: Effects on reading comprehension". *Scientific Studies of Reading*. 17(6): 435–452.

Elson, R. B. and D. P. Connelly (1995). "Computerized patient records in primary care: Their role in mediating guideline-driven physician behavior change". *Archives of Family Medicine*. 4(8): 698.

Fang, Z. (2016). "Teaching close reading with complex texts across content areas". *Research in the Teaching of English*. 51: 106–116.

Fiset, D., C. Blais, M. Arguin, K. Tadros, C. Ethier-Majcher, D. Bub, and F. Gosselin (2009). "The spatio-temporal dynamics of visual letter recognition". *Cognitive Neuropsychology*. 26(1): 23–35.

Fisher, D. and N. Frey (2014). "Contingency teaching during close reading". *The Reading Teacher*. 68(4): 277–286.

Fitchett, S. and A. Cockburn (2009). "Evaluating reading and analysis tasks on mobile devices: A case study of tilt and flick scrolling". *Proceedings of the 21st Annual Conference of the Australian Computer-Human Interaction Special Interest Group: Design: Open*. 24(7): 225–232.

Fletcher, J. M., G. R. Lyon, L. S. Fuchs, and M. A. Barnes (2019). *Learning Disabilities: From Identification to Intervention*. 2nd ed. The Guilford Press.

Fostick, L. and H. Revah (2018). "Dyslexia as a multi-deficit disorder: Working memory and auditory temporal processing". *Acta Psychologica*. 183: 19–28.

Fraser, C. A. (2007). "Reading rate in L1 Mandarin Chinese and L2 English across five reading tasks". *The Modern Language Journal*. 91(3): 372–394.

Fuchs, L. S., D. Fuchs, M. K. Hosp, and J. R. Jenkins (2001). "Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis". *Scientific Studies of Reading*. 5(3): 239–256.

Gabbard, J. L., J. E. Swan, and D. Hix (2006). "The effects of text drawing styles, background textures, and natural lighting on text legibility in outdoor augmented reality". *Presence: Teleoperators and Virtual Environments*. 15(1): 16–32.

Gaillard, W. D., L. M. Balsamo, Z. Ibrahim, B. C. Sachs, and B. Xu (2003). "FMRI identifies regional specialization of neural networks for reading in young children". *Neurology*. 60(1): 94–100.

Galliussi, J., L. Perondi, G. Chia, W. Gerbino, and P. Bernardis (2020). "Inter-letter spacing, inter-word spacing, and font with dyslexia-friendly features: Testing text readability in people with and without dyslexia". *Annals of Dyslexia*. 70(1): 141–152.

Gao, X., J. Dera, A. D. Nijhof, and R. M. Willems (2019). "Is less readable liked better? The case of font readability in poetry appreciation". *PLoS One*. 14(12): e0225757.

Germanò, E., A. Gagliano, and P. Curatolo (2010). "Comorbidity of ADHD and Dyslexia". *Developmental Neuropsychology*. 35(5): 475–493.

Goel, M., L. Findlater, and J. Wobbrock (2012). "WalkType: Using accelerometer data to accomodate situational impairments in mobile touch screen text entry". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2687–2696. DOI: 10.1145/2207676.2208662.

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

Graesser, A., D. Greenberg, A. Olney, and M. Lovett (2019). "Educational technologies that support reading comprehension for adults who have low literacy skills". In: *Wiley Handbook of Adult Literacy*. Wiley. 471–493. URL: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119261407.ch22.

Grainger, J., A. Rey, and S. Dufau (2008). "Letter perception: From pixels to pandemonium". *Trends in Cognitive Sciences*. 12(10): 381–387.

Grainger, J., W. J. Van Heuven, and P. Bonin (2004). "Modeling letter position coding in printed word perception". In: *The Mental Lexicon*. Nova Science, Publishers. 1–23.

Grohmann, B., J. L. Giese, and I. D. Parkman (2013). "Using type font characteristics to communicate brand personality of new brands". *Journal of Brand Management*. 20(5): 389–403.

Gugerty, L., R. A. Tyrrell, T. R. Aten, and K. A. Edmonds (2004). "The effects of subpixel addressing on users' performance and preferences during reading-related tasks". *ACM Transactions on Applied Perception*. 1(2): 81–101.

Gürtler, A. and C. Mengelt (1985). "Fundamental research methods and form innovation in type design compared to technological developments in type production". *Visible Language*. 19(1): 123.

Haas, C. and L. Flower (1988). "Rhetorical reading strategies and the construction of meaning". *College Composition and Communication*. 39(2): 167–183.

Hammoud, R. I. (2008). *Passive Eye Monitoring: Algorithms, Applications and Experiments*. Springer. URL: https://www.springer.com/gp/book/9783540754114.

Hancock, P. A., A. A. Pepe, and L. L. Murphy (2005). "Hedonomics: The power of positive and pleasurable ergonomics". *Ergonomics in Design*. 13(1): 8–14.

Hancock, P. A., B. D. Sawyer, and S. Stafford (2015). "The effects of display size on performance". *Ergonomics*. 58(3): 337–354.

Hendrickson, K. and K. L. Ailawadi (2014). "Six lessons for in-store marketing from six years of mobile eye-tracking research". In: *Shopper Marketing and the Role of In-Store Marketing*. Vol. 11. Emerald Group Publishing Limited. 57–74. DOI: 10.1108/S1548-64352014000011002.

Henriksen, B. S., I. H. Goldstein, A. Rule, A. E. Huang, H. Dusek, A. Igelman, M. F. Chiang, and M. R. Hribar (2020). "Electronic health records in ophthalmology: Source and method of documentation". *American Journal of Ophthalmology*. 211: 191–199. DOI: 10.1016/j.ajo.2019.11.030.

Hernandez, D. J. and J. S. Napierala (2013). "Early education, poverty, and parental circumstances among hispanic children: Pointing toward needed public policies". *Journal of the Association of Mexican American Educators*. 7(2): 30–39.

