©Copyright 2022
Qisheng Li

Conducting Volunteer-based Online Studies with People with Cognitive Disabilities

Qisheng Li

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:

Katharina Reinecke, Chair

James Fogarty

Adam Fourney

Program Authorized to Offer Degree:
Computer Science & Engineering

University of Washington

Abstract

Conducting Volunteer-based Online Studies with People with Cognitive Disabilities

Qisheng Li

Chair of the Supervisory Committee:
Katharina Reinecke
Paul G. Allen School of Computer Science & Engineering

To design accessible technology, it is vital in both academia and industry to solicit information from people with disabilities. Gathering information from those with cognitive disabilities is particularly important since the impact of these disorders differs by individual and over time. However, traditional recruiting methods, through gate-keepers such as local organizations, jeopardize generalizability due to the small numbers of such participants. Therefore, some researchers turn to online experiments, including those with volunteers, to increase participation and diversity. However, there is little understanding of whether volunteer-based online experiments are suitable for studying those with cognitive disabilities: Can we attract sufficient numbers of such volunteers? Can these studies be conducted rigorously? Can we provide benefits and support in return for volunteer participation?

In this dissertation, I address the trade-off between needing more diverse and larger numbers of participants and needing more control to conduct rigorous studies that target cognitive disabilities. Towards the goal, I (1) demonstrate the viability of volunteer-based online experiments in studying people with cognitive disabilities at scale, (2) identify how volunteer-based online studies related to cognitive disabilities are currently perceived by different stakeholders, i.e., participants and healthcare professionals, and (3) demonstrate the feasibility of using online experiments as a method to rigorously conduct studies that inform design guidelines of technologies for cognitive disabilities, in particular, dyslexia. Together, this work demonstrates the thesis statement: *Volunteer-based online studies can be*

conducted with people with cognitive disabilities in a way that is large-scale, self-motivating, helpful for participants, and enables rigorous experiments.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Number Page

LIST OF TABLES

ACKNOWLEDGMENTS

This work would not have been possible without the support of many wonderful people. First of all, I want to thank my advisor Katharina Reinecke. I joined UW CSE right after college with the minimum knowledge of what it means to be a Ph.D. student in Computer Science and in HCI. With her guidance throughout my Ph.D. career, I have learned how to conduct rigorous research, write elegant academic papers, and be open-minded and critical of research ideas. Reflecting on my Ph.D. journey, I have become more confident and independent and find myself part of the HCI community, because of Katharina's mentorship and support along the way. I would also like to thank Tim Althoff, James Fogarty, Adam Fourney, and Alexis Hiniker for being on my thesis committee. I appreciate all of their suggestions provided and the questions asked on my dissertation.

I am lucky to have done several industry internships at Adobe, Google, and Microsoft, and have had the opportunities to collaborate with many inspiring mentors: Zoya Bylinskii, Krzysztof Z. Gajos, Zeyu Jin, Eunyee Koh, Kevin Larson, Tak Yeon Lee, Tanya Matskewich, Meredith Ringel Morris, Justin Salamon, Jimmy Tobin, Katrin Tomanek, Subhashini Venugopalan, and Jason Yeatman. I want to thank them all for their mentorship and for providing me with invaluable internship experience.

I am also grateful to have made many friends in the HCI community along the way. I would like to show my love to everyone in the WildLab for all the lab lunches, exchange of research ideas, and sleepless deadlines that we have pulled together – especially to Nigini for helping me set up endless LabintheWild experiments, and to Tal for being my Ph.D. buddy, and for getting through every milestone together. Kudos to amazing undergraduates Josie, Lior, and Christina for their enthusiasm for learning about and contributing to my research projects. Thank Mingrui, Liang, Ruotong, Rock, Hao-Fei, Zheng, Yuhang, and many others for all the random chats about research, life, gossip, and food. There are just

too many people that I cannot list all of them here. I feel a sense of belonging to the HCI community because of every and each of you.

Finally, I would like to thank my friends who have always been the emotional support that help me cross the finishing line: Thank my officemates Aida, Sofia, and Saadia for including me in those NLP conversations; thank Xingfan, Sinan, Meredith, Beibin, Ellen, Jasmine, Jiayin, Leixin, Yezi, Zhaoqi, Yishuo for exploring the Pacific Northwest with me together, and for sharing the love from your fur-babies. Thank my college friends Yi, Chunyu, Di, Yifei, and many others for visiting me in Seattle, playing board games together, and for many encouraging conversations we had over the phone. You all have made my life more delightful and less lonely during the pandemic. Last but not least, I want to express my greatest gratitude to my family - my dad and mom - for their unconditional love, trust, and support. I hope they can be proud of their daughter being the first Dr. Li in the family and I hope to see them again soon. Of course, I will not forget to thank Sweet Potato, the little low rider, for being my family member since 2019. She has brought so much joy to my life for being cute, smart, mischievous, playful, and fearless.

Thank you all!

DEDICATION

To my parents, for their unconditional love, trust, and support.

Chapter 1

INTRODUCTION

Soliciting insights from people with disabilities through various research studies is vital in both academia and industry for the design of accessible technology. Yet many researchers struggle to recruit users that meet particular characteristics in sufficiently large numbers [53, 176] to capture population nuances. Traditional recruiting occurs through gatekeepers, such as local organizations or advocacy groups [20], or by establishing local participant pools, which can be time- and labor-intensive. These approaches also risk loss of generalizability and data quality if the same small population repeatedly participates in similar studies [206].

To study the needs of people with disabilities, some researchers have therefore turned to alternative recruiting methods. This includes using online labor markets, such as Prolific or Amazon Mechanical Turk (MTurk) [39, 213, 223], to support efficient recruitment of disabled populations at low cost. Participating in online experimentation can benefit those with disabilities: in addition to receiving financial compensation, they may find it more convenient and feasible than having to travel to a laboratory [33, 57, 244]. However, participating in online labor markets can also challenge people with disabilities due to usability problems and limitations of those platforms [37, 222, 244]. Some participants also struggle to find tasks that match their abilities [244].

Laboratory studies benefit from a supervised and more controlled environment, allowing setups that require specialized equipment, such as eye tracking devices. Online studies, in contrast, are unsupervised, with participants providing input on different devices and from different environments, creating uncontrolled factors that may affect testing reliability. Therefore, a trade-off arises between the need for a larger, more diverse participant pool and the need for more precise control of study implementation.

This dissertation evaluates the suitability of an alternative methodology for recruiting and gaining insights from people with disabilities: *volunteer-based online experiments.*

This approach, conducted on a variety of platforms (e.g., TestMyBrain.org, GamesWithWords.org, LabintheWild.org), usually provides personalized performance feedback in exchange for study participation. Previous experiments conducted on these platforms have been shown to attract more diverse participants than laboratory studies and those conducted on MTurk, in terms of age, education level, and geographic distribution [77,186]. Obtaining larger and more diverse sample sizes could extend the findings of smaller-scale laboratory studies, help researchers measure variability between people with specific disabilities, and assist with verifying results provided by people from diverse demographic backgrounds.

However, important questions about volunteer-based online experiments in the context of people with disabilities remain unanswered. For example, do these experiments attract sufficiently large numbers of participants with disabilities to robustly conduct comparative studies? How are these studies perceived by participants with disabilities? What types of studies can be rigorously conducted online with this population? Addressing these questions may shed light on how online experiments should be designed to attract large samples and provide adequately rewarding and engaging experiences for disabled participants.

In this dissertation, I focus my exploration on participants with *cognitive disabilities* since at least 20% of the world population experience a cognitive or mental disability at some time in their lives [2,217]. Examples of common cognitive disabilities include autism spectrum disorder (ASD), age-related cognitive decline, and dyslexia. Within the wide range of cognitive disabilities, I in particular focus on dyslexia, a specific learning disability that affects as many as 15-20% of the whole population with a cluster of language-related symptoms, especially reading [8]. I focus on the research of dyslexia because dyslexia occurs in people of all backgrounds and intellectual levels, and it is shown that an early diagnosis and intervention is extremely important to set people up for academic and career success [143]. Therefore, I am eager to explore how technological designs could facilitate early diagnosis and intervention to benefit this large population. Furthermore, the exact causes of dyslexia still remain unclear. The impact dyslexia has is also different for each person: it depends on the severity of the condition and the effectiveness of instruction or remediation. Therefore, I am inspired to contribute to the understanding of dyslexia with a large number of people with a variety of characteristics.

1.1 Thesis Contribution

This dissertation seeks to (1) examine the viability of volunteer-based online experiments in studying people with cognitive disabilities at scale, (2) identify how volunteer-based online studies related to cognitive disabilities are currently perceived by different stakeholders, i.e., participants and healthcare professionals, and (3) demonstrate the feasibility of using online experiments as a method to rigorously conduct studies that inform design guidelines of technologies for cognitive disabilities, in particular, dyslexia.

My work is guided by the following research questions (RQs):

- RQ1: Can we attract sufficiently large numbers of individuals with cognitive disabilities to participate in volunteer-based online experiments?
- RQ2: How are these volunteer-based online experiments perceived by people with cognitive disabilities and healthcare professionals?
- RQ3: How can we conduct rigorous, more controlled, and generalizable volunteer-based online experiments to study cognitive disabilities?

My thesis statement postulates that:

> *Volunteer-based online studies can be conducted with people with cognitive disabilities in a way that is large-scale, self-motivating, helpful for participants, and enables rigorous experiments.*

This dissertation makes the following contributions.

1.1.1 Replication and Extension of Prior Laboratory Studies

As researchers, we conduct rigorous research studies to replicate and extend prior work. At the intersection of Human-Computer Interaction (HCI) and accessibility research, an important question is whether we can repeat, extend, and generalize from studies that

include people with disabilities while conducting the studies online and with larger and more diverse populations.

Analyzing four online experiments on LabintheWild with a total of 355,656 participants, I show that volunteer-based online experiments that provide personalized feedback attract large numbers of participants with diverse disabilities and allow robust studies with these populations that replicate and extend the findings of prior laboratory studies. By additionally analyzing participants' feedback and forum entries that discuss LabintheWild experiments, I show that participants use the studies to diagnose themselves, compare their abilities to others', quantify potential impairments, self-experiment, and share their stories. I use these findings to inform design guidelines for online experiment platforms that adequately support and engage people with disabilities.

1.1.2 Different Stakeholder Perceptions of Online Tests

It is important to understand how online experiments involving people with cognitive disabilities are perceived by various stakeholders, including participants who take the experiments and interpret the results, as well as healthcare professionals who are experts in diagnosing and treating these conditions. Because these stakeholders benefit from online experiments for different reasons, recognizing these perspectives would help us design and deploy studies that benefit all constituencies.

To gather these perspectives, I conduct interviews to assess varied motivations for and experience using online tests. I find that online tests serve as an important resource to address shortcomings in support systems for people with professionally diagnosed or suspected cognitive disabilities. Interview results also uncover challenges and risks that prevent people with known or suspected health conditions from fully taking advantage of online tests. I also conduct interviews with healthcare professionals to learn their views on existing online tests, including whether and when they recommend these online studies, and what to be aware of when participating. Based on these findings, we discuss how online tests can be better leveraged to support people with cognitive disabilities before and after professional diagnosis.

1.1.3 Empirical Understanding of Dyslexia Using Online Experiments

Given the understanding of how online studies are perceived by these different stakeholders, an open question is whether the online experiment method could generate novel insights about how technology designs affect people's experiences, especially those with cognitive disabilities. This takes our investigation one step further than exploring previous study replications.

I focus my research on dyslexia, a learning disability that affects as many as 20% of the world's population [8]. I first conduct an online study to collect reading measurements, such as reading speed, reading comprehension, and perceived readability and aesthetics from people with and without dyslexia. I show how Reader View, a feature provided by many web browsers, improves people's reading speed and user experience and suggest guidelines for the design of websites and web browsers that better support those with divergent reading skills.

However, not every in-laboratory study can be easily replicated as an online experiment. For instance, some psycho-physical experiments that study dyslexia require precise control of the location and size of on-screen stimuli relative to human eyes. In these lab studies, participants usually use a chin rest to fixate their viewing distance for precise control. To enable such experiments outside of the lab, I invent the Virtual Chinrest, a method that measures a participant's viewing distance in the web browser by locating their blind spot; this makes it possible to automatically adjust stimulus configurations based on an individual's viewing distance. I validate the Virtual Chinrest in four in-laboratory and online studies, including the first large-scale study with 1153 participants in which it reveals that visual crowding, a phenomenon extremely sensitive to visual angle, varies between populations with varying ages and between people with and without dyslexia.

My contribution here also includes making Virtual Chinrest an open-sourced JavaScript library [1] and integrating it in jsPsych [2], a widely used JavaScript framework for creating behavioral experiments that run in a web browser.

---

[1] https://github.com/QishengLi/virtual_chinrest

[2] https://www.jspsych.org/7.0/plugins/virtual-chinrest/index.html

1.2 Thesis Overview

This thesis consists of nine chapters. Chapter 2 positions this work in the context of related research into the evolution of citizen science, volunteer-based online studies, current methods to study cognitive disabilities (in particular, dyslexia) in HCI, as well as related work in disability studies on people's motivation to participate in such studies.

The following five chapters are organized into two parts. Part 1 (Chapter 3-5) describes the viability of conducting volunteer-based online studies with people with cognitive disabilities from the perspectives of different stakeholders:

- Chapter 3 demonstrates, from **researchers' perspective**, that volunteer-based online studies could attract a large number of self-motivated participants with disabilities, and in particular, cognitive disabilities, as well as replicate and extend previous in-laboratory study results. I also present an initial exploration on what motivates people with disabilities to participate in such online studies with personalized feedback. This work was published at ASSETS'18 [135], collaboratively with Krzysztof Z. Gajos and Katharina Reinecke.

- Chapter 4 delves deeper into how **participants** feel about existing volunteer-based online studies, what attracts or hinders them from participation before and after diagnosis, and what additional enhancements can be made. This work was in collaboration with Josephine Lee, Christina Zhang and Katharina Reinecke, and won the Best Paper Award at ASSETS'21 [137].

- Chapter 5 explores **healthcare professionals'** opinions on utilizing the free and public online tests for people with (suspected) cognitive disabilities and the advantages/drawbacks of taking such online studies given existing healthcare systems.

Part 2 (Chapter 6-7) of the dissertation describes two works that empirically demonstrate that we can reliably study dyslexia and detect nuances between populations via volunteer-based online studies.

- Chapter 6 uses an online study that collects reading behavioral measurements to show the effectiveness of "Reader View," a tool that modifies web page layout and design to enhance readability, in improving the reading experience for people with and without dyslexia. This work was done jointly with Meredith Ringel Morris, Adam Fourney, Kevin Larson and Katharina Reinecke, and was previously published at CHI'19 [138].

- Chapter 7 introduces Virtual Chinrest, a novel method that makes it possible to automatically adjust online stimulus configurations according to each participant's viewing distance, producing reliable results for visual perception studies of dyslexia that are sensitive to display parameters. This work was completed in collaboration with Sung Jun Joo, Jason D. Yeatman and Katharina Reinecke, and was published in Scientific Reports'20 [136].

In Chapter 8, I summarize design implications for conducting online studies with people with disabilities based on lessons learned from different stakeholders – researchers, participants and healthcare professionals. I envision how the future of such online study platforms could fit into the overall support systems. I conclude in Chapter 9 by reviewing dissertation contributions.

Chapter 2

RELATED WORK

In this chapter, I discuss prior work on 1) the current practices and limitations when studying people with disabilities in laboratory studies as well as in online studies, 2) existing methods for conducting online studies that replicate prior findings, and 3) the barriers people with disabilities face when receiving support from professionals, and why they turn to online resources, including online studies. Some content in this chapter is taken and modified from previous publications [135,136,137,138].

2.1 Current Practices and Limitations When Studying People with Disabilities

Researchers usually strive to study large and representative samples to ensure generalizability and finding small effects. However, given that recruitment of specific populations is immensely difficult [53,176], most studies with people with disabilities and older adults are forced to rely on small numbers. Researchers often recruit through local organizations or advocacy groups [20], frequently establishing a local participant pool that can be used over time. For example, Johansson et al. [107] recruited participants with mental and cognitive disabilities through a local member-driven organization using snowball sampling. Similarly, the SiDE user pool [53,54] was established to facilitate accessibility studies with mostly elderly people. Researchers developed the SiDE pool for more than five years by travelling to different neighborhoods and repeatedly contacting potential participants and local communities. By 2014, 694 members from this pool had participated in one or more research studies. Maintenance of the user pool, however, requires several staff, much time and money [54].

Establishing such local participant pools is unavoidable if an experiment requires specific equipment or exhibits other characteristics that necessitate a supervised and controlled laboratory environment. For other experiments, researchers have developed and evalu-

ated alternative ways to recruit and study participants. For example, researchers increasingly recruit participants through online labor markets, such as Amazon Mechanical Turk (MTurk) [23, 35, 91]. Compared to traditional laboratory experiments, online studies offer faster and more effortless participant recruitment, as well as larger and more diverse samples [23, 80, 100, 148]. Despite initial concerns about the quality of data collected from unsupervised online workers, robust and validated data quality methodologies have been developed for conducting a broad range of experiments, yielding results comparable to those obtained in conventional laboratory settings [77, 80, 89, 121, 186].

Researchers have also used MTurk to conduct studies with people with disabilities and other specific populations. Tenenbaum et al. [223], for example, recruited 153 individuals with physical disabilities and studied how various impairment-related factors influence vocational self-efficacies. Carr [39] recruited 111 cancer survivors on MTurk and showed that the majority of participants (88.75%) were honest in their responses to a series of questions and return rates and test-retest reliabilities were high. Smith et al. [213] also concluded that MTurk is a good solution to sample hard-to-reach populations, such as people with low socioeconomic status, people with disabilities, or LGBT individuals.