Hiebert, E. H., P. D. Pearson, E. H. Hiebert, and P. D. Pearson (2010). *An Examination of Current Text Difficulty Indices with Early Reading Texts*. Reading Research Report No. 10-01.

Highsmith, C. (2020). *Inside Paragraphs: Typographic Fundamentals*. 2nd ed. Princeton Architectural Press.

Ho, C.-J., A. Slivkins, S. Suri, and J. W. Vaughan (2015). "Incentivizing high quality crowdwork". In: *Proceedings of the 24th International Conference on World Wide Web*. 419–429. URL: 10.1145/2736277.2741102.

Hoitash, R., U. Hoitash, and L. Morris (2021). "eXtensible business reporting language: A review and implications for future research". *AUDITING: A Journal of Practice and Theory*. 40(2): 107–132. DOI: 10.2308/AJPT-2019-517.

Holcomb, P. J. and J. Grainger (2007). "Exploring the temporal dynamics of visual word recognition in the masked repetition priming paradigm using event-related potentials". *Brain Research*. 1180: 39–58. DOI: 10.1016/j.brainres.2007.06.110.

Huang, Y.-M. and T.-H. Liang (2015). "A technique for tracking the reading rate to identify the e-book reading behaviors and comprehension outcomes of elementary school students". *British Journal of Educational Technology*. 46(4): 864–876.

Hudson, J. (2016). "Introducing OpenType variable fonts". *Medium*. URL: https://medium.com/variable-fonts/https-medium-com-tiro-introducing-opentype-variable-fonts-12ba6cd2369.

Huey, E. B. (1908). *The Psychology and Pedagogy of Reading*. Macmillan.

Hughes, L. and A. Wilkins (2002). "Reading at a distance: Implications for the design of text in children's big books". *The British Journal of Educational Psychology*. 72: 213–226. DOI: 10.1348/000709902158856.

Hutcheon, J. A., A. Chiolero, and J. A. Hanley (2010). "Random measurement error and regression dilution bias". *BMJ*. 340: c2289.

Idsardi, W. (1992). *The Computation of Prosody*. Doctoral Thesis, Massachusetts Institute of Technology. URL: https://dspace.mit.edu/bitstream/handle/1721.1/12897/27832131-MIT.pdf?sequence=2.

Ikeda, K. and M. S. Bernstein (2016). "Pay it backward: Per-task payments on crowdsourcing platforms reduce productivity". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 4111–4121. DOI: 10.1145/2858036.2858327.

International Dyslexia Association (2022). *Dyslexia Basics*. URL: https://dyslexiaida.org/dyslexia-basics/.

Jamieson, S. (2013). "Reading and engaging sources: What students' use of sources reveals about advanced reading skills". *Across the Disciplines*. 10(4): 1–22.

Jang-Jaccard, J. and S. Nepal (2014). "A survey of emerging threats in cybersecurity". *Journal of Computer and System Sciences*. 80(5): 973–993.

Jo, J., B. Kim, and J. Seo (2015). "EyeBookmark: Assisting recovery from interruption during reading". In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. Association for Computing Machinery. 2963–2966. DOI: 10.1145/2702123.2702340.

Johnston, P. (1984). "Prior knowledge and reading comprehension test bias". *Reading Research Quarterly*. 19(2): 219–239.

Joo, S. J., A. L. White, D. J. Strodtman, and J. D. Yeatman (2018). "Optimizing text for an individual's visual system: The contribution of visual crowding to reading difficulties". *Cortex*. 103: 291–301. DOI: 10.1016/j.cortex.2018.03.013.

Jung, T.-P., S. Makeig, C. Humphries, T.-W. Lee, M. J. Mckeown, V. Iragui, and T. J. Sejnowski (2000). "Removing electroencephalographic artifacts by blind source separation". *Psychophysiology*. 37(2): 163–178.

Just, M. A. and P. A. Carpenter (1987). *The Psychology of Reading and Language Comprehension*. Allyn and Bacon.

Kessler, B. and R. Treiman (2015). "Writing systems: Their properties and implications for reading". In: *The Oxford Handbook of Reading*. Oxford University Press.

Killingsworth, M. A. and D. T. Gilbert (2010). "A wandering mind is an unhappy mind". *Science*. 330(6006): 932.

Kim, S., M. A. Nussbaum, and J. L. Gabbard (2019). "Influences of augmented reality head-worn display type and user interface design on performance and usability in simulated warehouse order picking". *Applied Ergonomics*. 74: 186–193. DOI: 10.1016/j.apergo.2018.08.026.

Klein, S. A. (2001). "Measuring, estimating, and understanding the psychometric function: A commentary". *Perception and Psychophysics*. 63(8): 1421–1455.

Kleitman, N. (1923). "Studies on the physiology of sleep: I. The effects of prolonged sleeplessness on man". *American Journal of Physiology-Legacy Content*. 66(1): 67–92.

Klemm, P., S. Oeltze-Jafra, K. Lawonn, K. Hegenscheid, H. Volzke, and B. Preim (2014). "Interactive visual analysis of image-centric cohort study data". *IEEE Transactions on Visualization and Computer Graphics*. 20(12): 1673–1682.

Knaack, L., A.-K. Lache, O. Preikszas, S. Reinhold, and M. Teistler (2019). *Improving Readability of Text in Realistic Virtual Reality Scenarios: Visual Magnification Without Restricting User Interactions*. Proceedings of Mensch Und Computer. 749–753. DOI: 10.1145/3340764.3344902.

Knowles, M. S. (1970). *The Modern Practice of Adult Education: Androgogy Versus Pedagogy*. New York Association Press.

Ko, Y.-H. (2017). "The effects of luminance contrast, colour combinations, font, and search time on brand icon legibility". *Applied Ergonomics*. 65: 33–40. DOI: 10.1016/j.apergo.2017.05.015.

Kou, G., D. Ergu, C. Lin, and Y. Chen (2016). "Pairwise comparison matrix in multiple criteria decision making". *Technological and Economic Development of Economy*. 22(5): 738–765.

Krafka, K., A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba (2016). *Eye Tracking for Everyone*. 2176–2184. URL: https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Krafka_Eye_Tracking_for_CVPR_2016_paper.html.

Kumar, G. and S. T. L. Chung (2014). "Characteristics of fixational eye movements in people with macular disease". *Investigative Ophthalmology and Visual Science*. 55(8): 5125–5133.

Kutas, M. and K. D. Federmeier (2011). "Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP)". *Annual Review of Psychology*. 62: 621–647. DOI: 10.1146/annurev.psych.093008.131123.

Kutas, M. and S. A. Hillyard (1980). "Reading senseless sentences: Brain potentials reflect semantic incongruity". *Science*. 207(4427): 203–205.

Larson, K. (2007). "The technology of text". *IEEE Spectrum*. 44(5): 26–31.

LeCun, Y., Y. Bengio, and G. Hinton (2015). "Deep learning". *Nature*. 521(7553): 436–444.

Lee, D.-S., K.-K. Shieh, S.-C. Jeng, and I.-H. Shen (2008). "Effect of character size and lighting on legibility of electronic papers". *Displays*. 29(1): 10–17.

Lee, M. and S. Hong (2021). "Adequate Sample Sizes for a Three-Level Growth Model". *Frontiers in Psychology*. 12: 685496.

Lee, J., D. Moon, I. Kim, and Y. Lee (2019). "A semantic approach to improving machine readability of a large-scale attack graph". *The Journal of Supercomputing*. 75(6): 3028–3045.

Leek, M. R. (2001). "Adaptive procedures in psychophysical research". *Perception and Psychophysics*. 63(8): 1279–1292.

Legge, G. E. (2007). *Psychophysics of Reading in Normal and Low Vision*. Lawrence Erlbaum.

Legge, G. E., S.-H. Cheung, D. Yu, S. T. L. Chung, H.-W. Lee, and D. P. Owens (2007). "The case for the visual span as a sensory bottleneck in reading". *Journal of Vision*. 7(2): 9.

Legge, G. E., G. S. Rubin, and A. Luebker (1987). "Psychophysics of reading—V. The role of contrast in normal vision". *Vision Research*. 27(7): 1165–1177.

Legge, G. E., G. S. Rubin, D. G. Pelli, and M. M. Schleske (1985). "Psychophysics of reading—II. Low vision". *Vision Research*. 25(2): 253–265.

Lehavy, R., F. Li, and K. Merkley (2011). "The effect of annual report readability on analyst following and the properties of their earnings forecasts". *The Accounting Review*. 86(3): 1087–1115.

Levi, D. M. (2008). "Crowding—An essential bottleneck for object recognition: A mini-review". *Vision Research*. 48(5): 635–654.

Levitt, H. (1971). "Transformed up-down methods in psychoacoustics". *The Journal of the Acoustical Society of America*. 49(2B): 467–477.

Li, F. (2010). "Survey of the literature". *Journal of Accounting Literature*. 29: 143–165.

Li, J., R. K. Mantiuk, J. Wang, S. Ling, and P. L. Callet (2018a). "Hybrid-MST: A hybrid active sampling strategy for pairwise preference aggregation". URL: http://arxiv.org/abs/1810.08851.

Li, J., R. K. Mantiuk, J. Wang, S. Ling, and P. L. Callet (2018b). "Hybrid-MST: A hybrid active sampling strategy for pairwise preference aggregation". *Advances in Neural Information Processing Systems*: 31.

Li, J., J. Wang, M. Barkowsky, and P. L. Callet (2018c). "Exploring the effects of subjective methodology on assessing visual discomfort in immersive multimedia". *Electronic Imaging*. 2018(14): 1–6.

Li, Q., S. J. Joo, J. D. Yeatman, and K. Reinecke (2020). "Controlling for participants' viewing distance in large-scale, psychophysical online experiments using a virtual chinrest". *Scientific Reports*. 10(1): 904.

Lindlbauer, D., A. M. Feit, and O. Hilliges (2019). "Context-aware online adaptation of mixed reality interfaces". In: *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 147–160.

Ling, J. and P. Schaik (2007). "The influence of line spacing and text alignment on visual search of web pages". *Displays*. 28: 60–67.

Liu, T.-Y. (2007). "Learning to rank for information retrieval". *Foundations and Trends® in Information Retrieval*. 3(3): 225–331.

Lotem, A., G. Cohen, I. Horn, and M. Meiseles (2012). *U.S. Patent No. 8,239,951*. Washington, DC: U.S. Patent and Trademark Office.

Loughran, T. and B. McDonald (2010). *Measuring Readability in Financial Text*. SSRN ELibrary.

Loughran, T. and B. McDonald (2014). "Measuring readability in financial disclosures". *The Journal of Finance*. 69(4): 1643–1671.

Macfadyen, H. (2011). "The reader's devices: The affordances of ebook readers". *Dalhousie Journal of Interdisciplinary Management*. 7. DOI: 10.5931/djim.v7i1.70.

Majaj, N. J., D. G. Pelli, P. Kurshan, and M. Palomares (2002). "The role of spatial frequency channels in letter identification". *Vision Research*. 42(9): 1165–1184.

Mäkelä, V., R. Rivu, S. Alsherif, M. Khamis, C. Xiao, L. Borchert, A. Schmidt, and F. Alt (2020). *Virtual Field Studies: Conducting Studies on Public Displays in Virtual Reality*. DOI: 10.1145/3313831.3376796.

Margolin, S. J., C. Driscoll, M. J. Toland, and J. L. Kegler (2013). "E-readers, computer screens, or paper: Does reading comprehension change across media platforms?" *Applied Cognitive Psychology*. 27(4): 512–519.

Martelli, M., G. Filippo, D. Spinelli, and P. Zoccolotti (2009). "Crowding, reading, and developmental dyslexia". *Journal of Vision*. 9(4): 14.