While useful for researchers, there are also benefits of online experimentation for people with disabilities compared to participating in laboratory studies, such as the flexible time commitment, not having to rely on public transit, or being able to remain anonymous [33, 244]. People with disabilities will also often receive a sense of self-worth, self-efficacy and autonomy when participating in such studies [57]. They are motivated to contribute to scientific research [53] and to cognitively benefit from doing a task [33].

However, although online labor markets, such as MTurk, have been used to conduct accessibility research, researchers have found several usability problems that make it difficult to access for people with disabilities [37, 244]. For example, MTurk was found to violate multiple Web Content Accessibility Guidelines (WCAG 2.0) that may affect users with visual, cognitive, reading, physical or auditory disabilities [37, 222]. It is also often difficult for people with disabilities to identify which of the available tasks match their abilities, or to complete tasks within a specific time frame [244]. As a result of such barriers, the diversity of people on Mechanical Turk is still limited, both in terms of people with disabilities [37],

and in terms of age range [33].

2.2 Methods for Conducting Online Experiments

The push for online studies also evolves over time, especially being driven by increasing calls for large and more diverse participant sample and by the recent pandemic [12]. This transition from in-lab studies to online experiments raises concerns including but not limited to data quality (e.g. can online studies replicate results from previous in-lab studies?) and feasibility (e.g. what types of studies can we conduct reliably online?). In recent work, for example, Sauter et al. summarized the difficulties of conducting behavioral studies online and presented an overview of actively maintained solutions for the critical components of successfully online data acquisition: creating, hosting and recruiting [203]. A growing body of literature has also explored methodologies for conducting a broad range of experiments, and shown that online experiments yield results comparable to those obtained in conventional laboratory settings [15, 47, 52, 77, 135, 181, 182, 183, 186, 211, 243].

For instance, online experiments have been shown to accurately replicate the findings from behavioral experiments that rely on reaction time measurement [15, 47, 181, 182, 211, 243], rapid stimulus presentation [47, 77, 186] and learning tasks with complex instructions [47]. De Leeuw and Motz conducted a visual search experiment with interleaved trials implemented in both the Psychophysics Toolbox (in lab) and JavaScript (online) and showed that both software packages were equally sensitive to changes in response times [52]. Similarly, Reimers and Stewart demonstrated that two major ways of running experiments online, using Adobe Flash or JavaScript, can both be used to accurately detect differences in response times despite differences in browser types and system hardware (machines) [183]. Researchers have also investigated if web-based within-subjects experiments studying visual perception can accurately replicate prior laboratory results [89, 140]. These online experiments replicated prior laboratory results despite not being able to control for participants' viewing distance and angle.

2.3 *Why People With Disabilities Are Motivated to Participate in Online Studies*

Here I summarized the existing literature discussing the barriers people with mental and/or cognitive disabilities[1] face in receiving support and why they are motivated to participate in online tests.

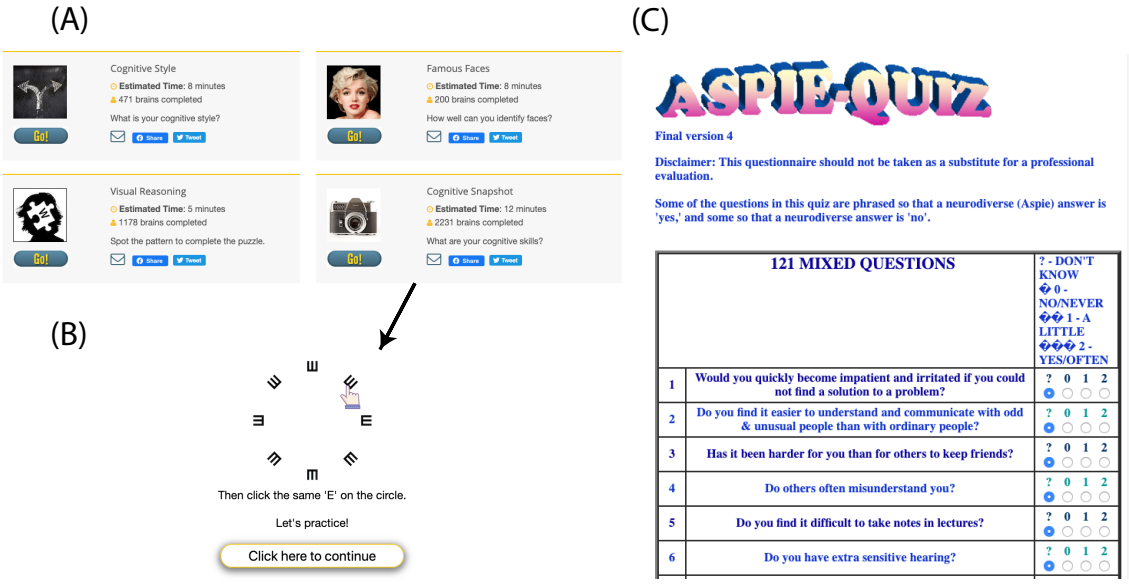


Figure 2.1: Examples of online tests that are used by people with cognitive or mental disabilities to assess themselves: (A) Examples of several cognitive assessment tests on TestMyBrain.org. (B) An example task in the "Cognitive Snapshot" test on TestMyBrain.org. (C) The Aspie-Quiz (`rdos.net/eng/Aspie-quiz.php`), a questionnaire developed by independent researcher Leif Ekblad. Its websites states that it evaluates neurodiverse traits in adults, which "can be used to give a reliable indication of autism spectrum traits prior to eventual diagnosis."

Common cognitive and mental disabilities include neurodevelopmental disorders, such as autism spectrum disorder (ASD) and attention deficit hyperactivity disorder (ADHD);

[1] A large subgroup of participants with disabilities who take online tests frequently [135]

mental disorders, such as borderline personality disorder (BPD) and depression; specific learning disorders, such as dyslexia and dyscalculia; and neurocognitive disorders caused by conditions like traumatic brain injury (TBI) [7]. Diagnosing these conditions is difficult due to imprecise diagnostic thresholds and high rates of comorbidity, which makes differentiating symptoms from co-existing cognitive impairments or medical conditions challenging [13,96, 154,231]. As a result, people often receive an insufficient explanation of their diagnoses and are frequently provided inadequate support and resources for interventions [96].

There are various reasons for why people may not seek professional help or receive a formal diagnosis. Common concerns include costs, the lack of insurance, unavailable or inconvenient care when needed, not knowing where to go, inadequate transportation, concerns about confidentiality and the belief that the treatment will not help [69,82,95,156]. Likewise, patients often feel that they can handle the symptoms themselves and do not consider their disorder as serious or recognize it as an illness [27,69,156]. Others refrain from acknowledging their disability due to public, perceived, and self-stigmatising attitudes towards mental conditions and cognitive disabilities. For instance, people with psychiatric disorders often feel embarrassed or uncomfortable to talk about their personal problems to others [239]. They have reservations towards talking to both strangers [235] and to people who they knew or knew they would have future dealings with [34,239].

People who suspect or know that they have a cognitive or mental disability frequently turn to online resources to receive more information, understand how their cognitive functions may affect their lives, and meet others with the same conditions [132,164]. Among these resources are online tests and assessments, which people with cognitive or mental disabilities (diagnosed or suspected) use to assess the severity of their cognitive impairment or compare their cognitive performance and behavioral functions to that of others [59,76,135]. Websites that offer such online tests (e.g., mybraintest.org, testmybrain.org, labinthewild.org, psychcentral.com) are often, but not always, based on scientific research and authored by healthcare professionals, yet are rarely suitable for diagnosing health conditions. Instead, they commonly serve the purposes of providing initial assessments and/or helping researchers study cognitive deficits, as exemplified in Figure 2.1. These tests assess behavioral and cognitive traits using either behavioral tasks or survey questions, followed

by a results page that tells participants where they stand.

2.4 LabintheWild

LabintheWild is an online experiment platform for conducting behavioral experiments and surveys with volunteers. Experiments enlist participants using short slogans, such as "Can we guess your age?", or "Test your social intelligence!". After completing an experiment, participants can view their personalized results to see how they compare to others. This personalized feedback is provided instead of financially compensating participants and serves four main purposes. First, it encourages participants to take part in experiments because it enables self-reflection and social comparison [94]. Second, it exposes participants to scientific concepts and increases their interest in research and scientific findings [167]. Third, it ensures data quality: Participants are intrinsically motivated to provide honest answers and exert themselves. Experiments conducted on LabintheWild and other volunteer-based experiment platforms produce reliable data that matches the quality of in-lab studies [77, 80, 186]. Fourth, the personalized feedback serves as a word-of-mouth recruitment tool, because participants share their results on social networking sites or other web pages [186].

LabintheWild avoids some of the limitations of paid online experiments by being openly available to anyone who wants to participate without having to sign up. This lowers the barrier for participation. There is also no need to collect identifiable participant information for reimbursement. As a result, volunteer-based online experiments such as those conducted on LabintheWild have the potential to recruit from over 3.2 billion people around the world who have Internet access [224]. Existing volunteer-based platforms have indeed proven to attract more diverse participant samples than in-lab experiments and those conducted on MTurk, with participants reporting wider age ranges, more diverse educational backgrounds, and a far more expansive geographic dispersion (see, e.g., [77, 85, 186] and Table 3.1).

Two previous studies on LabintheWild indicate the feasibility of conducting online experiments with volunteers who have a disability and/or who are elderly (included in Table 3.1): (1) A study of people's color differentiation ability, which showed that innate and acquired color vision deficiencies, but also situational lighting conditions, monitor settings, and demographics, can significantly impact how many colors on a given user interface someone can

distinguish [184]; (2) A study comparing human listening rates between sighted, low-vision, and blind people [32], which showed that the listening rate of visually impaired participants is significantly faster (334 words-per-minute) than the listening rate of sighted participants (297 words-per-minute) and that it increases with years of screen reader usage. In the next section, we build on this prior work to verify whether online experiments with volunteers are suitable for conducting high-quality, robust studies with people with disabilities and older adults.

Part I

UNDERSTANDING VOLUNTEER-BASED ONLINE STUDIES FROM DIFFERENT STAKEHOLDERS

Part I of this dissertation examines the viability of conducting volunteer-based online studies with people with cognitive disabilities from the perspectives of different stakeholders: Chapter 3 focuses on the researchers' point of view, demonstrating that volunteer-based online studies could attract sufficiently large numbers of participants with various disabilities, and they replicate and extend previous work that studied a variety of disabilities. Chapter 4 focuses on the perspective of participants, exploring how these studies contribute to people's support systems. Finally, Chapter 5 delves into healthcare professionals' opinions on taking the tests for people with cognitive disabilities.

Chapter 3

EXAMINING THE VIABILITY OF VOLUNTEER-BASED ONLINE TESTS FOR STUDYING COGNITIVE DISABILITIES

3.1 Introduction

The goal of this chapter is to evaluate the suitability of an alternative methodology for the recruitment and study of people with disabilities and older adults: volunteer-based online experiments. Such online experiments with volunteers are conducted on a variety of platforms (e.g., TestMyBrain.org, GamesWithWords.org, LabintheWild.org) and usually provide personalized performance feedback in exchange for study participation. Previous experiments conducted on LabintheWild have shown to attract more diverse participants than laboratory studies and those conducted on Mechanical Turk in terms of age, education level, and geographic distribution [186]. Obtaining larger and more diverse sample sizes could extend the findings of smaller-scale laboratory studies, enable us to measure the variability between people with specific disabilities and of various ages, and verify results with people from diverse demographic backgrounds. However, it remains unknown (1) whether volunteer-based online experiments attract sufficiently large numbers of participants with disabilities and older adults to robustly conduct comparative studies, and (2) why participants with disabilities participate in such studies. Knowing their motivations and needs may shed light on how online experiments should be designed to attract large samples and provide adequately rewarding and engaging experiences for these populations.

To answer these questions, we replicated four laboratory studies on LabintheWild, all of which offered tasks that were known to be impacted by various disabilities or age-related decline. All four experiments attracted people of diverse ages and with various disabilities. Of 355,656 participants that took part in the studies, 4,799 (1.35%) participants self-reported to have some kind of impairment; an additional 7,564 (2.25%) participants were above age 65. Using the data that we collected, we replicate and extend previous work that studied

dyslexia, cognitive decline, autism, and motor impairments.

To better understand the motivations and needs of participants with disabilities, we further analyzed the comments that some of them voluntarily provided at the end of LabintheWild experiments and forum entries that discussed LabintheWild experiments as related to various disabilities. The results suggest that LabintheWild attracts people with disabilities because it provides personalized performance feedback and social comparison at the end of its studies: Participants use the experiments to diagnose or confirm a suspected disability, or to test its severity or impact on other situations and tasks in daily life by comparing their performance to others. Based on these findings, we contribute design implications for online experiment platforms that better support these needs.

3.2 Study Replication on LabintheWild

We replicated four studies on LabintheWild, chosen to represent a broad range of tasks (see Table 3.1) and modified to suit an uncontrolled online environment as described in each study section. None of these studies were specifically targeted at people with disabilities, but open to anyone to participate. All studies were advertised on LabintheWild with a slogan and provided personalized feedback at the end of the experiment.

3.2.1 Study 1: Weather Prediction Study

Our first study is a modification of Knowlton et al.'s study from 1994 [120], known as the "Weather Prediction Task". The probabilistic classification learning task was developed to show that humans' implicit memory and explicit memory systems contribute to procedural learning skills at different stages. In contrast to the explicit (declarative) memory system, human's implicit (non-declarative) memory does not require conscious thought and is used in early stages of the procedural learning process.

Several researchers have since then shown that people with neuro-developmental disorders such as Tourette syndrome, Schizophrenia, or developmental dyslexia, perform less well in the Weather Prediction Task than non-disabled participants [72, 114, 115, 146]. In particular, Gabay et al. [72] showed that adults with dyslexia performed better in the Weather

Table 3.1: Overview of LabintheWild experiments that can be related to specific disabilities or age-related decline. The first four are presented in this paper. Sample sizes are the final numbers used in the analysis. * denotes that participants were not asked about their disabilities, but chose to mention them in comments at the end of an experiment (hence the lower numbers).

Study Name (citing original study, if any)	**Slogan**	**Related Disabilities**	**# Months online**	**Matched sample size**	**# of participants with disabilities**	**% female**	**age range (mean age, stdev)**
Weather Prediction [120]	How quickly do you learn?	Amnesia [120], Dyslexia [72], Tourette syndrome [115, 146]	22	3,786	328 (8.66%)	52.76	5-99 (m=25, sd=11.5)
Memory [218]	How fast is your memory?	Cognitive decline in elderly people [228]	40	18,026	173 (0.96%) above 65 years, 26 (0.14%) with disability*	N/A	6-99 (m=25, sd=13.5)
Social Intelligence [16]	Test your social intelligence!	High-functioning Autism, Asperger's Syndrome [16]	10	123,928	3,368 (2.72%), 75 (0.06%) with Autism or Asperger's Syndrome	48	4-98 (m=27, sd=12.5)
Fitt's Law	Can we guess your age?	Motor impairments due to age-related decline [110, 116, 233]	4	209,916	5685 (2.71%) above 65 years, 1077 (0.51%) with disability*	33.3	20-85 (m=35, sd=11.9)
Listening Rate [32]	How fast can you process words?	Vision Impairment [17]	2	453	143 (32%)	57.83	8-80 (m=34, sd=15)
Colorblindness [184]	Can we guess your color age?	Color vision deficiencies [26]	12	31,248	1,831 (3.85%)	70.6	5-94 (m=30, sd=15.2)

Prediction Task as the training extended, but overall they performed significantly less well than matched controls. This is the main result that we aim to replicate.

Procedure Just like in the original task, participants in our LabintheWild experiment were shown a series of cards displaying one of four particular geometric designs (circles, diamonds, squares, or triangles). Each trial showed one or more of these cards together (Figure 3.1a). Each geometric design was previously assigned to a particular weather outcome (rainy or sunny). Participants were asked to predict whether the cards suggested that the weather would be rainy or sunny. While participants had to guess at the beginning, they could learn over time from feedback showing whether their responses were correct or not.

After presenting an informed consent form and a demographic questionnaire, the experiment started with five practice trials, followed by 80 experimental trials (as opposed to 150 trials in [72]) that each elicited a participant's response to one or more of the four types of cards, followed by feedback on whether the response was correct or incorrect. The 80 trials were evenly divided into four blocks. Participants were then presented with a personalized results page showing their performance (forecasting accuracy in percent) in comparison to others. They also received a written explanation about the purpose of the experiment and about the meaning of implicit memory in daily life. Completion of the experiment took approximately 10 minutes.

Participants

Over the course of 22 months, 3,786 participants completed the experiment, ranging in age from 5–99 years (m=25, sd=11.5). Roughly half (52.75%) of participants were female. Asked whether they had any cognitive or neurological disabilities, 328 (8.66%) answered in the affirmative. 223 (68%) of those who answered yes provided details about their disability in an open-ended box provided underneath the question. The most common cognitive disabilities were Attention-deficit (Hyperactivity) Disorder (ADD/ADHD) (N=103, 2.7%), Dyslexia (N=81, 2.14%), Autism Spectrum Disorder (ASD) or Asperger's Syndrome (N=62, 1.64%), and Depression (N=30, 0.8%).