Mason, W. and S. Suri (2011). "Conducting behavioral research on Amazon's Mechanical Turk". *Behavior Research Methods*. 44(1): 1–23.

Massimi, M., R. Campigotto, A. Attarwala, and R. M. Baecker (2013). "Reading together as a leisure activity: Implications for E-reading". In: *Human-Computer Interaction – INTERACT*. Ed. by P. Kotzé, G. Marsden, G. Lindgaard, J. Wesson, and M. Winckler. Springer. 19–36. DOI: 10.1007/978-3-642-40480-1_2.

McCallum, D. R. and J. L. Peterson (1982). "Computer-based readability indexes". In: *Proceedings of the ACM '82 Conference*. 44–48. DOI: 10.1145/800174.809754.

McClelland, J. L. and D. E. Rumelhart (1981). "An interactive activation model of context effects in letter perception: I. An account of basic findings". *Psychological Review*. 88(5): 375–407.

McElree, B., G. L. Murphy, and T. Ochoa (2006). "Time course of retrieving conceptual information: A speed-accuracy trade-off study". *Psychonomic Bulletin and Review*. 13(5): 848–853.

McLaughlin, G. H. (1969). "SMOG grading-a new readability formula". *Journal of Reading*. 12(8): 639–646.

Microsoft HoloLens (2019). HoloLens 2 – Overview, Features, and Specs. URL: https://www.microsoft.com/en-us/hololens/hardware.

Millett, S. (2007). "Asian and pacific speed readings for ESL learners twenty passages written at the one thousand word level". *English Language Institute Occasional Publication*. 24. URL: https://citeseerx.ist.psu.edu/viewdoc/citations;jsessionid=4CD613749890C9C2E1E64CFFE7EAF77C?doi=10.1.1.163.4187.

Mills, C. B. and L. J. Weldon (1987). "Reading text from computer screens". *ACM Computing Surveys*. 19(4): 329–357.

Minakata, K., C. Oderkerk, and S. Beier (2020). "Low contrast in letter-stroke facilitates lexical identification". *Journal of Vision*. 20(11): 369.

Morris, R. A., K. Aquilante, D. Yager, and C. Bigelow (2002). *P-13: Serifs Slow RSVP Reading at Very Small Sizes, But Don't Matter at Larger Sizes*. Vol. 33. 244–247.

Mustonen, T., M. Olkkonen, and J. Hakkinen (2004). "Examining mobile phone text legibility while walking". In: *CHI '04 Extended Abstracts on Human Factors in Computing Systems*. 1243–1246. DOI: 10.1145/985921.986034.

Nam, S., Z. Bylinskii, C. Tensmeyer, C. Wigington, R. Jain, and T. Sun (2020). "Using behavioral interactions from a mobile device to classify the reader's prior familiarity and goal conditions". ArXiv: 2004.12016 [Cs]. URL: http://arxiv.org/abs/2004.12016.

Negahban, A. and C.-H. Chung (2014). "Discovering determinants of users perception of mobile device functionality fit". *Computers in Human Behavior*. 35: 75–84.

Ngiam, W. X., K. L. Khaw, A. O. Holcombe, and P. T. Goodbourn (2018). "Visual working memory for letters varies with familiarity but not complexity". *Journal of Experimental Psychology: Learning Memory, and Cognition*. 45(10): 1761.

Niantic, I. (2021). *Pokémon GO. Pokémon GO*. URL: https://pokemongolive.com/.

Niehorster, D. C., T. H. W. Cornelissen, K. Holmqvist, I. T. C. Hooge, and R. S. Hessels (2018). "What to expect from your remote eye-tracker when participants are unrestrained". *Behavior Research Methods*. 50(1): 213–227.

Nygren, E., M. Johnson, and P. Henriksson (1992). "Reading the medical record. II. Design of a human-computer interface for basic reading of computerized medical records". *Computer Methods and Programs in Biomedicine*. 39(1–2): 13–25.

Oderkerk, C. A. T. and S. Beier (2022). "Fonts of wider letter shapes improve letter recognition in parafovea and periphery". *Ergonomics*. 65(5): 753–761.

O'Donovan, P., J. Lībeks, A. Agarwala, and A. Hertzmann (2014). "Exploratory font selection using crowdsourced attributes". *ACM Transactions on Graphics*. 33(4): 92:1–92:9.

Olson, K. (2010). "An examination of questionnaire evaluation by expert reviewers". *Field Methods*. 22(4): 295–318.

Olulade, O. A., E. M. Napoliello, and G. F. Eden (2013). "Abnormal visual motion processing is not a cause of dyslexia". *Neuron*. 79(1): 180–190.

Ophir, E., C. Nass, and A. D. Wagner (2009). "Cognitive control in media multitaskers". *Proceedings of the National Academy of Sciences*. 106(37): 15583–15587.

Orlosky, J., K. Kiyokawa, and H. Takemura (2013). "Dynamic text management for see-through wearable and heads-up display systems". In: *Proceedings of the 2013 International Conference on Intelligent User Interfaces*. 363–370. DOI: 10.1145/2449396.2449443.

Osterhout, L. and P. J. Holcomb (1992). "Event-related brain potentials elicited by syntactic anomaly". *Journal of Memory and Language*. 31(6): 785–806.

Owsley, C. (2011). "Aging and vision". *Vision Research*. 51(13): 1610–1622.

Paolacci, G. and J. Chandler (2014). "Inside the Turk: Understanding mechanical Turk as a participant pool". *Current Directions in Psychological Science*. 23(3): 184–188.

Papoutsaki, A., P. Sangkloy, J. Laskey, N. Daskalova, J. Huang, and J. Hays (2017). "WebGazer: Scalable webcam eye tracking using user interactions". In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16)*. 3839–3845.

Parker, A., T. Slattery, and J. Kirkby (2019). "Return-sweep saccades during reading in adults and children". *Vision Research*. 155: 35–43. DOI: 10.1016/j.visres.2018.12.007.