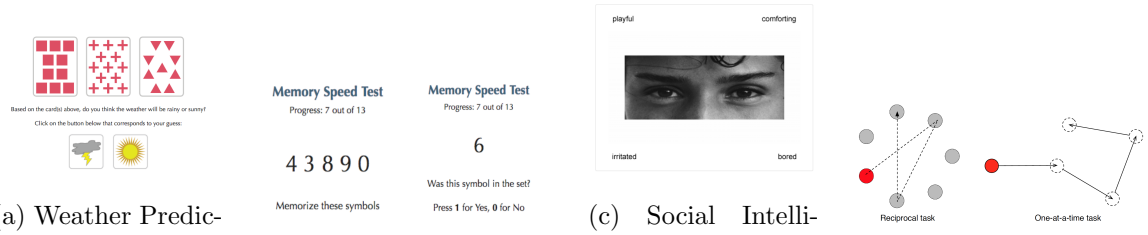
Analysis

For analysis, we first excluded 319 (8.4%) participants who self-reported that they had taken the test before, seven participants who achieved a zero percentage correct rate in at least one of the four blocks, 435 (11.2%) people who did not answer the question on whether they have any cognitive disabilities, and 226 participants who reported having one or more cognitive disabilities other than dyslexia, to further control for effects of other cognitive disabilities. We included participants aged 18 - 30 to match the age distribution of the participants from Gabay et al.'s [72] study (N=30, age range 18-30). The final number of participants was 1,654, including 46 (2.8%) who indicated having dyslexia.

Following the analysis procedure presented in [72], we conducted an ANOVA comparing the performance of non-disabled participants with those participants who self-reported having dyslexia. We modeled Block (trials 1-20, 21-40, 41-60, 61-80) as a within-subject factor, Dyslexia (dyslexia vs. non-dyslexia) as a between-subject factor, and a Dyslexia by Block interaction. Mean proportion of correct answers was the dependent variable.

Results

Our results showed a significant main effect of Block ($F(3, 6608) = 10.62, p < .001$), suggesting that for participants with dyslexia and for those without, accuracy improved as the training extended. This confirms previous results that all participants learned gradually to associate cues with the appropriate outcome [72,120]. Our results also showed a main effect of Dyslexia ($F(1, 6608) = 6.90, p < .01$, Cohen's d = .17). Participants with dyslexia overall achieved a significantly less accurate forecasting accuracy (m=55%, sd=12%) than those without dyslexia (m=57%, sd=11%, independent two-tailed t-test: $t_{(139)} = 2.57, p = 0.01$), confirming [72]. However, in contrast to Gabay et al.s finding [72], there was no Dyslexia by Block interaction effect ($F(3, 6608) = 1.35, p = .26$), meaning that both people with dyslexia and those without learned the probabilistic relationships at similar pace. Our findings extend prior work by indicating that the difference in pace between dyslexics and controls found by Gabay et al. might not hold for all people with dyslexia.



(a) Weather Prediction Study (b) Memory Study (c) Social Intelligence Study (d) Fitt's Law Study

Figure 3.1: Overview of the stimuli used in four of our LabintheWild experiments.

3.2.2 Study 2: Memory Study

Our second study is a replication of Sternberg's experiment [218], which demonstrated that the response time of retrieving an item from working memory is linearly proportional to the number of items stored in memory. Follow-up work demonstrated cognitive decline in elderly people, i.e. that the reaction time of elderly participants increases at a higher rate with the number of items held in working memory than is the case for younger people [228]. This is the main result that we aim to replicate. The finding is supported by the so-called "complexity effect", which implies that when the complexity of a task increases, performance differences between young and elderly people become larger [174].

Procedure The experiment began with an overview of the study, an informed consent form, and an optional demographic questionnaire, followed by the main experiment consisting of 12 experimental blocks. Each block presented a sequence of 1-6 randomly chosen symbols (digits and uppercase English alphabet letters) to memorize, followed by 3 positive (i.e., containing a symbol from the original set) and 8 negative probes, in random order (Figure 3.1b). Participants were asked to determine whether the symbol in the probe had been part of the original sequence. Each set size between 1 and 6 was represented in two experimental blocks (resulting in a total of 12 blocks). The experiment took 8 minutes to complete.

Table 3.2: Linear regression predicting reaction time for participants who are 22 or older from age, set size, and their interaction in the Memory study. Adjusted $R^2 = 0.1645, p < .0001, F(3, 828857) = 54380, p < .0001$.

Variable	**Est.**	**SE**	**t-value**	$Pr(>\|t\|)$	
(Intercept)	334.97	2.10	159.76	$< .001$	***
set_size	66.53	0.54	122.71	$< .001$	***
age	5.15	0.06	88.14	$< .001$	***
age × set_size	0.04	0.02	2.70	< 0.01	***

Participants

Over the course of around 40 months, 18,026 participants completed the study (see also Table 3.1). They ranged in age from 6 to 99 (m=25 years, sd=13.5). Since we were interested in age-related cognitive decline, we did not ask any question related to potential disabilities (but 26 participants voluntarily commented having a disability that they thought might explain the performance in this task).

Results

To prepare the data for analysis, we excluded 1,438 (8%) participants who indicated having technical difficulties or having cheated. We then excluded trials that resulted in extreme outliers of response time, computed as the median + 3×IQR (2111 ms), which might indicate a distraction from the test.

To analyze whether the slope increase in reaction time across set sizes is steeper for elderly participants than for young ones (which would confirm the complexity affect), we conducted a multiple linear regression with reaction time as the dependent variable and age, set size, and an interaction effect between age and set size as independent variables. We included 7,363 participants aged 22 or older (because performance peaks at about age 21 and our aim was to model age-related decline). We modeled both set size and age as continuous variables given that our large number of participants allowed us to analyze the

effect for all ages (rather than binning them into discrete "young" and "old" groups as in [228]).

Our results show that the reaction time increases with set size ($\beta = 66.53, t = 122.71, p < .001$), confirming Sternberg's original results of the test [218]. In addition, we found a significant interaction effect between age and set size (Table 3.2), demonstrating the complexity effect [174]. This confirms the results of [228] who found that reaction time increases at a higher rate as a function of memory load in elderly than in young people. We extend this result by showing that this interaction effect is true for ages 22-99.

3.2.3 Study 3: Social Intelligence Study

Originally developed by Baron-Cohen et al. as the "Reading the Mind in the Eyes" test [16], the study showed that compared to non-disabled adults, people with Asperger's Syndrome or High-functioning Autism were less likely to recognize people's emotions by looking at images of their eyes. This is the main result that we aim to replicate.

Procedure Participants first saw a brief description of the study, agreed to the informed consent, and were then presented with instructions of the task. They were given a practice trial that included feedback on the accuracy of their response. Just like in the original task, participants in our LabintheWild experiment were shown 36 trials, each showing one image containing only a person's eyes (Figure 3.1c). For each image, participants were asked to choose one of four words that best expresses the emotion the eyes are showing. At the end of the study, participants were shown their number of correct answers compared to the average of 26 as reported in the original study [16]. Completion of the experiment took approximately 8 minutes.

Participants

131,840 participants completed the study within ten months. We excluded 7,912 participants who had taken the study before for a total of 123,928 participants. Participants were between 4-98 years old (m=29.5, sd=12.2), with 48% identifying as female. In our demographics questionnaire, 3,368 (2.72%) participants disclosed having at least one type

of disability. We included 75 (0.06%) participants who mentioned having Autism spectrum disorder (ASD) or Asperger's Syndrome (a milder form of ASD) in our study, excluding the remaining participants with disabilities.

Results

We conducted an independent two-tailed t-test to compare the performance of people with ASD to those without. Participants with ASD received significantly lower scores (m=22.92, sd=4.63) than non-disabled participants (m=26.29, sd=4.60, $t_{(74.039)} = 4.23, p < .001$, Cohen's d = .73). The results confirm those of Baron-Cohen et al. [16], who had found means of 22 (sd=6.6) for ASD participants (N=15) and 26.2 (sd=3.6) for non-disabled controls.

3.2.4 Study 4: Fitt's Law Study

Our last study was designed to study age-related effects of pointing performance using the ISO 9241-9 standard Fitt's Law task [99]. Much prior work has found that people's pointing performance when using a mouse declines with age. For example, older adults have lower peak speeds (the maximum speed during a movement) than younger adults [110], they have longer verification times (the time interval between the end of a movement inside a target and the beginning of the click) [233], they make more pauses of 100ms [110] and their normalized jerk (fluctuations in the speed profile of the movement) is higher [116]. Our goal is to replicate these results.

Procedure Participants were first asked to agree to an informed consent form and read through brief instructions, which included a request to perform the pointing tasks as quickly and accurately as possible. They were then presented with a total of 80 trials, divided into ten blocks, in which they had to perform five tasks each of the following two types: (1) Reciprocal tasks, in which targets were arranged in a circle, and subsequent targets appeared in red in a predictable manner (this was based on the ISO 9241-9 standard [99]), and (2) one-at-a-time tasks, in which only one target was visible at a time. Subsequent targets appeared in a random direction (Figure 3.1d). Target sizes (10, 15, 25, 40, and 60 pixels)

Table 3.3: Previously reported age-related decline of motor performance and the results of our Fitt's Law study.

Finding	**Our Results**	**Supported?**
Older adults have lower peak speeds than young adults [110].	$\beta = -0.0026, F_{1,209913} = 15370, p < 0.0001$	Yes
Older adults have longer verification times than young adults [233].	$\beta = 0.0422, F_{1,204509} = 46634, p < 0.0001$	Yes
Older adults make more pauses of 100ms than young adults [110].	$\beta = 0.0016, F_{1,204512} = 11116, p < 0.0001$	Yes
Normalized jerk is higher for older adults than for young adults [116].	$\beta = 0.0039, F_{1,204419} = 14349, p < 0.0001$	Yes

and distance between targets (75–400 pixels) were varied between tasks. After completing the ten blocks, participants were presented with a demographics questionnaire before seeing their personalized results. The results included a "guess" of their age, predicted with the help of a linear regression model that included several movement features of participants from a previous dataset. Participants were able to reveal their actual age underneath the prediction on the results page. Completion of the experiment took approximately 5 minutes.

Participants

More than 540,000 participants completed the experiment within four months that it was online. To match the sample to those in prior work, we only report on 209,916 participants who used a mouse, who revealed their actual age after seeing our predicted age, and who were between 20-85 years old. The resulting sample had a mean age of 35 (sd=11.9) with 33% female. Young adults were well-represented (e.g., over 10,000 individuals aged 27); the least represented were subjects at age 84 (N=24).

Results

We conducted multiple linear regressions with age modeled as a continuous independent variable and the dependent variable being our different measures of interest. All movement variables were calculated following the procedures in related work [110,116,233]. Table 3.3 shows that all results were consistent with prior work. We additionally extend prior results by showing a continuous age-related decline of motor abilities between ages 20-85.

3.2.5 Summary

Our results show that LabintheWild studies accurately replicate the main findings of prior laboratory studies with larger samples of specific disabilities and ages. We also extended previous work with novel findings that were made possible because of the large scale and more diverse samples. Together, these results suggest that conducting volunteer-based online experiments is a suitable methodology for efficiently studying older adults and people with disabilities.

When developing LabintheWild, we made specific design decisions, such as foregoing the necessity for people to create an account (thus avoiding a sign-up barrier and preserving anonymity), and providing social comparison and sharing opportunities at the end of each study. As we will show in the next section, these design decisions are among the main reasons why LabintheWild attracts large numbers of participants, including people with disabilities.

3.3 Motivations and Needs of Participants with Disabilities

We were additionally interested in finding out what motivates participants with disabilities to take part and what their needs are. Knowing this can lead to insights into how to better design online experiment platforms for this particular population to ensure their continued, and perhaps increased participation, and to ensure that such experiments are mutually beneficial to participants and researchers. We therefore analyzed

1. *comments* that participants provided voluntarily at the end of LabintheWild experiments in response to a generic question, such as "Do you have any comments or

feedback?". We included comments from the six experiments listed in Table 3.1 if they were either made by a participant who self-reported having an impairment, or if the comment itself revealed details about an impairment that may have not been asked about in the demographics form.

2. *forum entries* that discuss LabintheWild experiments. A broad search of LabintheWild mentions on social networking sites, forums, and via search engines revealed 10 platforms that include discussions of LabintheWild experiments in 16 different forums (see Table 3.4).

To find out what the main needs and motivations are that participants share in the comments and on external forums, two researchers generated initial codes for a subset of comments and forum entries, discussed the codes, and then coded all entries. We then iteratively clustered codes into themes following the thematic analysis method [79].

Some of the quotes presented below have been slightly modified for readability. If the quote was taken from a comment in an LabintheWild experiment, the specific experiment and participant number is noted in brackets (with the exception of the Listening Rate Test, which only recorded session IDs).

3.3.1 Results

The analysis revealed four major themes related to participants diagnosing a disability, comparing results, experimenting, and explaining results to themselves and to the researchers.

Testing the Effects of a (Suspected) Disability Our first theme showed that many participants interpret their performance in the context of their (suspected) impairment. Many forum entries and comments in LabintheWild experiments suggest that people either have received a medical diagnosis of their disability but are unsure whether it affects other functions, or they suspect they might have a disability and are trying to find out if that is indeed the case. An example of the latter is an entry in the FitMisc forum, a forum about "Fitness, Memes & Motivation", in which one user started a thread with the title "SRS ANSWERS, how do I know if I'm autistic?" and then added several follow-up posts: "Srsly

how do I know this?" and "But like what is the science behind it? Is there mild autism, are there different subgenres etc?". Responses to these questions were overwhelmingly sarcastic, but one forum user responded with a link to LabintheWild's Social Intelligence test:

> Take this test: http://socialintelligence.labinthewild.org/. You have to guess what emotions the picture of eyes are showing, as that provides an indicator of your social intelligence and if you have an autism spectrum disorder.

The fact that the test has previously been connected to autism is not actually revealed in the LabintheWild version, showing that participants sometimes make these connections either based on prior knowledge or based on their own assumptions. To confirm a suspected disability, many forum users seem to appreciate having been provided such links to LabintheWild studies. For example, in response to seeing a link to the same test on the Furaffinity forum, a user wrote:

> okay I'm taking this. For the record I recently learned about autism and I'm like 99% sure that I am on the spectrum. I can't afford to go to a psychologist but the evidence from my infancy through my childhood and now in my adult life screams autism or something. For so long I struggled with myself not knowing why I seemed very...delayed emotionally and socially and I have ADHD and sensory disorder and disassociate as well. so it just all came together. Of course, I am a girl, so it goes widely unnoticed in quiet little girls.

This particular user later revealed the score she received as 25/36, which is slightly lower than the average result of 26.4 that non-disabled female participants achieved in Baron-Cohen et al.'s original study [16], but higher than the average score of 21.9 that was found for participants with Asperger's syndrome, which is a form of high-functioning autism.

As mentioned above, other participants are often certain that they have a disability, perhaps because they have previously received a medical diagnosis. Despite knowing about it, their comments frequently indicated that they are unsure what other functions their

disability might affect. They take LabintheWild experiments to test these boundaries. For example, after completing the Weather Prediction Study and seeing his results, one participant wrote:

> One of the more exciting tests! [...] I did take longer than average to learn, according to the results graph; I wonder if this has anything to do with my ASD. (Weather Prediction Study, P2888)

A participant in the Colorblindness Test additionally explained her results by reporting on a previous accident and subsequent surgery that she suspected had impacted her color vision:

> I had damage to my retina due to a car accident, airbag in the face. I had a retinal peel and then eight months later that caused cataracts so I had a lens implant. I have noticed my colour vision is less perfect than before, and I still have a blind spot in the macula. Combine that with a lousy and very old monitor and poor indoor light (circuit breaker is out so I can't turn on another light – well, I'm not as good as I used to be. (Colorblindness, P15177)

We observed a similar kind of sense-making and using results to test a disability in forums. Related to this, participants also publicly discuss and compare their results to others' as described in the next section.

Comparison of Results To test the effects and severity of their disability, our analysis showed that participants desire comparisons to others with similar diagnoses. For this, they turn to external forums (see Table 3.4). Most forum threads that discuss LabintheWild experiments start with someone posting a link to a specific test, often proposing that the test might be relevant to people with a specific disability. For example, in r/ADHD, Reddit's ADHD subreddit, one poster wrote:

> Do you focus on the big picture or the fine details? Online psychology test (X-Post from r/Psychology, thought it'd be interesting to see ADHDer results!)

Table 3.4: List of forums that discuss LabintheWild experiments and thread lengths. * indicates experiments that were not designed to test and have not previously been found to relate to a disability.

Website	**Forum**	**LabintheWild experiment discussed**	**# of replies**
Crunchyroll.com	Autism	Social Intelligence Test	81
Elkoy.org	Autism	Social Intelligence Test	23
Fitmisc.net	Autism	Social Intelligence Test	59
Furaffinity.net	Autism	Social Intelligence Test	42
Reddit.com	ADHD	Frame-Line Test*	63
Reddit.com	Autism	Frame-Line Test*	38
Reddit.com	BPD	Social Intelligence Test	49
Reddit.com	Sociopath	Social Intelligence Test	38
Psychforums.com	Narcissism	Social Intelligence Test	12
Schizophrenia.com	Schizophrenia	Multitasking Test*	8
Schizophrenia.com	Schizophrenia	Thinking Style Test*	6
Schizophrenia.com	Schizophrenia	Listening Rate Test	8
Supforums.com	Autism	Social Intelligence Test	128
Testyourmight.com	Asperger's Syndrome	Social Intelligence Test	68
Wrongplanet.net	Autism	Social Intelligence Test	12
Wrongplanet.net	Autism	Multitasking Test*	18

The test that this person was referring to is LabintheWild's Frame-Line test, advertised on LabintheWild as "Are you more Eastern or Western?" because it has previously been shown to detect cross-cultural differences in perception between people in the U.S. and Japan [119]. On Reddit, the original poster later explained why they thought this test might relate to ADHD:

> So I bet there is more differences culturally here, but it's been said many times that ADHD causes difficulty focusing on small details, I wonder if this test shows up that difference.