Peer, E., L. Brandimarte, S. Samat, and A. Acquisti (2017). "Beyond the Turk: Alternative platforms for crowdsourcing behavioral research". *Journal of Experimental Social Psychology*. 70: 153–163. DOI: 10.1016/j.jesp.2017.01.006.

Pelli, D. G., C. W. Burns, B. Farell, and D. C. Moore-Page (2006). "Feature detection and letter identification". *Vision Research*. 46(28): 4646–4674.

Pelli, D. G. and K. A. Tillman (2007). "Parts, wholes, and context in reading: A triple dissociation". *PLoS One*. 2(8): e680.

Pennington, B. F. (2006). "From single to multiple deficit models of developmental disorders". *Cognition*. 101(2): 385–413.

Perea, M. and P. Gomez (2012). "Subtle increases in interletter spacing facilitate the encoding of words during normal reading". *PloS One*. 7(10): e47568.

Peterson, N. N., C. E. Schroeder, and J. C. Arezzo (1995). "Neural generators of early cortical somatosensory evoked potentials in the awake monkey". *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*. 96(3): 248–260.

Petrosky, A., D. Bartholomae, and T. Petrosky (2010). *Ways of Reading: An Anthology for Writers*. Bedford/St. Martin's.

Pew Internet Center (2013). *New Reading Data from the NEA's Survey of Public Participation in the Arts*. Pew Research Center: Internet, Science & Tech. URL: https://www.pewresearch.org/internet/2013/10/02/new-reading-data-from-the-neas-survey-of-public-participation-in-the-arts/.

Piepenbrock, C., S. Mayr, and A. Buchner (2014). "Smaller pupil size and better proofreading performance with positive than with negative polarity displays". *Ergonomics*. 57(11): 1670–1677.

Pires, I. M., N. M. Garcia, N. Pombo, F. Flórez-Revuelta, S. Spinsante, and M. C. Teixeira (2018). "Identification of activities of daily living through data fusion on motion and magnetic sensors embedded on mobile devices". *Pervasive and Mobile Computing*. 47: 78–93. DOI: 10.1016/j.pmcj.2018.05.005.

Plöchl, M., J. P. Ossandón, and P. König (2012). "Combining EEG and eye tracking: Identification, characterization, and correction of eye movement artifacts in electroencephalographic data". *Frontiers in Human Neuroscience*. 6. DOI: 10.3389/fnhum.2012.00278.

Powell, S. L. and A. D. Trice (2020). "The impact of a specialized font on the reading performance of elementary children with reading disability". *Contemporary School Psychology*. 24(1): 34–40.

Pušnik, N., A. Podlesek, and K. Možina (2016). "Typeface comparison—Does the x-height of lower-case letters increased to the size of upper-case letters speed up recognition?" *International Journal of Industrial Ergonomics*. 54: 164–169. DOI: 10.1016/j.ergon.2016.06.002.

Qian, L., J. Gao, and H. V. Jagadish (2015). "Learning user preferences by adaptive pairwise comparison". *Proceedings of the VLDB Endowment*. 8(11): 1322–1333.

Raichle, M. E., A. M. MacLeod, A. Z. Snyder, W. J. Powers, D. A. Gusnard, and G. L. Shulman (2001). "A default mode of brain function". *Proceedings of the National Academy of Sciences*. 98(2): 676–682.

Ramus, F., S. Rosen, S. C. Dakin, B. L. Day, J. M. Castellote, S. White, and U. Frith (2003). "Theories of developmental dyslexia: Insights from a multiple case study of dyslexic adults". *Brain*. 126(4): 841–865.

Rasinski, T. V., S.-C. Chang, E. Edmondson, J. Nageldinger, J. Nigh, L. Remark, K. S. Kenney, E. Walsh-Moorman, K. Yildirim, W. D. Nichols, D. D. Paige, and W. H. Rupley (2017). "Reading Fluency and College Readiness". *Journal of Adolescent and Adult Literacy*. 60(4): 453–460.

Rasinski, T. V., N. D. Padak, C. A. McKeon, L. G. Wilfong, J. A. Friedauer, and P. Heim (2005). "Is reading fluency a key for successful high school reading?" *Journal of Adolescent and Adult Literacy*. 49(1): 22–27.

Ravula, S. (2021). "Text analysis in financial disclosures". Preprint ArXiv: 2101.04480.

Rayner, K. (1975). "The perceptual span and peripheral cues in reading". *Cognitive Psychology*. 7(1): 65–81.

Rayner, K. (1998). "Eye movements in reading and information processing: 20 years of research". *Psychological Bulletin*. 124(3): 372–422.

Rayner, K., ed. (2012). *Psychology of Reading*. 2nd ed. Psychology Press.

Rayner, K., M. S. Castelhano, and J. Yang (2010). "Preview benefit during eye fixations in reading for older and younger readers". *Psychology and Aging*. 25(3): 714.

Reed, A. V. (1973). "Speed-accuracy trade-off in recognition memory". *Science*. 181(4099): 574–576.

Reeves, B., N. Ram, T. N. Robinson, J. J. Cummings, C. L. Giles, J. Pan, A. Chiatti, M. Cho, K. Roehrick, X. Yang, A. Gagneja, M. Brinberg, D. Muise, Y. Lu, M. Luo, A. Fitzgerald, and L. Yeykelis (2019). "Screenomics: A framework to capture and analyze personal life experiences and the ways that technology shapes them". *Human-Computer Interaction*. 36(2): 150–201.

Reeves, B., T. Robinson, and N. Ram (2020). "Time for the human screenome project". *Nature*. 577: 314–317. DOI: 10.1038/d41586-020-00032-5.

Reicher, G. M. (1969). "Perceptual recognition as a function of meaningfulness of stimulus material". *Journal of Experimental Psychology*. 81: 275–280. DOI: 10.1037/h0027768.