Participants answered by posting their own results, such as "I got like a 72 on the first one and a 42 on the second..." or "100 on the first and 61 on the second. Really interesting". Forums usually contain long chains of replies from other forum participants who share their own scores. The majority of them are shared without further comments, but forum participants occasionally add further details, such as this post in r/ADHD:

> Fascinating. I got a perfect score of 100 on the first test and a crappy 42 on the second test. I'm not surprised, I already knew I suck at judging absolute length, I didn't know I was so good on relative length, though. Given that I am autistic I was expecting the exact opposite result.

A series of posts on various forums further showed how participants openly reveal having received relatively low scores, probably benefiting from the anonymous environment of such forums. A user on Furaffinity's Autism forum, for example, described their score and experience with the Social Intelligence test the following way:

> 2/36 [...] I logged in just to say, wow this is impressive. I had no idea you can tell how the person feels just by looking at their eyes. All I could tell was the people in that test were looking at something, people have different eye shapes and different ways of looking (i.e some turn their eyes, some their head, some do both when looking at something at their side). [...]

Another user shared a similar experience when replying to others' scores in the Social Intelligence test on r/Sociopath:

> 14 [out of 36], they all looked the same only ones i could tell were when the eyebrows were heavily impressioned.

In response to this post, another person offered a potential diagnosis by saying "Do you have autism? They have low cognitive empathy whereas sociopaths have low affective empathy."

Providing such interpretations of other people's scores was common in forums. Responding to a user in the Schizophrenia forum who posted their results in the Multitasking Test, another user wrote:

> Interesting that your attention on the clicking task was very different from the average. You slowed down much less than me when remembering multiple symbols. Both of us were below average for that, with me being considerably so. I hope others will try this.

On r/BPD, the subreddit for Bipolar Personality Disorder, users also discussed and questioned previous medical diagnoses because they seemed to contradict their performance in an experiment. For example, one user wrote, referring to their result in the Social Intelligence test "Wow, I got 35 out of 36. [...] I'm surprised because I have Aspergers and find it difficult to read people.", to which someone else replied:

> Are you sure the Aspergers isn't a misdiagnosis? I was misdiagnosed with it for a while. BPD and ASD sometimes have superficial similarities, but since ASD is characterized by underdeveloped theory of mind and BPD is characterized by overdeveloped theory of mind, I'm not sure someone can really be both. (I could of course be wrong though.)

Several forums that included comparisons between results also contained entries that summarized the results. For example, one poster in r/sociopath responded to a question

"whether sociopaths score higher or lower than average on this test" with "Looks like we are all over the place, and it depends more on the person." Similarly, a post in r/ADHD about the LabintheWild Frame-Line task contained a score and a general assessment of how that compares to others in the subreddit:

> 97 vs 48. That's quite a big difference, but it's quite similar to what you guys report. Seems like there is a difference between members of this subreddit and the general population.

In addition, some of the entries revealed a desire to find such opportunities of comparison to other people with a similar disability on LabintheWild itself. For instance, a user discussing the Frame-Line test in the ADHD subreddit wrote:

> [...] i would be interested in seeing scores broken down by people with ADHD and comparing it across countries. if they are hypothesizing culture affects perception, i'd be interested in seeing if it correlated with how severe one's ADHD is perceived as well as if overall the big picture vs. details was more strongly linked to culture than an ADHD diagnosis.

Self-experimentation Those participants who seemed to be sure of their disability frequently suspected that interventions, such as hearing aids or medication, could change their performance. A participant in the Memory Test, for instance, suggested that their lack of medication might have affected his results:

> Because my ADHD caused my reactions to be more jittery reactions and trigger happy sensations rather than me not knowing. [...] So I think me not being medicated for ADHD was my problem and the data could be fixed if i were to be properly medicated by a practitioner. (Memory, P16924)

A blind and hearing-impaired participant in the Listening Rate test more directly indicated an interest in testing her ability to understand text at different listening speeds with and without her hearing aids:

> [...] I did this without my hearing aids. I think it would be interesting to see how I'd do if I'd chosen to put in my aids before doing this. (Listening Rate)

That participants try to make sense of their disability via self-experimentation was also occasionally the case in forum entries. In the ADHD subreddit, for instance, a user responded to other people's scores in the Frame-Line test:

> just took it again after meds. first time: 77 (big picture) vs 45 (details) second time (after meds): 100 (big picture) vs 37 (details). I would say the big picture relative test is probably easier as it's relative lengths, but I was surprised that my score on details went down. of course there's a lot of bias. namely i've already taken the test so i've had practice and maybe knowing my details score was low the first time i over-corrected [...] edit: also just glancing at other people's scores, it looks like the really high scorers on big picture (¿90) seem to have a much larger gap between their details score than more "average" scorers. of course i'm just grasping straws. is interesting though!

Providing Context to Explain Results Another frequent theme that we discovered was that participants indirectly put their performance results in context by providing much detail on their disability. Two participants in the Memory Test, for example, talked about their short-term memory loss and their strategies for compensating:

> At 16 years of age I suffered an indented fracture of the skull which caused ongoing short term memory loss. I have had to compensate by committing tasks to lists. This has enabled me to excel in my career, IT. To let go of the enormous effort to recall from memory has enabled me to achieve through focusing on innovation. (Memory, P12312)

> I have a medical condition that causes short term memory deficit. I am studying Mandarin Chinese to exercise my brain. In the very short term, like your test, I think I do okay, but having to go back to previous sets or longer range of time; items get lost easier from my memory... (Memory, P7344)

The latter comment also relates to one of our previous themes, that participants often use LabintheWild experiments to test the extent of their disability.

Participants shared similar details in forums, where they often used descriptions of their condition to explain their scores to others. In r/psychology, for example, one user wrote:

> [...] I have the big picture appreciation as well and I too suck at long term planning. It usually means I easily form a grand idea of how something should be but it's too abstract for me to actually be able to make a plan and execute :/

Participants also frequently provided seemingly unrelated information that put the results in context. After participating in the Memory Test, a female participant commented:

> I used to have major depression and suicidal tendencies. The outcome [of medication] received for years of depression was chemical imbalances in my body and symptoms similar to Anhedonia. Activities from yesterday feels like a dream. Tangible memories become vague. Taste and smell senses are not clear, I can hardly taste food flavor and I usually need to focus hard to figure out the flavors. (Memory, P10269)

Apart from putting their own performance in context, participants comments also suggest their desire to share details about their condition and engage in a conversation. A participant in the Social Intelligence Test, for example, asked:

> I think I did well... I would like to do bad. I've Aspergers and am supposed to be bad at recognizing that kind to things... But I am as well a painter and depend on being able to paint - for example - expressive eyes. But what if my Aspergers diagnosis is wrong??? (Social Intelligence, P53892)

The comment also indicates the participant's struggle with their medical diagnosis and their need for further confirmation. We observed a similar need to receive advice in other participants' comments and on forums.

3.4 Discussion and Design Implications

The goal of our work was to validate online experimentation with volunteers as an alternative methodology for conducting studies with people with disabilities and elderly people. Analyzing four replication studies conducted on LabintheWild, we showed that these studies attract people with a range of different disabilities and ages. The results of these experiments confirm and extend previous laboratory results, suggesting that LabintheWild experiments studying people with disabilities and older adults result in high data quality. Our studies show that volunteer-based online experiments are a viable methodology for conducting such experiments.

However, our experience with these studies also revealed room for improvement. Recruiting larger numbers of participants with disabilities and elderly people in a shorter amount of time would be desirable, as would be a more targeted recruitment of people with specific types of disabilities and age groups. Doing so will require designing inclusive online experiment platforms that support people with these characteristics and provide them with rewarding experience.

As a first step in this direction, we investigated why participants with disabilities currently take part in the studies and how they could be better supported. The most prominent finding of this analysis is that participants search for and use LabintheWild experiments as diagnostic tools. With the help of the experiments, participants test whether they have a disability and, if they are already aware of a specific disability, they test its severity and what other tasks and situations it might affect. In many cases, the experiments that participants used for such self-experimentation and comparison were not actually designed to test a disability; instead, participants hoped to find out whether such seemingly unrelated tasks might also be affected by a specific condition. These results show that the personalized feedback and opportunity for social comparison at the end of each study are key reasons why people with disabilities are attracted to LabintheWild.

Our findings point to a number of potential improvements for the design of LabintheWild and other online experimentation platforms:

3.4.1 Validate Experiments for Specific Disabilities

A risk of using LabintheWild's experiments as diagnostic tools is that participants might read too much into their results and potentially misdiagnose themselves or others. This risk is especially severe given that participants often use experiments that have not been previously found to relate to a disability as tools for assessment. To address participants' need for diagnosing themselves and testing the severity of their disability, it will be essential to provide validated tests. Such experiments could be existing ones that have proven to be reliable for assessing specific disabilities. Validated tests could also be developed on demand. In fact, an exciting future avenue would be to enable participants to state the need for specific tests. Researchers could point them to existing resources (e.g., via a library of experiments related to specific disabilities) or develop new experiments that address this need.

3.4.2 Support Comparison to Specific Groups

Our analysis also revealed a desire to receive personalized feedback that allows specific comparison to others with a similar disability. Instead of providing comparisons to the average person, as is currently common on LabintheWild and other volunteer-based online experiment platforms, the personalized feedback could be presented with a choice of a comparison group. Of course, this requires bootstrapping the data with sufficient results from a specific group, which may or may not be available from prior literature. One solution would be an integrated model, in which participants who want to compare themselves to a specific group can recruit others, and results are then communicated back to anyone who signed up to receive specific results about this group with a time-delay.

3.4.3 Allow for Self-experimentation

Similarly, we found that participants frequently use LabintheWild experiments for self-experimentation, both longitudinally and within a short time frame, such as before and after taking specific medications. To support this, online experiment platforms should facilitate keeping test results from multiple study runs of the same participant and providing access

to a personal profile that allows reviewing these results. While many volunteer-based online experiment platforms refrain from using log-ins, participants who are interested in having access to such profiles could create a (privacy-preserving) account after participation.

3.4.4 Involve Participants in the Recruitment

Our four example studies demonstrated the feasibility of serendipitously recruiting diverse people with disabilities, but it would be desirable to facilitate more targeted recruitment of people with specific disabilities to increase sample sizes in shorter amounts of time. We showed evidence that people recruit each other through forums; but reaching these forums in the first place is a challenge. Online experiment platforms could work with specific populations to achieve this aim, similar to what we described above: Previous participants could be encouraged to recruit others with similar disabilities through their social networks and specific forums. Reward mechanisms could be the subsequent possibility of comparing to others, contributing to science, gaining a sense of self-worth [57], or co-authorships offered for involvement in the larger research cycle, similar to Stanford's crowd research project [225].

3.4.5 Provide Opportunities for Discussion

Participants occasionally mentioned not having a physician, psychologist, or psychiatrist to turn to. They therefore turn to LabintheWild experiments to assess a disability, risking misdiagnosing themselves as mentioned above, but also risking being left alone with results that might be perceived as troubling. Many online experiments providing personalized results therefore add disclaimers on their results pages that state the purpose of a specific test and that it should not be used for medical diagnoses. However, our analysis suggests that the problem is not that participants are not aware that experiments are often designed for a different purpose or insufficient tools for medical diagnoses, but that they have an otherwise unsatisfied need for finding out where they stand and how their disability relates to other tasks. Embracing such experimentation at a personal and at a community level would be a better approach.

We also found that the common one-way communication when participants leave comments insufficiently addresses participants' need for dialog. This finding supports previous

work [167], which found that online experiment volunteers often use the comment box to start a dialog with the researchers — usually to inquire about the research background of a study or its goals. We extend this finding by showing that people with disabilities additionally use the comment boxes to report on specific medical diagnoses, life events, or compensation strategies. The motivation behind this is two-fold: participants explain their performance in a given experiment to themselves and to the researcher, but they also commonly seem to share this information to simply talk to someone.

The need for bi-directional communication with the researchers is currently unsupported in most, if not all, online experiments. In our eyes, it raises the urgent question of how researchers can provide answers, support, and debriefing information to the large numbers of participants in online experiments who may need it. To address this, Oliveira et al. [167] suggested an internal forum where participants and researchers can exchange their thought and ideas with others. But there are problems with this approach specific to people with disabilities: First, some of the comments might have to be answered by an expert with specific (medical) expertise, who may or may not be the researcher or other participants. Second, our analysis of external discussion forums suggests that participants feel comfortable revealing and discussing their disabilities and experiment results within their community, such as within a subreddit on a specific disability. If online experiment platforms provided internal forums, they should therefore enable subforum discussions between groups of people who identify with each other. Medical questions could be flagged and redirected to a crowd of experts or knowledgeable citizen scientists. A future version of LabintheWild could connect participants with questions to such expert groups in real-time (e.g., using an approach as in VizWiz, an application that enables answering visual questions [25]). How to train and motivate such expert groups to provide this support will be exciting new research in crowdsourcing and citizen science.

In summary, our work validated online experimentation with volunteers as a viable alternative for studying older adults and people with disabilities. In contrast to MTurk and laboratory studies, the potential of these studies is not yet fully exhausted; We hope that our design implications will inspire researchers to explore ways for improving the recruitment, engagement, and support of volunteer participants with disabilities.

Chapter 4

PARTICIPANTS' OPINIONS ON ONLINE TESTS

4.1 Introduction

Around 20% of the U.S. population, and at least 1 in 3 people around the world, have experienced a cognitive or mental disability at some time in their lives [2,217]. Common cognitive and mental disabilities include neurodevelopmental disorders, such as autism spectrum disorder (ASD) and attention deficit hyperactivity disorder (ADHD); mental disorders, such as borderline personality disorder (BPD) and depression; specific learning disorders, such as dyslexia and dyscalculia; and neurocognitive disorders caused by conditions like traumatic brain injury (TBI) [7]. Diagnosing these conditions is difficult due to imprecise diagnostic thresholds and high rates of comorbidity, which makes differentiating symptoms from co-existing cognitive impairments or medical conditions challenging [13,96,154,231]. As a result, people often receive an insufficient explanation of their diagnoses and are frequently provided inadequate support and resources for interventions [96]. There are also several factors that impede people from seeking a professional diagnosis in the first place, including concerns about the costs or confidentiality, a lack of transportation or knowledge of where to go, and doubts about the effectiveness of a potential treatment [69,82,95,156].

People who suspect or know that they have a cognitive or mental disability frequently turn to online resources to receive more information, understand how their cognitive functions may affect their lives, and meet others with the same conditions [132,164]. Among these resources are online tests and assessments (short: online tests), which people with cognitive or mental disabilities (diagnosed or suspected) use to assess the severity of their cognitive impairment or compare their cognitive performance and behavioral functions to that of others [59,76,135]. Websites that offer such online tests (e.g., mybraintest.org, testmybrain.org, labinthewild.org, psychcentral.com) are often, but not always, based on scientific research and authored by healthcare professionals, yet are rarely suitable for di-

agnosing health conditions. Instead, they commonly serve the purposes of providing initial assessments and/or helping researchers study cognitive deficits. These tests assess behavioral and cognitive traits using either behavioral tasks or survey questions, followed by a results page that tells participants where they stand.

While prior work shows that online tests are perceived as useful by people with disabilities, including those with cognitive or mental disabilities [135], it is unknown whether and how effectively online tests contribute to healthcare and general support systems for people with diagnosed or suspected conditions. What benefits do online tests provide for people with cognitive or mental disabilities? When are online tests most helpful? What are the associated risks, and what may prevent people with cognitive or mental disabilities from participating in online tests? Answering these questions is the first step towards our long-term goal of designing online tests that supplement other resources provided to people with cognitive or mental disabilities.

To shed light on these questions, we conducted 17 semi-structured interviews; 13 with people who have been previously diagnosed with cognitive and/or mental disabilities, and four with people who suspect they may have a condition. Our results revealed that online tests are an important, and previously mostly unrecognized, resource both before and after diagnosis. Before a diagnosis, people use the tests to evaluate whether they may have a cognitive or mental disability, especially when they face barriers that prevent them from getting diagnosed. For them, online tests either provide sufficient confirmation, reducing the need for a professional diagnosis, or they constitute the first step towards getting a diagnosis. After diagnosis, online tests can often fill the gaps left open by people's professional diagnoses, namely the lack of explanations about the severity of their conditions, what behavioral or cognitive functions may be affected, and whether the condition may change over time. As such, one of the main benefits of online tests is that they support people in navigating the impacts of their health conditions and in establishing their disability identity. Our results also revealed a number of challenges that prevent people with suspected or known cognitive or mental disabilities from fully taking advantage of online tests. Based on these findings, we contribute design implications for online tests that could better support people with cognitive or mental disabilities while mitigating risks of misinterpretation, trust, and

replacement of professional diagnoses.

Terminology

We use "mental and cognitive disabilities" as an umbrella term for common mental health conditions and cognitive disabilities, according to the Accessible Writing Guide of SIGACCESS [1]. In the medical field, these conditions are called "psychiatric disorders" [7], which was a term occasionally adopted in HCI. Therefore, following best practices for reconciling naming conventions in different fields [195], we refer these population as "people with cognitive and/or mental disabilities" when we broadly talk about how one's cognitive or behavioral functioning has been affected by cognitive or mental conditions, as well as how their lives have been impacted by the related societal barriers, throughout the paper; we refer to "psychiatric disorders" when specifically speaking within the contexts of the medical field, mostly in the *Related Work* section.

4.2 Related Work

In most cases, receiving a professional diagnosis by a certified healthcare professional or psychiatrist is of utmost importance for any cognitive or mental health condition as it may lead to the development of treatment plans and interventions. Ideally, a professional diagnosis should be obtained as early as possible in a person's life to mitigate potential development of anxiety and depression that can also result in complications with schooling and employment [75,98,161]. In the following, we describe the current literature on 1) how professional diagnosis and self-diagnosis of psychiatric disorders are situated in the healthcare communities, 2) previously found barriers towards receiving a professional diagnosis, 3) the status quo of receiving a professional diagnosis and interventions, and 4) work in the field of HCI towards supporting people with psychiatric disorders, including automated diagnosis tools.

4.2.1 Professional Diagnosis vs. Self-Diagnosis in Psychiatry

Diagnosis has long been a dominant topic of discussion and debate in the psychiatric field. In a general framework, psychiatric disorders refer to disturbances of personal experience, social

behavior, and bodily function [66]. Therefore, the concept of diagnosis is not only medically constructed, but hugely affected by the political, economic and cultural factors [5, 66]. Due to the controversial criteria for defining and diagnosing most psychiatric disorders and its complicated societal impacts [71], receiving a formal diagnosis of psychiatric disorders has its pros and cons. On one hand, a professional diagnosis can help people identify empirically supported treatments, qualify people for insurance reimbursement, facilitate self-understanding, self-legitimation and self-enhancement, and reduce anxiety [3, 166]. On the other hand, however, a psychiatric diagnosis can also have negative consequences, such as stigmatization [3]. Because the process of diagnosing psychiatric disorders is inherently subjective due to its heavily reliance on clinical interviews, a diagnosis can be invalid or unreliable if the clinicians are inexperienced, biased, or blind to the complexity of life and human nature [3, 71].

When seeking an alternative to the traditional professional diagnosis, people often turn to online communities or online self-assessment tests, as both resources provide much more easily accessible consultation for those in need [70, 162]. Online mental health communities operate as an informal medical consultancy for the undiagnosed, where members recommend online diagnostic or quasi-diagnostic instruments to each other and respond to the requests for help with described behaviors [78]. This interaction, however, remains a degree of reverence for professional expertise, as the medical consultancy of participants often include disclaimers such as "I'm not an expert." For people who face barriers that make formal mental consultation impossible or at least very unlikely, online mental health tests become a convenient tool to perform self-diagnosis. For instance, Lewis explored self-diagnosis experience of autism spectrum disorder in adults: most individuals took online self-tests for ASD when they started to doubt themselves and found a "fit" in the criteria [133]. Online self-diagnostic resources are also favored by mental health professionals themselves. An interview study revealed that psychology students who performed self-diagnosis frequently rely on online resources, including online tests [4]. Their academic background and professional knowledge protected them from purely trusting the results of online tests and allowed them to take the tests as supplemental, educational resources.

4.2.2 Barriers and Stigma Associated with Receiving a Professional Diagnosis

There are various reasons for why people may not seek professional help or receive a formal diagnosis. Common concerns include costs, the lack of insurance, unavailable or inconvenient care when needed, not knowing where to go, inadequate transportation, concerns about confidentiality and the belief that the treatment will not help [69,82,95,156]. Likewise, patients often feel that they can handle the symptoms themselves and do not consider their disorder as serious or recognize it as an illness [27,69,156]. Others refrain from acknowledging their disability due to public, perceived, and self-stigmatising attitudes towards mental conditions and cognitive disabilities. For instance, people with psychiatric disorders often feel embarrassed or uncomfortable to talk about their personal problems to others [239]. They have reservations towards talking to both strangers [235] and to people who they knew or knew they would have future dealings with [34,239].

4.2.3 Diagnosis and Interventions of Psychiatric Disorders

Despite the large number of people suffering from psychiatric disorders, diagnosing such disorders is difficult. The Diagnostic and Statistical Manual of Mental Disorders (DSM) [7] and International Statistical Classification of Diseases (ICD) [168], serves as the principal authority used by clinicians and researchers for psychiatric diagnoses and classification in the United States and internationally. In the most recent DSM-5 and ICD-10, diagnostic criteria is listed for each of the disorders, and it is often memorized by trainees in psychiatry and other fields for certification exams [96].

Because multiple changes have been made to the diagnostic criteria throughout different editions of DSM and ICD, the diagnosis of many psychiatric disorders is at times confusing, even for specialists [196]. Moreover, the diagnostic criteria are primarily categorical rather than quantitative (or dimensional), therefore lacking concrete diagnostic thresholds or descriptions of what is typical [95,96], clinicians are forced to make a judgement call, often based on a "clinical significance" criterion that is included with the symptom lists of many disorders. This risks adding subjectivity to the nature of assessment and denying milder symptom presentations [111,208]. The "discontinuity" of diagnostic criteria could

also affect the accuracy of the diagnosis, since symptoms may vary in severity with time and developmental and environmental factors [38, 113, 123].

In addition, evidence has found excessive and scientifically premature splitting of disorders, resulting in high comorbidity rates in clusters of related illnesses, thus, making the diagnosis for each disorder even harder [112]. In the same vein, criteria for disorders are sometimes over-specified so that patients do not precisely match any criteria and receive a diagnosis of Not Otherwise Specified (NOS) [96], leading to unpredictable implications for treatment intervention [104]. In Table 4.1 and Table 4.2, we provide examples of common psychiatric disorders, their definition, prevalence, and the state-of-the-art treatment and prevention strategies. Like the ambiguity in diagnosing psychiatric disorders, prior studies reveal that treatment and prevention strategies often yield equivocal efficacy, as summarized in Table 4.1 and Table 4.2.

4.2.4 Assistive Technologies for People with Psychiatric Disorders

Assistive technologies, computer-mediated systems, and design frameworks for mental health and disabilities have long been of interest to the human-computer interaction (HCI) community. For instance, Sonne et al. developed an assistive technology design framework for people with ADHD [215]. Sanches et al. reviewed 10 years of HCI literature on mental disorders, showing that most innovation took place in automated diagnosis [201]. For instance, prior work has investigated computer-mediated automated diagnosis tools, such as speech-base psychosis detection [21], emotion and disposition recognition [226], which are used to detect and identify psychiatric disorders in clinical settings [24]. Similarly, Hafiz et al. showed that internet-based cognitive assessment tools (ICAT) can be used to screen for cognitive impairment in clinical settings [83]. Researchers have also developed systems that utilize behavioral data such as mouse operations [220], search log, sensor data [105] as well as biofeedback data such as heart rate [201], to facilitate automated diagnosis. Though a wide range of computational psychiatry approaches have been studied and deployed in clinical settings, they are not accessible to the majority of the population.

Prior work has also investigated how online resources and collaborative technologies play

an important role in supporting people with mental disorders and cognitive disabilities. For instance, technology has played an important role in facilitating mental health peer support [170]: people often turn to online communities and social media to self-disclose about their conditions for emotional well-being [11,51,163,202,229], and to seek information, emotional support, and advice [14,134,177,178]. However, the stigma around having these disorders can often hold people back, or even become the source of more severe stress-related illnesses [93,139,163,202].

Furthermore, technology has provided people with (suspected) disabilities a space to learn more about themselves through online experimentation. Li et al. found that many people with various disabilities use online tests on the volunteer-based experiment platform LabintheWild [186] to diagnose themselves, compare their abilities to others, quantify potential impairments, self-experiment, and share their own stories with researchers [135]. Li et al. additionally analyzed comments from participants and online forum entries where people discussed the tests retroactively, but did not host interviews to find out how online tests may supplement the support systems that provided through healthcare, family, and other online resources [135]. In this paper, we aim to shed light on this question by examining the role of online tests in supporting people with psychiatric disorders.

4.3 Methods

Our study was guided by two primary research questions:

> **RQ1**: How do online tests support people with cognitive and mental disabilities, and how do they contribute to existing support systems?
>
> **RQ2**: What are the opportunities and challenges of using online tests for people with cognitive or mental disabilities?

To answer these questions, we conducted semi-structured interviews with 17 participants between February and April 2020. All participants were recruited from online forums with topics related to cognitive or mental disabilities where online tests are frequently shared: 13 from *Reddit* (r/anxiety, r/autism, r/BPD, r/dyscalculia, r/dyslexia, r/TBI) and four from

Wrong Planet. After obtaining the permission from moderators, we posted our recruiting advertisement on these forums, asking people to sign up via a screening survey. Eligibility for the interview required participants to be at least 18 years old. Of the 17 participants, 15 interviewees were from the USA, one was from Australia, and one was from Canada. Eight interviewees identified as male, eight as female, and one as non-binary. As for their levels of education, nine of them had graduated or were attending college, five were graduate students, two completed high school, and one completed army technology school. Most (13) of the interviewees were full-time employees or students while four of them were currently unemployed. Participants' self-reported disabilities and diagnosis status are presented in Table 4.3.

The first and second authors conducted the remote, semi-structured interviews via Google Meet and Zoom. Interviews were audio-recorded and transcribed verbatim with permission. The length of the interviews ranged from 23 to 60 minutes and averaged around 35 minutes. Participants received $10 upon completion of the interview. The study was approved by our institution's Institutional Review Board (IRB) and was performed in accordance with the relevant guidelines and regulations.

We used the constant comparative method to identify patterns in the data and ensure theoretical saturation [44]. An initial coding pass was completed after nine interviews, in which two transcripts were coded by three authors independently in order to develop a codebook. The entire research team then met to refine the preliminary codebook, discuss and modify ambiguous codes, and discuss the data, including early themes we saw emerging. We then continued conducting interviews until we had reached theoretical saturation. Two authors subsequently coded all of the transcripts independently while discussing and modifying the codebook to reconcile ambiguities on an ongoing basis. All 17 interviews were coded at least twice by two or three authors individually. We discussed any discrepancies until reaching consensus. We did not, however, calculate the inter-rater reliability (IRR), as the primary goal of the coding process was not to achieve complete agreement, but to eventually yield overarching concepts and themes [151].

After coding all interviews, all authors conducted multiple sessions of thematic analysis [79] of the interviews, using affinity diagramming to uncover themes of various levels.

We present our themes and results in the following section. Some of the participant quotes have been edited slightly and shortened to improve readability.

4.4 Results

Through our interviews, we found that online tests can fill gaps left open in the support systems for people with cognitive or mental disabilities. We organized our results around four overarching themes: 1) online tests can support people who suspect they have a cognitive or mental disability by removing barriers to professional diagnosis and by fostering an acceptance of their disability; 2) online tests can supplement professional diagnoses by providing additional information and support; 3) online tests provide a basis of connection with other people, and 4) the helpfulness of current online tests is mitigated by issues with trust, difficulties with (over-)interpreting results, confirmation bias, and a lack of connection with other resources, such as online communities and healthcare professionals.

4.4.1 Online Tests Provide Support Pre-Diagnosis

Our first theme revealed that online tests can be helpful for people who suspect that they may have a mental or cognitive disability. Our interviewees often used online tests as a first step to learn more about themselves, especially when a professional diagnosis was out of reach – which turned out to be a common issue.

Several interviewees mentioned struggling to discover how to receive a professional diagnosis as a key difficulty of the diagnostic process. P14, for instance, who suspects he may be on the autism spectrum, said:

> *It's not so much a question of why did you not get a diagnosis or why did you not want diagnosis. It's a question of the steps to get a diagnosis not being exactly clear. (P14)*

P14, instead, did significant research into the difficulties that a person on the autism spectrum might face, contemplated how those difficulties may relate to his own life, and took many online autism tests, all of which indicated he was *likely* on the autism spectrum. He later commented:

> *[The online tests] made me confident enough in my own knowledge to expect that, if I was to speak with a diagnostician, I probably would receive the diagnosis of autism. (P14)*

Other participants were hindered from seeking professional help due to a lack of access (e.g., clinics, transport, cost), a finding which is consistent with previous research [69, 82]. For instance, P17, who suspects she may have dyscalculia, confided in us that "*The testing is expensive. I don't have these resources, and I don't know anyone in person who can help me.*" Instead, online tests provided her with a way to "*help quantify if I even have dyscalculia on any base level, [...], so at least I feel validated enough that I might go see [a therapist].*"

Similarly, P16, who also suspects herself of having dyscalculia, mentioned that online tests and other online resources already gave her sufficient information, obfuscating the need for a costly, professional diagnosis:

> *It's not something that my insurance covers, you know, so I'm worried that it's something that's a major expense to just confirm something that I know to be true. (P16)*

Adding another barrier, P2 pointed out the lengthy time it took her to get professionally diagnosed with Autism:

> *The psychologist that I went to is really difficult to get into, because there aren't enough psychologist specializing in women and girls, especially adult women. (P2)*

During the time of waiting, she turned to online tests to assess herself:

> *I did a couple of those [online diagnostic tests], and scored fairly high. [...] Yeah, I found that quite helpful. (P2)*

Taking online tests during this period re-affirmed her curiosity and motivations to get diagnosed, leading her to ultimately accredit her diagnosis to the tests.

Online tests were also helpful for interviewees whose family members stood in the way of getting a professional diagnosis. In fact, we found that our interviewees sometimes had to rely on family members to make a professional diagnosis possible, either through providing the means to consult a professional or acting as a necessary reference for the professional. Despite this dependency, family members were not always willing to participate. P8, for example, first realized she might be different from others when she was 12, but did not seek professional help until college because her mom "*has always been someone that denied things being wrong even though she is a social worker herself.*" Instead, P8 started using online tests to understand herself better:

> *I've taken like every psychometric quiz that exists. They definitely make you self-reflect a little bit, just trying to understand yourself. (P8)*

The theme of parents denying that their children have a disability was also reflected in P14's comments, who suspected he was on the autism spectrum but never received a formal diagnosis, in part because his mother's lack of participation in the process:

> *My mom was very, very much against the idea that I might be autistic. I went through every single one of the criteria of both autism and Asperger's disorder, and she said, oh wow, those match exactly. And then I told her what they were for. And she said, no, you're definitely not autistic. And she didn't want to participate. So it's very difficult to get someone to participate in the diagnostic process, when they're so averse to diversity. That diversity to even considering the possibility [was frustrating] because she always knew that I was different than other people. But she would claim that it was just because I was smarter than other people. (P14)*

Like P8, P14 also used online tests to assess himself, but he additionally used the results to try and convince his mother that he may have ASD. Although he did not end up obtaining a professional diagnosis, online tests provided him with what he felt was sufficient information.

P14's experience also shows at what stage online tests may be most useful to people who suspect they may have a cognitive impairment or mental disability. Similar to others, he sought out online tests primarily when he first started to realize he might be different, as he was having a particularly difficult time with job interviews:

> *I did [online tests] much more frequently when it was closer to that time than I have over the past few years because it was when something is new, you're kind of focusing on it, you're wanting to learn about it. (P14)*

P2, who took online tests about autism prior to seeing a healthcare professional, also emphasized that online tests became less interesting for her after her diagnosis:

> *I don't really do them anymore. It was sort of pre-diagnosis when I was wondering and up in the air a little bit, but now I don't really take them. (P2)*

Our analysis also revealed that online tests can act as a meaningful resource, providing ways of understanding and coping with their potential cognitive or mental disabilities without having to experience the perceived risks associated with professional diagnosis, such as for privacy concerns, fear of confirming what may be perceived as negative news, or fear of being labeled. P15, for example, feared a professional diagnosis because she did not want to receive an official label, which may result in being treated differently than others. By taking multiple online tests, such as the face blindness test and the autism spectrum quotient test, and discussing the results with others on Wrong Planet, she was able to learn more about how autism affects her life. The test results confirmed her suspicions that she may be on the autism spectrum and allowed her ways of managing how autism may affect her life, without having to receive an official diagnosis:

> *Just being an adult where I can go and see, you know, professionals and have a therapist and things, I've come to more understand myself in these nuances [of ASD]. Now I'm less concerned about looking for a diagnosis or labels so much as just learning skills to deal with things. (P15)*

In summary, the path to obtaining a professional diagnosis is paved with obstacles that prevent people from getting diagnosed early or even at all, ranging from a resistance in the family, fear of costs and being labeled, or privacy concerns. Participants therefore took online tests as a first step towards understanding their suspected cognitive impairments or mental health conditions and seeking professional help.

4.4.2 Online Tests Provide Support after a Professional Diagnosis

Our second theme exposed that online tests can fill some of the gaps left by a lack of support after people receive a professional diagnosis and could even help forming a new identity. Those participants who had previously been diagnosed with a mental or cognitive condition commonly felt that they did not receive enough information or support to understand how the condition might affect their lives and how they can mitigate the negative impact. For example, P11, who was diagnosed with dyscalculia, said: "*I was actually given by the diagnosis, honestly, not much*". Likewise, P8, who was diagnosed with major depressive disorder and bipolar disorder, said "*I was given literally nothing.*"

In particular, interviewees repeatedly raised frustrations over not receiving information about improving their conditions. Their diagnoses were often conveyed as a static condition that cannot be changed. This created a sense of hopeless and felt like "*a lifetime sentence of failure*", as P7 described it. P5, who was diagnosed with borderline personality disorder, revealed to us:

> *It would have been nice to be told that this is the treatment for it. With BPD it took a long time for me to realize that I wasn't destined to live like this forever. And I don't think that was communicated to me very well. They are just like, this is what you have. (P5)*

The lack of information at the time of diagnosis was also apparent in P3's conversation with us, who had been diagnosed with schizoid personality disorder when he was a child, and with ADD and Asperger's Syndrome in his adulthood. Referring to his therapist, he said:

> *[...] they didn't really talk to me about it [schizoid personality disorder] at all. And later in life, like much later, I had to research that on my own. And as for the shrink, his words to me were like, well, I'm sorry, sir, but there's nothing much that I can do to help you. (P3)*

Interviewees were also disappointed about receiving no or only little information about the nuances of their disability, such as how it might express itself in particular situations or what other cognitive functions it may affect. One participant noted:

> *There are so many symptoms of BPD, it can be really difficult to figure out which one is the most urgent to address. (P5)*

To reduce this complexity and better understand specific aspects of their diagnosis, some of our interviewees turned to online tests. P1, for example, communicated to us his doubts about his professional diagnosis of autism and that he did not believe many of the symptoms applied to him. Talking about the time after his diagnosis, he said:

> *I took tests just because of curiosity, procrastination, and just wondering what happened. Also there is a tendency among a lot of autistic people to doubt their diagnosis: "Am I like that? Is it correct? I can totally handle this." (P1)*

As such, online tests helped P1 develop an acceptance of his disability over time by discovering how it expresses itself and delineating which parts of his cognitive and behavioral functions are typical.