Reichle, E. D. (2021). *Computational Models of Reading: A Handbook*. Oxford University Press.

Reid, L., M. Reid, and A. Bennett (2004). "Towards a reader-friendly font: Rationale for developing a typeface that is friendly for beginning readers, particularly those labelled dyslexic". *Visible Language*. 38: 246–259.

Rello, L. and R. Baeza-Yates (2016). "The effect of font type on screen readability by people with dyslexia". *ACM Transactions on Accessible Computing*. 8(4): 15:1–15:33.

Rello, L. and R. Baeza-Yates (2017). "How to present more readable text for people with dyslexia". *Universal Access in the Information Society*. 16(1): 29–49.

Rello, L., G. Kanvinde, and R. Baeza-Yates (2012). "Layout guidelines for web text and a web service to improve accessibility for dyslexics". In: *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility*. 1–9. DOI: 10.1145/2207016.2207048.

Rello, L., M. Pielot, and M.-C. Marcos (2016). "Make it big!: The effect of font size and line spacing on online readability". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 3637–3648. DOI: 10.1145/2858036.2858204.

Rello, L., M. Pielot, M.-C. Marcos, and R. Carlini (2013). "Size matters (spacing not): 18 points for a dyslexic-friendly Wikipedia". *10th International Cross-Disciplinary Conference on Web Accessbility*. 10: 2461121–2461125. DOI: 10.1145/2461121.2461125.

Riehmann, P., M. Hanfler, and B. Froehlich (2005). "Interactive Sankey diagrams". *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.* 233–240. DOI: 10.1109/INFVIS.2005.1532152.

Rothwell, J. (2020). *Assessing the Economic Gains of Eradicating Illiteracy Nationally and Regionally in the United States.* Barbara Bush Foundation for Family Literacy. URL: https://www.barbarabush.org/wp-content/uploads/2020/09/BBFoundation_GainsFromEradicatingIlliteracy_9_8.pdf.

Rubin, G. S. and G. E. Legge (1989). "Psychophysics of reading. VI. The role of contrast in low vision". *Vision Research.* 29(1): 79–91.

Ružický, E., J. Lacko, J. Štefanovič, J. Hlaváč, and M. Šramka (2020). "Processing and visualization of medical data in a multiuser environment using artificial intelligence". *2020 Cybernetics Informatics (KI)*: 1–5. DOI: 10.1109/KI48306.2020.9039890.

Rzayev, R., P. W. Wozniak, T. Dingler, and N. Henze (2018). *Reading on Smart Glasses: The Effect of Text Position, Presentation Type and Walking.* Vol. 9. DOI: 10.1145/3173574.3173619.

Sabatini, J. P., J. Shore, S. Holtzman, and H. S. Scarborough (2011). "Relative effectiveness of reading intervention programs for adults with low literacy". *Journal of Research on Educational Effectiveness.* 4(2): 118–133.

Samek, W., T. Wiegand, and K.-R. Müller (2017). "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models". ArXiv: 1708.08296 [Cs, Stat]. URL: http://arxiv.org/abs/1708.08296.

Sanford, E. C. (1888). "The relative legibility of the small letters". *The American Journal of Psychology.* 1(3): 402–435.

Sawyer, B. D., B. Wolfe, J. Dobres, N. Chahine, B. Mehler, and B. Reimer (2020). "Glanceable, legible typography over complex backgrounds". *Ergonomics.* 63(7): 864–883.

Scaltritti, M., A. Miniukovich, P. Venuti, R. Job, A. De Angeli, and S. Sulpizio (2019). "Investigating effects of typographic variables on webpage reading through eye movements". *Scientific Reports.* 9(1): 12711.

Schildbach, B. and E. Rukzio (2010). "Investigating selection and reading performance on a mobile phone while walking". In: *Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services*. 93–102.

Schneegass, C. and F. Draxler (2021). "Designing task resumption cues for interruptions in mobile learning scenarios". In: *Technology-Augmented Perception and Cognition*. Ed. by T. Dingler and E. Niforatos. Springer International Publishing. 125–181. URL: 10.1007/978-3-030-30457-7_5.

Schnell, T., L. Yekhshatyan, and R. Daiker (2009). "Effect of luminance and text size on information acquisition time from traffic signs". *Transportation Research Record: Journal of the Transportation Research Board*. 2122(1): 52–62.

Seaboyer, J. and T. Barnett (2019). *New Perspectives on Reading and Writing Across the Disciplines*. DOI: 10.1080/07294360.2019.1544111.

Shatz, I. (2017). "Fast, free, and targeted: Reddit as a source for recruiting participants online". *Social Science Computer Review*. 35(4): 537–549.

Shaywitz, M. D. S. and M. D. J. Shaywitz (2020). *Overcoming Dyslexia, A New and Complete Science-Based Program for Reading Problems at Any Level*. 2nd ed. Vintage Books, a division of Random House, Inc.

Sheedy, J. E., M. V. Subbaram, A. B. Zimmerman, and J. R. Hayes (2005). "Text legibility and the letter superiority effect". *Human Factors*. 47(4): 797–815.

Sheppard, S., S. Nobles, S. Kajfez, A. Palma, K. Crowley, M. Jordan, and S. Beier (2022a). "Influences of font format on reading comprehension: Implications of font personalization in K-8 students". In: *Society for the Scientific Study of Reading (SSSR) Annual Conference*.

Sheppard, S., S. Nobles, A. Palma, S. Kajfez, M. Jordan, K. Crowley, and S. Beier (2022b). "The influence of font format and font format personalization on comprehension in child and adolescent readers". Under Review.

Siegenthaler, E., Y. Bochud, P. Bergamin, and P. Wurtz (2012). "Reading on LCD vs e-Ink displays: Effects on fatigue and visual strain." *Ophthalmic and Physiological Optics*. 32(5): 367–374.

Siegenthaler, E., P. Wurtz, P. Bergamin, and R. Groner (2011). "Comparing reading processes on e-ink displays and print". *Displays*. 32(5): 268–273.