Similarly, other participants described their motivation for taking tests as being "*part of the awareness of knowing myself.*" (P4) and "*to find out more about myself and my capacities.*" (P3). Online tests helped them know themselves better and form a disability identity – an important step in adapting to a disability [60].

Like other participants, P3 also perceived online tests as something that helped him get a sense that there was something he could do about his diagnosis. For example, he described using online tests to mitigate some of his symptoms:

> *[Taking online tests] is the chance to quickly and easily learn something. [...] I guess it's a form of brain exercise for me. (P3)*

Using online tests as a form of intervention, such as to exercise the brain, was rare among our interviewees (likely due to the type of tests our interviewees reported taking), but has been found to be a common theme in participants' comments on online testing platforms, such as LabintheWild [135].

What was more common in our interviews was to employ online tests for keeping track of changes in their mental state and ability. This form of longitudinal self-experimentation appeared to be especially valuable for people who experience long-term effects, such as memory loss. For instance, P12 who was diagnosed with traumatic brain injury 20 years ago, took online tests to test how his memory has been affected:

> *I wanted to know what's changed in the last 20 years and even taking a quiz on things that I thought I knew was troubling. (P12)*

Similarly, P8, who was first diagnosed with major depressive disorder (MDD) in 2015, and then bipolar II disorder in 2017, told us that she has been taking the same online tests every one to two weeks over the course of the past two years:

> *[I keep taking] the Myers-Briggs Type Indicator or more popular standardized ones, seeing like, I took it two years ago, what did I get? versus now? Have I changed? I like thinking about these questions and how my experiences have changed who I am, especially now, you know, I graduated high school five years ago, and now I'm graduating with my masters. My life has changed so much in a short period of time, so I've obviously changed a lot in a short period of time. (P8)*

Testing the malleability of their cognitive abilities with the help of online tests was described as a way to gain insights into their disabilities and overcome the feeling of helplessness. Interviewees especially emphasized the importance of this support for adults, as professional interventions are usually focused on children.

To summarize our second theme, participants often felt insufficiently supported by their diagnosis alone and found that online tests could help fill this gap by furthering confidence in a previous diagnosis, explaining nuances that a binary diagnosis could not, and by providing a tool for the self-tracking of health conditions.

4.4.3 Online Tests Facilitate Communal Attachment

Another theme that emerged from our analysis is that online tests often provide people with the opportunity to connect and share their experiences with each other, thereby facilitating the process of communal attachment in which people start feeling part of a community [60]. One of our participants who was diagnosed with bipolar disorder described how she used a combination of a Facebook group and online tests to help her process her diagnosis and get to know herself better:

> *I have a [Facebook] group that we have like 15 people in it, and we do personality tests and stuff all the time, and we always share things and talk about it. (P8)*

By discussing the results of online tests on disability-specific online forums, such as Reddit, Wrong Planet, or Facebook, online tests were valued as a starting point to generate conversations. Our interviewees described they often received confirmation and encouragement by posting tests themselves and/or engaging in these discussions, which made them feel more positive about their disability. Having taken and shared the Wisconsin Card Sorting Test (WCST) on Reddit, a neuropsychological test that assesses perseveration (i.e., the ability to switch ideas or responses) and abstract thinking, P12 commented:

> *I was glad to post the study on Reddit. I was glad to be validated in that somebody read it, somebody understood it, somebody thought it was something. (P12)*

Sharing the studies created a sense of community – participants appreciated that they could support others by inviting them to take the same test and by discussing the results. For example, P17 said:

> *I feel like [sharing the studies on Reddit] does create a sense of community, just because you get to talk about something that you all have access to and can only interpret within the same context. (P17)*

What is noteworthy here is that our interviewees frequently pointed out that online tests gave them a reason to start a conversation in an online community and that these conversations often led to a comparison of people within that community. This is important because current online tests only rarely provide comparisons to others, and if they do, it is often reduced to a comparison with a general population, including neurotypical participants. For P13, this is a shortcoming of current online tests. When talking about his online test results, he expressed that it would be valuable to know "*if it was an extremely similar result based on the severity of their TBI.*"(P13). Online communities allowed participants to receive this more precise comparison to a group of people that mattered to them. For example, P14 described to us how he discussed his results of an online test with people on the Wrong Planet Autism Community Forum:

> *So what I was able to gain was that my results were very, very much in line with the majority of other people's results within those discussions [on Wrong Planet]. It was as close to a confirmation that I could find. Basically the test listed multiple different dimensions where you seem to be a match for all the criteria. It seemed that given all of these different groups, I matched in the majority of those groups, so, there was a lot of confirmation within the discussion. (P14)*

While some participants were wary of fully trusting the words of others on online forums, especially those without active moderation, this participant found the combination of online tests and online community served as a way to self-diagnose and forgo a professional diagnosis. Consistent with the findings of Giles and Newbold [78], this outlines how the combination of these two resources enable people to come to terms with their disabilities by facilitating a way of communal attachment.

4.4.4 The Challenges of Online Tests

While our previous three themes emphasized how online tests can support both people with suspected and diagnosed cognitive or mental disabilities and provide the basis for them to connect with others, we, additionally, saw a fourth high-level theme emerging: online tests are far from perfect. A few of our interviewees even mentioned actively avoiding such tests for a variety of reasons. Here we lay out three key pitfalls of current online tests that emerged from our analysis.

Trust in online tests One common issue raised by our interviewees was the difficulty of finding trustworthy and helpful tests. Not knowing whether to trust a test was sometimes a deterrent for participants who feared for their privacy. P8, for instance, talked about ramifications of taking potentially dubious online tests:

> *Having [...] certain information on the internet that can technically be accessed by anybody can be dangerous for you when it comes to insurance. (P8)*

After seeing an abundance of online resources that "*explain the borderline personality disorder thing in such an archaic way*", P5 concluded that she refrained from taking any online tests that are related to BPD altogether:

> *I usually try and avoid [online tests] because like, I never found one that I thought was credible, and I was just like very trying to be careful with the kind of the internet content [that I pay attention to]. (P5)*

Those interviewees that used online tests were often wary of "*recreational type of tests, such as buzzfeed-like quizzes*", and instead tried to find tests that they could trust using various heuristics. Looking for tests of dyscalculia, P17, for example, heavily relied on the URL to determine the tests' credibility:

> *That having a trustworthy URL may be linked to a society or something like a university or like a trustworthy source. You know, I'm not going to take a quiz from a link that says "dyscalculia is dumb.com". (P17)*

Our conversation with another interviewee, P2, underlined the subjectivity of determining whether a test is trustworthy. Asked how she determines a test's credibility, she answered:

> *[I'm] more attracted to the ones that looked more professional and looked more like they were designed by professionals.*

Interestingly, these conversations highlighted the struggle for finding appropriate and trustworthy online tests, but also showed how people are on their own in identifying what makes tests trustworthy.

(Over-)interpretation of the results In addition to worrying about the difficulties in determining which test to trust, participants also sometimes struggled to interpret the results, and consolidate the results with their assumptions. For example, P15 suspected that her inability to recognize people was due to having autism, and therefore, took a face recognition test to find out:

> *Hilariously, I scored in the 98th percentile in terms of being good at recognizing faces. So my inability to recognize my family members outside of context, I still don't understand. I don't know if it's because that test only scores your short term memory or more because of other reasons, like they only use a certain number of faces or something. (P15)*

P15 felt that the result did not align with her assumptions about the symptoms of autism and struggled to find a reason for her high score. She was also surprised that the test did not confirm her struggles with recognizing faces, showing how participants can over-estimate how generalizable tests are to a variety of situations.

Very similarly, we found that confirmation bias played a role in whether someone trusted and accepted test results. For instance, P10 told us that he only occasionally took online tests related to dyslexia — but that he would only trust the results if they confirmed his prior dyslexia diagnosis and what he already knew about dyslexia or himself:

> *This is coming from someone who knows they have it, has known they've lived for it forever. I feel like I would trust the result if it told me what I already knew. (P10)*

Other participants confirmed having issues with trusting results of online tests and explained when they were more likely to believe the results. For example, P7, who has been living with bipolar disorder, generalized anxiety disorder, and social phobia for more than 20 years, told us:

> *I probably would have to see results from other people and get a large study, to be confident of the veracity of any particular test. (P7)*

Similar to P7, P12, who was diagnosed with traumatic brain injury (TBI), also emphasized the importance of seeing his online test results in the context of others to aid his interpretation:

> *I think it's very important that somehow people really ought to get a baseline for just general capabilities, because trying to figure out where you were without being able to qualify where you were, is really difficult. (P12)*

Presenting the results in the context of neurotypical participants was also mentioned as important to ensure that people do not overreact, as a quote from P7, who is on the autism spectrum, exemplified:

> *I think that [comparing to others] would be very interesting. It would let me know if I'm overreacting if I compare myself to a control. I'll know then where I was, where I stand in any particular situation. (P7)*

Taken together, these findings emphasize the difficulties of interpreting results and the important role of surrounding information, such as comparisons to others. The following subtheme further underlines that online tests cannot be seen as a stand-alone solution.

Current online tests do not provide a way forward Another challenge that our analysis revealed was that online studies often fell into the same trap as professional diagnoses: People often felt left alone with the results and did not know what to do with them. Our interviewees emphasized the need for providing additional resources and follow-up advice. When asked about how online tests could be improved, P17, who suspects they have dyscalculia, answered:

> *It'd be kind of crappy to get a result that says you have to struggle and then leave you stranded, you know, on a lifeboat all alone. You have spent your whole life [suspecting something is wrong] which is probably why you're taking the quiz in the first place. (P17)*

Other interviewees confirmed that the results of online tests seemed to often confirm and reinforce that they were struggling, rather than provide a way forward to deal with the struggle. This is in line with suggestions by one of our participants, P11, to provide pointers on how to connect with a psychologist and/or how to get a professional diagnosis:

> *I don't know how practical it is that maybe somebody kind of popped up, [...] like a psychologist nearby that could help you, or just give a location on a map [...]. But then it kind of comes off like sponsored [...] I feel like just giving more options for resources [would be helpful]. (P11)*

This further emphasizes the shortcomings of current online tests, which are seen as disconnected from the professional healthcare system and do not provide a straightforward path towards finding other resources or obtaining a professional diagnosis. However, P11 also pinpointed one of the difficulties of connecting tests and providers, describing it as a risk for the test being perceived as sponsored. In the following, we will discuss our overall results in the context of such challenges and provide potential solutions for online tests to better support people with cognitive or mental disabilities.

4.5 Discussion and Design Implications

In this paper, we showed that online tests provide an opportunity to supplement, and to some extent replace, resources that are otherwise out of reach for people with suspected or known cognitive or mental disabilities. Our interviews have revealed that online tests are already contributing to the support system for people with cognitive and mental disabilities.

In particular, we found that our participants predominantly use online tests before (and sometimes instead of) a professional diagnosis. Getting professionally diagnosed was often described as out of reach, due to cost and access issues or because of resistance within their own families. To work around such barriers, our interviewees use online tests to validate their own suspicions and justify the need for a professional diagnosis to both themselves and their family members. With our interviewees often turning to online tests as a first step towards professional diagnoses, we can see that these often relatively informal and anonymous tests play a unique role in the support systems of people with disabilities: a way of slowly and informally introducing people to their disability without the potential risks perceived by an official, inescapable professional diagnosis. People can choose to believe the results of an online test, but, as our interviews have shown, there is a way out by disputing a test's validity. As such, online tests suffer from confirmation bias, but at the same time, our data shows that this might be their strength given that it allows people to slowly develop an acceptance of their disability. A professional diagnosis should of course provide final confirmation, but it should also come with enough resources to help a person accept a potential positive diagnosis of a cognitive or mental disability and move forward with a treatment plan.

We also hope to push towards a norm of including and providing more attention to individuals who self-diagnose disabilities, than it is now. On one hand, our community can think about including people who self-diagnosed disabilities in studies, which could help achieve sufficient N to detect medium and small effects, but we would always encourage researchers to treat self-diagnosis and professional diagnosis as two levels in the analysis. On the other hand, there is insufficient work to know whether and in which cases online tests could fail and how the self-diagnosis results compare to professional diagnoses. Therefore,

future work of rigorous clinical trials would be needed to assess this.

Our results also show the value of online tests post-diagnosis. This is similar to the findings in Li et al. [135] and Oliveira et al. [167], who showed that participants in online tests provided on LabintheWild frequently try to better understand their disability. We extend this prior work by showing that the tests are also used for the purpose of validating a professional diagnosis and for exploring what other behavioral or cognitive functions may be affected. Participants in our interviews commonly described this as finding out what their capacities are and what the symptoms of their disability are in comparison to others. Similarly, they were often given no information as to the malleability of their disability over time, instead perceiving it like an unchangeable "lifetime sentence of failure", as one of our interviewees put it. Online tests support them in establishing a personal disability profile by participating in a range of tests and comparing their personalized results to others. Interviewees also use online tests to track how their disability expresses itself over time, which confirms the finding in previous work that such tests are sometimes used for self-experimentation [135]. Both of these activities are likely supporting the process of establishing a person's disability identity, which, according to our results, is a gap that conventional resources available to people with disabilities often leave open.

While these findings are very encouraging, our interviews also lay open a number of challenges that online tests will need to overcome to improve their utility for people with cognitive and mental disabilities. In the following, we will discuss these challenges in the context of their implications for the design of future tests. For each design implication, we first state the implication that the finding brings, and then explain the finding from our interview.

Design Implication 1: By integrating high-quality online tests that assess cognitive and mental disabilities into professional healthcare systems, more people could benefit from taking these tests.

Our findings are encouraging in that they indicate online tests often provide a pathway to obtaining a professional diagnosis. While such tests cannot replace a professional diagnosis, they can point out who may be at risk and additionally raise awareness of specific

disabilities, which may also help advocate normalization of disabilities more generally [240]. It is important to note that such tests would need to be rigorously and carefully developed to avoid pitfalls, such as over-interpretation of the results. Therefore, one possible solution is to partner with the medical community.

By better integrating online tests into professional healthcare systems, online tests can assist in reducing barriers to obtaining a professional diagnosis and serving as a first step towards it. Tests developed by researchers and doctors could include pointers to resources such as how to find an adequate healthcare professional for a formal diagnosis. Such resources could increase access to professional diagnosis and empower online experimenters to continue taking steps towards understanding their (suspected) health conditions through credible means. However, one of our interviewees raised the issue of perceiving tests as sponsored if connected to specific healthcare resources. Therefore, providing a choice and more general pointers to professional healthcare resources, such as to a database of psychiatrists, may be a solution. Partnering with hotlines and other services available for people with cognitive or mental disabilities may also be a way of providing online test participants with immediate, in-person support if needed.

Design Implication 2: Standardized guidelines should be developed for the design of tests and for communicating the test results, before verified tests could be promoted publicly and confidently.

Of course such ubiquitously available tests carry a number of risks. Our interviewees confirmed a perhaps unsurprising fact that current online tests are frequently untrustworthy. Indeed, a quick web search for "online test" surfaces a number of scientifically questionable tests. Exacerbating this problem, people also commonly overestimate the diagnostic abilities of such tests, or they relate a specific test to their disability despite no indication that it is designed to assess or diagnose related behavioral or cognitive functions [135]. Because of these risks, it could be helpful to develop efficient ways to verify online tests for potential participants, such as by developing a set of heuristics that indicate scientific validity. Verified tests could be made available on a single platform that could be promoted in schools and in online communities commonly accessed by people who suspect they have a cognitive

or mental disability. Such platform could also employ user ratings that convey perceived helpfulness. In addition, it will be beneficial to develop a set of guidelines that tell participants what to expect, who developed the test, what the test can and cannot do, and how to interpret the results. A key to the guidelines will be to research and develop language that prevents participants from over-interpreting the results, such as by communicating uncertainties and offering additional resources. Note that there may not be a one-fits-all rule, but that these guidelines can be broken down by types of the disabilities or other criteria. Providing test designers (both researchers and others) with guidelines and best practices for the development of these tests and for communicating results is perhaps the most important first step before we can confidently promote such tests.

Design Implication 3: Online tests should ideally rely on representative baseline data to provide participants with nuances of their conditions and with comparison to a specific group of people.

An additional disadvantage of current online tests that our work uncovered is that they often insufficiently support people's desire to understand the nuances of their conditions and how their symptoms compare to others. Just like professional tests for assessing or diagnosing disabilities, online tests lack (normative) baseline data to provide an individual with comparison to a specific group of people, such as those without a disability, or people of the same age group with the same diagnosis. Creating tests that can provide such comparisons and provide information about the nuances of the conditions (e.g. the severity of various symptoms) would require testing a large number of people, which is difficult, but not impossible. In [73], for example, the experiment platform LabintheWild [186] was used to collect normative data from 250k healthy individuals and develop classifiers for accurate detection of Ataxia and Parkinsonism. The resulting system can compare individuals' performance to the baseline data of a specific age between 5 and 80 years old. Similar data collection efforts to develop predictions of severity levels and to provide comparisons to other people with similar demographics could be employed for cognitive or mental disabilities too.

Design Implication 4: Online tests should align with the affirmative model of

disability by highlighting a test participant's strengths and providing additional resources that describe positive examples.