Sihoe, A. D. L. (2015). "Rationales for an accurate sample size evaluation". *Journal of Thoracic Disease*. 7(11): E531–E536.

Slattery, T. J. and K. Rayner (2013). "Effects of intraword and interword spacing on eye movements during reading: Exploring the optimal use of space in a line of text". *Attention, Perception, and Psychophysics*. 75(6): 1275–1292.

Snell, J., S. van Leipsig, J. Grainger, and M. Meeter (2018). "OB1-reader: A model of word recognition and eye movements in text reading". *Psychological Review*. 125(6): 969.

Soleimani, H., S. Ketabi, and M. R. Talebinejad (2008). "The noticing function of output in acquisition of rhetorical structure of contrast paragraphs of iranian EFL university students". *Linguistik Online*. 34(2): Article 2. DOI: 10.13092/lo.34.527.

Soleimani, H. and E. Mohammadi (2012). "The effect of text typographical features on legibility, comprehension, and retrieval of EFL learners". *English Language Teaching*. 5(8): 207–216.

Sorenson Duncan, T., C. Mimeau, N. Crowell, and S. H. Deacon (2020). "Not all sentences are created equal: Evaluating the relation between children's understanding of basic and difficult sentences and their reading comprehension". *Journal of Educational Psychology*. 113(2): 268–278.

Spearman, C. (1908). "The method of 'right and wrong cases' ('constant stimuli') without Gauss's formulae". *British Journal of Psychology, 1904–1920*. 2(3): 227–242.

Spencer, M. and R. K. Wagner (2017). "The comprehension problems for second-language learners with poor reading comprehension despite adequate decoding: A meta-analysis". *Journal of Research in Reading*. 40(2): 199–217.

Spyridakis, J. H. and M. J. Wenger (1991). "An empirical method of assessing topic familiarity in reading comprehension research". *British Educational Research Journal*. 17(4): 353–360.

Srivastava, N., R. Jain, J. Healey, Z. Bylinskii, and T. Dingler (2021). "Mitigating the effects of reading interruptions by providing reviews and previews". In: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–6. DOI: 10.1145/3411763.3451610.

Stevens, J. P. (2012). *Applied Multivariate Statistics for the Social Sciences*. 5th ed. Routledge.

Stevens, S. S. (1946). "On the theory of scales of measurement". *Science*, New Series. 103(2684): 677–680.

Storer, K. M. and S. M. Branham (2019). "That's the way sighted people do it: What blind parents can teach technology designers about co-reading with children". In: *Proceedings of the 2019 on Designing Interactive Systems Conference*. 385–398. DOI: 10.1145/3322276.3322374.

Sweeney, M. A. (2018). "Audience awareness as a threshold concept of reading: An examination of student learning in biochemistry". *Research in the Teaching of English*. 53(1): 58–79.

Taptagaporn, S. and S. Saito (1990). "How display polarity and lighting conditions affect the pupil size of VDT operators". *Ergonomics*. 33(2): 201–208.

Taylor, S. E. (1965). "Eye movements in reading: Facts and fallacies". *American Educational Research Journal*. 2(4): 187–202.

Tejero, P., B. Insa, and J. Roca (2018). "Increasing the default interletter spacing of words can help drivers to read traffic signs at longer distances". *Accident Analysis and Prevention*. 117: 298–303. DOI: 10.1016/j.aap.2018.04.028.

The Yale Center for Dyslexia & Creativity (2022). *Dyslexia FAQ*. URL: https://dyslexia.yale.edu/dyslexia/dyslexia-faq.

Tinker, M. A. (1946). "The study of eye movements in reading". *Psychological Bulletin*. 43(2): 93–120.

Tinker, M. A. (1963). *Legibility of Print (Z250 A4 T5)*. Iowa City, IA: Iowa State University Press.

Tracy, W. (1986). *Letters of Credit: A View of Type Design*. London: Gordon Fraser.

Treisman, A. M. and G. Gelade (1980). "A feature-integration theory of attention". *Cognitive Psychology*. 12(1): 97–136.

Treutwein, B. and H. Strasburger (1999). "Fitting the psychometric function". *Perception and Psychophysics*. 61(1): 87–106.

U.S. Department of Education (2022). *NAEP Long-Term Trend Assessment Results: Reading and Mathematics*. URL: https://www.nationsreportcard.gov/highlights/ltt/2022/.

Valliappan, N., N. Dai, E. Steinberg, J. He, K. Rogers, V. Ramachandran, P. Xu, M. Shojaeizadeh, L. Guo, K. Kohlhoff, and V. Navalpakkam (2020). "Accelerating eye movement research via accurate and affordable smartphone eye tracking". *Nature Communications*. 11(1): Article 1. DOI: 10.1038/s41467-020-18360-5.

van der Mark, S., P. Klaver, K. Bucher, U. Maurer, E. Schulz, S. Brem, E. Martin, and D. Brandeis (2011). "The left occipitotemporal system in reading: Disruption of focal fMRI connectivity to left inferior frontal and inferior parietal language areas in children with dyslexia". *Neuroimage*. 54(3): 2426–2436.

van Engen-Verheul, M. M., L. W. Peute, N. F. de Keizer, N. Peek, and M. W. Jaspers (2016). "Optimizing the user interface of a data entry module for an electronic patient record for cardiac rehabilitation: A mixed method usability approach". *International Journal of Medical Informatics*. 87: 15–26. DOI: 10.1016/j.ijmedinf.2015.12.007.

Vellutino, F. R., D. M. Scanlon, S. Small, and D. P. Fanuele (2006). "Response to intervention as a vehicle for distinguishing between children with and without reading disabilities: Evidence for the role of kindergarten and first-grade interventions". *Journal of Learning Disabilities*. 39(2): 157–169.

VIVE (2018). "VIVE Pro Eye Overview". URL: https://www.vive.com/eu/product/vive-pro-eye/overview/.