We found that one of the challenges of online tests is that they are frequently perceived as "downers", i.e., as a way of confirming what many already suspected without providing a positive path forward. This is counterproductive to the affirmative model of disability [221], which promotes a more positive view of disability and has the goal of people focusing on their strengths rather than on personal tragedies. A good example for refocusing the discussion of a disability on its strengths are books, such as "*The Gift of Dyslexia: Why Some of the Brightest People Can't Read and How They Can Learn*" [50], which describes success stories of people living with dyslexia. To align online tests with the affirmative model of disability, online tests would need to diversify and test several behavioral and cognitive functions in order to emphasize those in which a person may excel. In addition, support to see their own strengths may be provided by including additional resources that outline a path forward which does not exclusively focus on low-performance functions.

Design Implication 5: Online tests should support participants in sharing and discussing their results with others by providing links to appropriate online communities and to specific threads discussing a certain online test whenever available.

Helpful for working towards an affirmative model of disability and supporting people's creation of a disability identity is connecting them with others in a similar situation. Our participants suggested that online tests gave them a reason to discuss their disability in online communities and made them feel more connected to others. However, to do so, they had to find an appropriate online community and introduce the test there. An obvious solution to this problem may be to create online testing websites that offer a forum for an immediate discussion of results, similar to what has been proposed in [135]. If the forum allowed anonymous posts to preserve privacy, we believe this could indeed better support participants in sharing and discussing their results with others. But there is something to be said about keeping online tests and online communities separate: Online communities are already established and many of them that are specific to certain disabilities, e.g.,

the subreddit r/ADHD or WrongPlanet.com, to have lively discussions with many long-term members. Instead of offering yet another forum or online community, a more fruitful approach for online testing websites could be to partner with, or to simply point participants to appropriate online communities. Ideally, a link would not simply lead participants to the online community's homepage, but rather to the specific thread that discusses a test.

4.6 Limitations and Future Work

While our work contributes exciting insights into the role of online tests for identity formation, it is only a first step towards our larger goal of better supporting people with cognitive or mental disabilities. Because we recruited participants for our interview study from online communities on a variety of cognitive and mental disabilities, the findings presented here are specific to people who currently use these online communities and thus, either suspect or know that they have a health condition. As such, our findings cannot shed light on the opportunities and challenges of online tests for people who do not suspect that they have a disability or for those who refrain from using online communities, for example because they may not yet have started the process of accepting their disability. Our choice to recruit from online communities was made because prior work had reported that online test participants often share their results in online communities; however, future work could broaden our findings by studying a broader sample of people with cognitive or mental disabilities, including those who do not necessarily use online communities.

Another limitation is that the majority of our participants came from the U.S., with one from Australia and Canada, respectively. Although the two non-U.S. participants found the benefits and the limitations of online tests to be the same in our analysis, our findings may largely reflect gaps in the American health care system for supporting people with cognitive or mental disabilities. We do believe cultural differences exist; for example, stigmatization differs across cultures and so does the acceptability of seeking out professional diagnoses. Culture has also been shown to be a leading diagnostic factor in cognitive and mental disabilities in previous work [5]. Therefore, online tests may play different roles within different cultures, societies or mental health care systems. An interesting direction of future work could be a larger survey study that sheds light on the variations across countries and

reveals a potential relationship between mental health care systems and the usefulness of online tests for people with cognitive or mental disabilities.

Likewise, self-selection bias may also impact the generalizability of our findings. People with mental disorders, for example, might have been reluctant to respond to our call because of potential prior experiences with stigma, marginalization, and oppression [232]. Those people may also refrain from using online tests because of similar fears, especially if online tests do not make it 100% clear that they do not collect identifiable data.

The work we presented here shows that online tests are often perceived as helpful by people with cognitive or mental disabilities and that they provide opportunities for forming a disability identity which a professional diagnosis and resources provided by the healthcare system often do not. However, there is a risk that online tests could be *perceived* as helpful while they are actually not, or worse, that they could be worsening a participants' state. An urgent next step therefore needs to investigate which online tests are truly helpful for people with cognitive or mental disabilities from the perspective of healthcare providers AND from the perspective of test takers. Studying this question with a large sample of online tests (with various degrees of scientific quality) may also reveal heuristics for developing best-practice guidelines for tests that are truly useful.

4.7 Conclusion

This paper contributes insights into the use of online tests by people with cognitive and mental disabilities as a first step towards better supporting them pre- and post-diagnosis. Our findings from 17 interviews with people with a variety of cognitive and mental health conditions (both suspected but undiagnosed and professionally diagnosed) showed that one of the main values of online tests is that they address shortcomings in the support of people with cognitive and mental disabilities, such as difficulties obtaining and justifying a professional diagnosis, a lack of information about the nuances of a disability, and a lack of continuous support provided by healthcare providers. Most importantly, our findings revealed that online tests are an important resource for developing a disability identity for people with suspected or known conditions. By contributing a discussion of challenges that current online tests pose, we hope to lay the foundation for future research efforts that

leverage the advantages of online tests and maximize their benefit to people with cognitive and mental disabilities.

Table 4.1: Summary of definition, prevalence, the state-of-the-art treatment and prevention of common psychiatric disorders. The prevalence statistics is cited from National Institute of Mental Health (NIMH) if not otherwise specified.

Disorder/ Disability	**Definition [7]**	**Prevalence in the U.S. [2]**	**Treatment & Prevention**
Attention-Deficit/ Hyperactivity Disorder (ADHD)	A persistent pattern of inattention and/or hyperactivity-impulsivity that interferes with functioning or development.	11% (4-17 years old); 8.7% (adolescents); 4.4% adults;	Medication can effectively treat ADHD symptoms [230].
Autism Spectrum Disorder (ASD)	Persistent deficits in social communication and social interaction, along with restricted, repetitive patterns of behavior, interests, or activities.	1.9% (8-year-olds)	No efficient therapeutic interventions for cor symptoms for ASD [65].
Bipolar Disorder (BD)	A group of brain disorders that cause extreme fluctuation in a person's mood, energy, and ability to function.	2.9% (adolescents); 2.8% (Adults)	Pharmacological and non-pharmacological approaches yielded mixed results [214].
(Borderline) Personality Disorder (BPD)	A group of brain disorders that cause extreme fluctuation in a person's mood, energy, and ability to function.	1.4% (adults)	Dialectical Behavioral Therapy (DBT) is effective in treating BPD [56]; effectiveness of pharmacological treatment is unknown [84].
Dyscalculia	A specific learning disability affecting the normal acquisition of arithmetic skills, a brain-based disorder.	6% [36, 207]	No effective treatment; interventions focus on specific training and instruction [157].
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 [9].	15-20% [8]	No effective treatment; interventions are education-based, focusing on spelling, visuo-attention, visual perception, etc. [74, 175]

Table 4.2: Table 4.1 continued.

Disorder/ Disability	**Definition** [7]	**Prevalence in the U.S.** [2]	**Treatment & Prevention**
Generalized Anxiety Disorder (GAD)	Excessive anxiety and worry (apprehensive expectation), occurring more days than not for at least 6 months, about a number of events or activities, such as work or school performance.	2.7% (adults); 2.2% adolescents;	Cognitive-behavioral therapy (CBT) [196, 197] is found to be efficacious; medication can be used to reduce symptoms [90].
Major Depressive Disorder (MDD)	Persistent feelings of sadness and hopelessness, lose interest in activities, physical symptoms such as significant weight change, diminished ability to think or concentrate.	7.1% (adults); 13.3% (adolescents)	Commonly treated with antidepressant medications and psychological therapies [117].
Social Anxiety Disorder (SAD)	Persistent fear of one or more social or performance situations in which the person is exposed to unfamiliar people or to possible scrutiny by others.	1.9% (8-year-olds)	Same as above (GAD).

Table 4.3: Interviewees' demographic and diagnostic information

ID	Gender	Age	Disability/Disorder	Diagnosed
P1	M	18 - 30	Autism	Y
P2	F	40 - 50	ADHD, Autism	Y
P3	M	50 - 60	ADD, Asperger's Syndrome, schizoid personality disorder	Y
P4	F	40 - 50	Autism, learning disorder, generalized anxiety disorder	Y
P5	F	18 - 30	Borderline personality disorder	Y
P6	M	18 - 30	Borderline personality disorder	Y
P7	M	50 - 60	Bipolar disorder, generalized anxiety disorder, social phobia	Y
P8	F	18 - 30	Bipolar disorder, major depressive disorder	Y
P9	F	30 - 40	Dyslexia	Y
P10	M	18 - 30	Dyslexia	Y
P11	F	18 - 30	Dyscalculia	Y
P12	M	50 - 60	Traumatic brain injury	Y
P13	M	30 - 40	Traumatic brain injury	Y
P14	M	30 - 40	Autism	N
P15	F	30 - 40	Autism	N
P16	F	30 - 40	Dyscalculia	N
P17	Non-binary	18 - 30	Dyscalculia	N

Chapter 5

HEALTHCARE PROFESSIONALS' OPINIONS ON ONLINE TESTS

5.1 Introduction

In addition to understanding participants' views of such online studies, it is critical to also find out how those online studies are perceived from a professional's point of view and how these online tests are positioned in the professional diagnostic systems. To do so, we conducted semi-structured interviews with six psychiatrists who have experience diagnosing a variety of cognitive disabilities.

We have learned that subjective rating scales and task-based assessments (which are commonly seen in online tests) are critical components during the diagnosis of cognitive disabilities, but clinical interviews with the ability to communicate with and observe the patients is indispensable to make an accurate diagnosis and provide appropriate treatments. By reviewing examples of online self assessments, our interviewees all agreed that these tests can be a good starting point for people to start exploring about their conditions, but definitely not an ending point. Instead, the tests should provide additional resources for people to seek clinical diagnosis. This is aligned with one of the shortcomings of current online tests perceived by our participants in Chapter 4, that the results are often disconnected from the professional healthcare systems and do not provide a straightforward path towards obtaining professional help. Finally, our interviewees commented on what key components make a test valid, again consistent with our findings from Chapter 4 and from previous work [186].

5.2 Methods

We conducted semi-structured interviews with six interviewees in July 2020. All interviewees were recruited by emails and all are scholars and psychiatrists from the greater Seattle area. Unlike surgeons who often only specialize in diagnosing and performing surgeries in one area, psychiatrists usually diagnose and treat a variety of cognitive disabilities. We have

Table 5.1: Interviewees' demographic and occupational information

Participant ID	Gender	Clinical and Research Foci
P1	M	Integrated Care, ADHD, Trauma-related Disorders
P2	F	Child and Adolescent Psychiatry
P3	M	Adult Psychiatry, Telepsychiatry, PTSD
P4	F	Autism Spectrum Disorder
P5	F	Child and Adolescent Psychiatry, Autism Spectrum Disorder
P6	F	ADHD

listed the detailed information about the interviewees, including examples of their clinical and research interests, in Table 5.1.

During the interviews, we began by asking interviewees to describe the process of diagnosing psychiatric conditions of their clinical focuses. We then asked about the resources they would recommend for people who cannot get diagnosed or who are still in the pre-phase of diagnosis. We next sought to understand interviewees' opinions on the existing online studies whose target audience are people with suspected cognitive or psychiatric conditions. We showed each interviewee two to three examples of online tests, to probe their opinions on what is trustworthy and beneficial and what is not about these tests. Finally we asked interviewees how they would improve the online tests shown to them.

Interviews were conducted remotely and lasted 15-20 minutes. They were recorded and then transcribed with permission. The study was approved by our institution's Institutional Review Board (IRB) and was performed in accordance with the relevant guidelines and regulations.

We went through the transcripts and coded them for themes using an open coding approach [40]. Through multiple iterations along with periodic discussions with the rest of the research team, the following major themes were selected. Because of the low number of interviewees, our findings should be regarded as indicative.

5.3 Results

5.3.1 Diagnostic Procedures in Clinical Settings

We first summarized our findings on how the interviewees make diagnoses of psychiatric disorders of their expertise, in order to better understand the status quo of diagnostic procedures in clinic, and thus where the online self-assessments are positioned from the clinical perspective.

First of all, we found that the diagnostic procedures of cognitive disabilities vary depending on multiple factors, including but not limited to type (of the disorders) and age. Even for the same condition within the same age group, our interviewees follow slightly different protocols to make a diagnosis. For example, three out of six (P2, P4, P5) of our interviewees specialize in the diagnosis of autism spectrum disorder for children. P2 described the diagnostic procedure as following: the patient will first complete a set of patient, family and school information about developmental history, symptoms, and measures that cover not only the chief complaint. Then it is followed by a 90-min visit reviewing the information, a second 45-min visit for the autism diagnostic interview with the parents, and the autism diagnostic observation schedule (i.e. ADOS, the interaction with the child with different tasks). For children with neuro/psych dysfunctions, they will look at IQ, memory and executive function with much more detailed measure, etc. P4 described a similar procedure: starting with an interview with parents/teachers about medical, educational and living history, then followed by a ADOS and a measure of adaptive functions. They will then meet with another team member to get a conclusion on the diagnosis.

However, all of our interviewees mentioned that patients' subjective ratings are not sufficient; they have to combine the self-report information with the objective measurements, observations and tasks. For instance, P1 described the importance of clinical diagnosis on ADHD as following:

> The key is clinical diagnosis. People shouldn't just make a diagnosis based on someone filling out a rating scale. And then you also have to have clear signs of functional impairment. [...] We do so by asking patients to do continuous

> performance tasks (CPTs), qualitatively talking to teachers and parents. Or you can get some of that qualitative data on some of these rating scales that actually have scales for functional impairment. But the challenge always is all the data points are from someone's subjective experience. So you know, they are not truly objective measures because they are asking someone to say what they think.

In summary, our interviewees all described certain forms of unsupervised measurements, such as subjective ratings and self experiments, as parts of the diagnostic procedures, but conducting clinical interviews with the ability to communicate and observe the patients is critical to distinguish between the environment factors and the actual cognitive dysfunctions that actually lead to the diagnosis.

5.3.2 Opinions on Online Tests

After reviewing examples of existing online tests that people took and discussed in online forums, our interviewees expressed a mixed attitude towards taking these online tests as a way to understand people's (suspected) cognitive disabilities.

Challenges and Concerns

One frequently mentioned concern from our interviewees is that a majority of cognitive disabilities can be reflected in many different aspects, the areas surfaced by some of the online tests is too limited, therefore it might result in inaccurate conclusions. Taking autism spectrum disorder as an example, P4 described that people with autism can be affected in many different aspects: sensory sensitivities, repetitive behaviors, repetitive body movements, restricted interests, peer relationships, etc. When P4 was shown the Social Intelligence test [1], she commented:

> there's so much more to autism spectrum disorder than just reading other people's facial expressions and reading people's emotions from their eyes. So yeah,

[1] `http://socialintelligence.labinthewild.org/mite/`, study 3 in [135].

> I think it's hard to take that alone and say, Okay, looks like I have a really high risk, or it looks like I have autism because I didn't do well on this particular test, you have to take into account so much more than just that.

As we learned, the Social Intelligent test is examining the brain's ability to process social information received from the eyes, but it is only one of the symptoms of autism. There are many more that are not evaluated by the test. Similar concerns are raised by other interviewees as well. For instance, after reading through RDOS Aspie quiz [2], a popular autism self assessment tool discussed in various autism forums [135], P5 expressed:

> I would want to see the sensitivity and specificity of whatever items they're using to diagnose. And my worry is that if you know they're not very sensitive or specific, that they'll be over or under diagnosing in a predictable way. Which, you know, you're focusing on kind of over interpreting [...] They're giving probably diagnosis but it's probably not sensitive enough to be (diagnostic).

Our interviewees also expressed concerns that one online test might not be suitable for everyone with the same disability – there are so many factors that affect how the condition expresses itself, such as age, gender and comorbidity, and therefore, some questions become essential for one subgroup of the population while not applicable for others. For example, P4 was presented a self-test with a series of questions for parents to find out whether their child's symptoms resemble those of children diagnosed with autism spectrum disorder. Taking one question "Does your child point with one finger to ask for something or to get help?" as an example, P4 commented on the test:

> we wouldn't ask them this question for all age groups. So that's one thing that it's not really adapted for age. I'm looking at the pointing thing I know for parents of teenage years, 14 or 15, it's really hard for them to think about does their child point because they've kind of gone past that developmental stage. Whereas if you're two or three years old, that's a very important question to

[2] http://rdos.net/eng/

> help distinguish, you know, autism or something else. Point, seems funny, but it's very important. But it's not so important when you're 15.

Thus, it does not specify an age range that this test is suitable for, and might lead to inaccurate conclusions for some participants. In addition, a few of our interviewees also mentioned the importance of comparing people's results with others who have similar experiences when making a diagnosis, which most of the existing online tests do not provide. As P6 mentioned:

> This is why we still say the clinical judgment of providers is really important. The clinical provider can compare that child against other children and can kind of see maybe it's more environment than it's actually the child and might make me make that diagnosis or not.

"The online tests can be a good starting point, but definitely not an ending point."

Most interviewees mentioned that people cannot just rely entirely on the self-report online surveys to do a self-diagnosis, but it certainly can be a first step to motivate people to learn more about themselves and their conditions. For instance, P1 said:

> you know, I think most of those types of things should be a starting point, not like the end point, if they can start you on a journey of self-discovery, go talk to your doctor, go read a book about this in more detail. But if they're describing themselves as the end point, then I think that's misinformation because human beings are way too complex to have everything solved for them by taking a three or four minute test online.

Similarly, other interviewees confirmed that the online tests would be helpful if they provide resources and recommendations on how people can move forward, rather than solely relying on the test results. As P2 mentioned:

> I think it'd be helpful to have something that says, your scores are elevated, and the next step, we would recommend to refer yourself, you know, to a psychiatrist, someone who could evaluate and maybe have some specific resources like, here are some people, here are some options for where you could go to try and get that evaluation.

What Makes a Valid Test

In this following section, we briefly lay out key factors that should be included in a valid test discussed by the interviewees. Our findings are well aligned with those from the previous chapters pointed out by test participants.