Wallace, S., Z. Bylinskii, J. Dobres, K. Kerr, S. Berlow, R. Treitman, N. Kumawat, K. Arpin, M. Miller, J. Huang, and B. D. Sawyer (2022a). "Towards individuated reading experiences: Different fonts increase reading speed for different Individuals". *ACM Transactions on Computer-Human Interaction (TOCHI)*. 29(4). DOI: 10.1145/3502222.

Wallace, S., J. Dobres, Z. Bylinskii, and B. D. Sawyer (2022b). "Space for readability: Effects of reading speed from individuated character and word spacing". *Journal of Vision*. 22.

Wallace, S., R. Treitman, J. Huang, B. D. Sawyer, and Z. Bylinskii (2020a). "Accelerating adult readers with typeface: A study of individual preferences and effectiveness". In: *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–9. URL: 10.1145/3334480.3382985.

Wallace, S., R. Treitman, N. Kumawat, K. Arpin, J. Huang, B. Sawyer, and Z. Bylinskii (2020b). "Towards readability individuation: The right changes to text format make large impacts on reading speed". *Journal of Vision*. 20(10): 17.

Wang, Y., Y. Gao, and Z. Lian (2020). "Attribute2Font: Creating fonts you want from attributes". *ACM Transactions on Graphics*. 39(4): 69:69:1–69:69:15.

Warm, J. S., R. Parasuraman, and G. Matthews (2008). "Vigilance requires hard mental work and is stressful". *Human Factors*. 50(3): 433–441.

Watson, A. B. and D. G. Pelli (1983). "Quest: A Bayesian adaptive psychometric method". *Perception and Psychophysics*. 33(2): 113–120.

Watson, A. and S. Wallace (2021). "Improving reading outcomes using digital reading rulers for readers with and without dyslexia". *Journal of Vision*. 21(9): 2650.

Wei, C., D. Yu, and T. Dingier (2020). "Reading on 3D surfaces in virtual environments". In: *IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 721–728.

Wheeler, D. D. (1970). "Processes in word recognition". *Cognitive Psychology*. 1(1): 59–85.

Whitney, C. (2001). "How the brain encodes the order of letters in a printed word: The SERIOL model and selective literature review". *Psychonomic Bulletin and Review*. 8(2): 221–243.

Wisiecka, K., K. Krejtz, I. Krejtz, D. Sromek, A. Cellary, Lewandowska, B., and A. Duchowski (2022). "Comparison of webcam and remote eye tracking". In: *2022 Symposium on Eye Tracking Research and Applications*. 1–7. DOI: 10.1145/3517031.3529615.

Wisotzky, E. L., J.-C. Rosenthal, P. Eisert, A. Hilsmann, F. Schmid, M. Bauer, A. Schneider, and F. C. Uecker (2019). "Interactive and multimodal-based augmented reality for remote assistance using a digital surgical microscope". In: *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 1477–1484.

Wolf, M. (2018). *Reader, Come Home: The Reading Brain in A Digital World*. Harper.

Wolfe, B., A. Kosovicheva, S. Stent, and R. Rosenholtz (2021). "Attentional cueing in the world: Temporal and spatiotemporal cues for road hazards". *Journal of Vision*. 21(9): 2218.

Wolfe, B., B. D. Sawyer, and R. Rosenholtz (2020). "Toward a theory of visual information acquisition in driving". *Human Factors*. 64(4): 694–713. DOI: 10.1177/0018720820939693.

Wolfe, J. M. (2021). "Guided search 6.0: An updated model of visual search". *Psychonomic Bulletin and Review*. 28: 1–33.

Wolfe, J. M., K. R. Cave, and S. L. Franzel (1989). "Guided search: An alternative to the feature integration model for visual search". *Journal of Experimental Psychology: Human Perception and Performance*. 15(3): 419–433.

Wong, C. A., V. A. Miller, K. Murphy, D. Small, C. A. Ford, S. M. Willi, J. Feingold, A. Morris, Y. P. Ha, J. Zhu, W. Wang, and M. S. Patel (2017). "Effect of financial incentives on glucose monitoring adherence and glycemic control among adolescents and young adults with type 1 diabetes: A randomized clinical trial". *JAMA Pediatrics*. 171(12): 1176–1183.

Xiong, Y.-Z., E. A. Lorsung, J. S. Mansfield, C. Bigelow, and G. E. Legge (2018). "Fonts designed for macular degeneration: Impact on reading". *Investigative Ophthalmology and Visual Science*. 59(10): 4182–4189.

Xu, R. and D. Wunsch (2005). “Survey of clustering algorithms”. *IEEE Transactions on Neural Networks*. 16(3): 645–678.

Yamabe, T. and K. Takahashi (2007). “Experiments in Mobile User Interface Adaptation for Walking Users”. *The 2007 International Conference on Intelligent Pervasive Computing (IPC)*: 280–284. DOI: 10.1109/IPC.2007.94.

Yeung, A. W., T. K. Goto, and W. K. Leung (2018). “Readability of the 100 most-cited neuroimaging papers assessed by common readability formulae”. *Frontiers in Human Neuroscience*. 12: 308. DOI: 10.3389/fnhum.2018.00308.

Yeykelis, L., J. J. Cummings, and B. Reeves (2014). “Multitasking on a single device: Arousal and the frequency, anticipation, and prediction of switching between media content on a computer”. *Journal of Communication*. 64(1): 167–192.

Zhang, X., Y. Sugano, and A. Bulling (2019). “Evaluation of appearance-based methods and implications for gaze-based applications”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13. DOI: 10.1145/3290605.3300646.

Zhou, S., H. Jeong, and P. A. Green (2017). “How consistent are the best-known readability equations in estimating the readability of design standards?” *IEEE Transactions on Professional Communication*. 60(1): 97–111.

Zineddin, A. Z., P. M. Garvey, R. A. Carlson, and M. T. Pietrucha (2003). “Effects of practice on font legibility”. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. 47(13): 1717–1720.