First, our interviewees argued that online tests should be developed from validated research, and that evidence should be explicitly explained to the participants. When reviewing the child autism test from ADDitude [3], P4 commented:

> What I like about this one is it says the self test was adapted from the Modified Checklist for Autism in Toddlers - Revised (M-CHAT-R). That is a highly used instrument that has pretty good sensitivity and specificity and ultimately predicting an autism diagnosis.

Similarly, P3 said he would first look for a strong affiliation, "to see that it's associated with a well respected institution that has a strong affiliation, say with the American Psychiatric Association, or the PTSD National Center, or something." In addition, comprehensive consent forms and warnings about how to use the test were mentioned by P1, P5 and P6. For example, P1 said:

> They should be very clear about what is the limitations of extrapolating, you know, whatever result is found to that individual person. I think they should clearly say what the limits are, say, that this may provide important information for you, but it's not suitable to use as a diagnosis. If you have concerns about

[3] https://www.additudemag.com/autism-spectrum-disorder-symptoms-test-children/

this, you should follow up with your health care provider, just have a normal warning about the limits.

Interestingly, these conversations with healthcare professionals suggested how to find appropriate and trustworthy online tests, and they resonated with how participants (from Chapter 4) are on their own in identifying what makes tests trustworthy.

5.4 Discussion

Despite all the barriers, seeking professional help is still the optimal choice for people with (suspected) cognitive disabilities to better understand how their cognitive functions affect their lives and what next steps they should take. Thus, in this chapter, we explored where online tests, as an alternative for people who do not (yet) have the access to the professional healthcare systems, are positioned in the clinical system from the healthcare providers' point of view. Through the interviews with six psychiatrists who diagnose and treat a variety of cognitive disabilities, our main finding is that well-designed online tests can be a helpful starting point for people to explore their cognitive conditions, but online tests could definitely not replace a professional diagnosis.

Our findings in this chapter are largely consistent with what we found in the previous chapter during the interviews with participants who take advantage of online tests. The doctors confirmed that online tests could be a helpful tool for people who just start to explore themselves. However, similar to how the participants from Chapter 4 use various heuristics to find trustworthy tests, the doctors also emphasized that they would only trust online tests that are developed from validated research at well-known institutions and that present clear, thorough consent forms which guide participants to properly interpret the test results. An exciting direction of future work could be exploring ways that help participants validate online tests with the knowledge from the experts, from other participants who have taken the tests before, or from other people who has more information about the tests, in addition to the participants having to replying on their own judgements.

In addition to confirming what we have learned from the previous chapters, we also gained knowledge on whether/how online tests (or those with similar formats) are part of the

clinical diagnostic procedures, from a professional perspective. We learned that psychiatrists usually make a diagnosis based on a variety of source information, including but limited to the patient's self-report complaints, medical/educational/developmental history, interviews with the patient's family, the communication and observation of neuro-psychological tests such as CPTs. Therefore, the psychiatrists confirmed that the researchers should make it clear the online tests should only be indicative of suspected cognitive disabilities, rather than diagnostic. One possible direction of future work could be exploring the possibility and ways to integrate the online tests as part of the official diagnostic process, providing people with additional access to the professional healthcare systems.

Last but not least, we learned that the psychiatrists would also consider the person's environmental and demographic background when making a diagnosis. Thus, online tests are limited in their ability to infer contextual information from the individuals. They usually only focus on one aspect of the cognitive abilities or would only be appropriate for certain sub-populations (e.g. some symptoms are only applicable for children rather than adults). Therefore, it is important to examine ways to effectively communicate the limitation of the test results and further make the online tests adaptive as future research.

While this work contributes to the overall understanding of how online tests contribute to the support systems of people with cognitive disabilities from the psychiatrists' perspectives, similar to the limitations of our work in Chapter 4, we only interviewed healthcare professionals from the U.S. and even one region of the U.S. where the healthcare systems are well-developed. Thus, the diagnostic procedures might be different in other parts of the country or around the world. Online tests may also play different roles within different cultures, societies and mental health care systems.

Part II

EMPIRICAL VOLUNTEER-BASED ONLINE EXPERIMENTS ON DYSLEXIA

In addition to examining the feasibility of using volunteer-based online studies for studying cognitive disabilities and understanding how these studies are perceived by participants and healthcare professionals, it is also important to conduct online studies that generate new knowledge about cognitive disabilities. Therefore, in Part II of the dissertation, I describe two works that empirically demonstrate that we can reliably study dyslexia, a common cognitive disability, and detect nuances between populations via volunteer-based online experiments on LabintheWild: Chapter 6 examines the effectiveness of web browser "Reader View" on user experience for people with and without dyslexia. Chapter 7 introduces and validates Virtual Chinrest, a novel method in web browsers which enables visual perception studies of dyslexia that are sensitive to display parameters.

Chapter 6

WEB-BROWSER READER VIEWS

6.1 Introduction

Reading information on screen and in web browsers has increasingly taken the place of traditional ways of reading. However, digital content can impede people's reading fluency and comfort due to visual clutter, advertisements, or a lack of contrast [192,212]. These problems are exacerbated for people with reading difficulties, such as dyslexia, which is a cognitive disorder that impacts people's reading ability in various ways. People with dyslexia (˜15-20% of the world's population [8]) often have difficulties organizing language and eliminating non-relevant elements when finding and understanding information on websites [158].

Inspired by one of the strategies that some people with reading difficulties employ for better web readability – using the browser's Reader View [158] – this paper explores how Mozilla Firefox's Reader View affects people's reading performance and user experience. More specifically, we endeavor to answer the following research questions:

RQ1: When and how does Reader View change a webpage?

RQ2: How does Reader View impact reading performance, perceived readability, and the user experience compared to the standard presentation of a website?

RQ3: Do people with dyslexia benefit more from the Reader View than those without?

The first research question (RQ1) was motivated by the fact that there is insufficient documentation on what the Reader View does for any of the popular browsers. This also means that website designers do not currently know how to design websites that could be transformed to Reader View pages. We therefore characterized which modifications are performed by the Reader View in the Firefox web browser [159] by inspecting its open-source code and by quantifying the difference in visual designs between standard webpages

and their Reader View version. We show that Reader View is only triggered on 2% of homepages and 41% of their child pages in our sample of 1100 webpages, and that the low percentage is mainly due to insufficient word count in the main content section. We also show that the Reader View reduces both images and text by around a third, while doubling the use of uniformly colored areas, including white space. This results in significantly lower visual complexity and colorfulness compared to the original websites.

It is also important to empirically validate the utility of reading tools, given that they appear in most browsers now. To better understand the impact of the visual modifications that Reader View makes on people's reading performance and their subjective preferences (RQ2 & RQ3), we conducted an online study with 391 participants (42 who self-reported having been diagnosed with dyslexia). We found that Firefox's Reader View had a significant effect on reading speed: The readers read content in Reader View 5% faster than on standard websites. In addition, while dyslexic participants consistently rated the readability and user experience of webpages lower than participants without dyslexia, both groups rated the readability and aesthetics of the Reader View higher than the standard webpages. This suggests that the low visual complexity of Reader View websites benefits reading performance, perceived readability, and user experience.

6.2 Related Work

Individual Differences In Reading Abilities

Differences in readers' skills and text comprehension are a result of multiple factors, such as working memory capacity [49], word-identification and comprehension skills [106], and age, which was shown to negatively correlate with reading speed. For example, older subjects (aged 65-75) read significantly slower than younger subjects (aged 25-35) [179] with people's reading speed decreasing between 20 and 88 years of age from 103 to 76 words per minute (WPM) [198]. Lott et al. demonstrated a similar decline in reading speed between ages 58 to 102 [141].

People also experience reading difficulties due to cognitive disabilities. People with dyslexia, for example, have difficulties learning to read and spell and struggle to achieve

Table 6.1: Design Guidelines that have been suggested to improve webpage readability for the average reader and for people with dyslexia. The column on the right indicates whether and how the Firefox Reader View applies these guidelines. * [155] summarized previous research and suggested to avoid formatting texts in large-width columns, which contradicts the other two work cited.

Guideline	**Reference**	**Avg. Reader**	**People with Dyslexia**	**Firefox Reader View**
Use section headings to organize the content	[155]		✓	✓
Limit the amount of content on a page to avoid scrolling	[155]	✓		×
Avoid using italics in the main body of the text	[155]		✓	✓
Avoid underlining large blocks of text	[155]		✓	✓
Use text size larger than 14pt	[155, 191]	✓	✓	✓
Use gray-scale foreground text and off-white background	[155, 192]	✓	✓	×
Use a plain, evenly spaced sans serif font	[155, 190]	✓	✓	✓
Let the user change text, background colors, and enlarge text	[42]	✓	✓	✓
*Increase the number of characters per line (contradict. findings)	[62, 155, 193]	✓	✓	×
Avoid unnecessary images, ads, and animations	[212]	✓		✓
Increase character spacing	[108, 191, 193]	✓	✓	×
Increase image size	[68]	✓	✓	✓

the levels of reading fluency as their non-dyslexic peers [209]. Impairments with any stages of processing, including visual processing (e.g. [87, 108]), phonological and orthographic processing (e.g. [31, 128]), and semantic processing (e.g. [160, 204]) can cause difficulties with reading.

On-screen Readability

Screen reading has been found to be 10-30% slower than reading on paper [124]. While earlier studies (from the pre-tablet era) indicate that on-screen reading performance (measured by traditional metrics such as reading speed, accuracy and comprehension) lags behind paper [55], greater equivalence is being achieved now as computer technology rapidly develops and more sophisticated comparative measures are used (e.g. [122, 165]). However, text that is surrounded by images and advertisements, such as on websites, has been found to reduce reading performance compared to text without. The distraction by even static ads occurs through overt fixations toward ads rather than as covert processing of ads during reading [212].

While website readability has generally been a concern, people with dyslexia consistently rate websites as less readable than people without dyslexia [158]. They perceive websites as too dense and cluttered, with the choice of font type, font size, color, and contrast between colors impacting their ability to find information [158]. Fourney et al. further found that perceived readability is impacted by several lexical and aesthetic webpage features for people with dyslexia, such as average line length, ratio of text appearing in and out of sentences, and average image size [68].

Prior work has developed design guidelines for improving on-screen and website readability through better text presentation for average readers and people with dyslexia (see Table 6.1). For instance, Miniukovich et al. have developed 8 core design guidelines to improve website readability for people with dyslexia, such as by using larger fonts, narrower columns, and avoiding underlining and italics [155]. Many of these guidelines are based on prior work that evaluated how text readability is affected by font properties (e.g. [189, 191]), text styles (e.g. italics, underlining [10, 103]), as well as character, line, and paragraph spac-

ing (e.g. [42,193]). Previous work also evaluated guidelines for average readers (see [62] for a review). However, due to inconsistencies in measurements and populations between studies with people with and without dyslexia, some of these guidelines are contradictory [62], and others lack rigorous evaluations and replications.

Technologies for Improving On-Screen Reading

A growing number of technologies have begun to emerge to improve on-screen reading. The Reader Views available in major web browsers are one example. Microsoft's Learning Tools additionally help users with reading challenges by offering features such as reading text aloud, highlighting text on screen word-by-word, and providing alternative spacing and fonts [153]. Similarly, the word processing software SeeWord provides users with some control over how information is displayed [81]. Many people with reading disabilities rely on tools that read a screen's content aloud (such as *Dragon Naturally Speaking* or even screen reader technologies designed for people with visual impairments) [158]. While this solution is helpful to many users, audio speech is almost twice as slow than the reading speed of an average person [149]. Hence, improving reading rather than providing such workarounds is still advantageous.

6.3 How "Reader View" Changes Websites

To answer our first research question, when and how a browser's Reader View changes a webpage, we analyzed the open-source implementation of Firefox Reader View [159]. We first characterized the steps Reader View takes to transform webpages by inspecting its source code, and then analyzed the visual changes Reader View triggers using a quantitative comparison of website image features. We selected Firefox Reader View because it is the only open-source implementation of the Reader View among major web browsers [6,152], which allows us to better understand any changes it triggers and enables reproducibility of our results. Note that our analysis focused on web browser Reader View of PC devices. We did not compare against mobile view because mobile websites usually contain different page layout, and people behave differently on mobile devices than on a PC.

6.3.1 Characterizing the Reader View in Mozilla Firefox

Mozilla provides an open-source, standalone version of the readability library used for Firefox Reader View [1]. The main file, *Readability.js* (obtained on 07/15/2018), transforms the original HTML Document to a structured, well-formatted one for better readability. Mozilla Firefox lets users toggle between Reader View and the original webpage by pressing a Reader View icon in the address bar. The address bar only appears if a page has been determined to be transformable into Reader View, a process that we describe next.

Transformability Decision

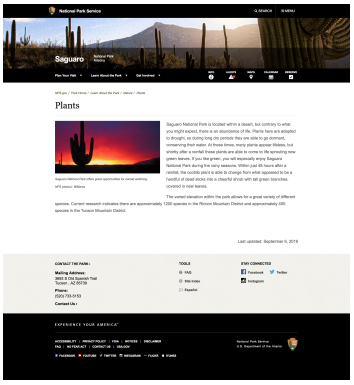
Readability.js first determines whether a webpage has enough textual content to be transformed into its Reader View. The process begins by assembling a list of all document nodes that indicate paragraphs, pre-formatted text, or content divisions that contain line breaks (HTML nodes: `<p>`, `<pre>`, or `<div>` + `<br>`, respectively). It then iterates over each node, adding the character length of the node's text content to an accumulator and skipping nodes unlikely to be content-bearing. Specifically, *Readability.js* skips nodes whose CSS class names or id attributes contain substrings such as "banner," "comment," or "header" – except in cases where the class names or ids also contain substrings such as "article," "body," or "main." *Readability.js* also skips nodes that contain certain list item configurations (HTML elements `<li>` + `<p>`). If the accumulator reaches a value greater than 560 characters at any point, then the iteration terminates and *Readability.js* enables the Reader View button.

Content Decisions

If a webpage was deemed transformable into Reader View, the following steps are triggered to decide on the content that will be included:

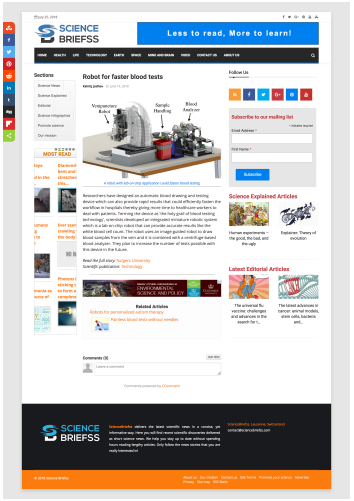
1. Parse an HTML string and build a JavaScript implementation of the document (DOM). *Readability.js* includes its own lightweight DOM parser.

[1] https://github.com/mozilla/readability

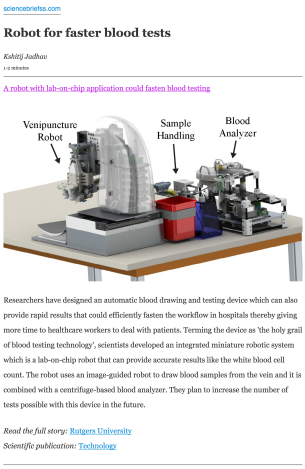


(a) Standard Webpage

(b) Reader View



(c) Standard Webpage



(d) Reader View

Figure 6.1: Two example webpages (a) & (c) and how they are rendered in Firefox's Reader View (b) & (d), respectively.

Table 6.2: Comparison of image metrics between standard websites and their reader view versions. Image metrics were calculated using the VizWeb open-source library [145]. Significance levels: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Image Metric	**Explanation**	**Standard Website** Mean (sd)	**Reader View** Mean (sd)	**t-value (DF=279)**
Visual complexity [187]	A model of perceived website complexity (range 1-10).	6.89 (1.54)	4.25 (0.70)	26.56***
Number of image areas	The number of image areas (adjacent images count as one).	7.35 (5.99)	2.24 (2.30)	13.94***
Number of text groups	The number of horizontal groups of text characters.	21.45 (12.92)	8.15 (2.61)	17.23***
Colorfulness [187]	A model of perceived website colorfulness (range 1-10)	3.97 (1.29)	2.40 (0.58)	20.95***
Saturation	The average pixel value in the HSV color space for saturation.	34.55 (19.75)	7.83 (6.58)	23.03***
Number of quadtree leaves	Recursive division of a website screenshot into quadrants (leaves) using color entropy as a criterion for further division.	73.73 (45.73)	37.84 (8.65)	13.06***

2. Prepare the DOM to be scraped, including stripping "script" and "style" tags, and handling bad markup by replacing two or more successive `<br>` elements with a single `<p>`, and replacing tags `<font>` to `<SPAN>`.

3. Extract the content that is most likely to be the main document content from the DOM tree using a variety of metrics, such as CSS class and id name (e.g., positive class names can be "content" and "main," unlikely candidates have names such as "menu," "sidebar," or "social"), element types (e.g., removing `<tag>` elements and nodes with empty text), and accumulated content score for each node (e.g., adding points for every 100 characters or for any commas within the paragraph) recursively added up from the child nodes in the DOM tree. Consequently, Reader View pages rarely contain menus, advertisements, logos, social sharing buttons, or sidebars, unless these elements are added to the HTML using different class and id names than what *Readability.js* looks for.

4. Prepare the article node for display: Clean out elements such as iframes, textareas, buttons, single-cell tables, and elements that have more images than paragraphs. Then return the content wrapped up in a `<div>`.

5. Finally, run any post-processing modifications to article content as necessary, for instance, fixing relative URLs and cleaning class attributes from every element except for the ones that *Readability.js* sets itself.

Style Decisions

With the content wrapped into one section (`<div>`), Firefox renders the webpage using a predefined style sheet [2]. The default is a Serif font with font-size 20px, content width of 660px, a line-height of 35.2px, and a white background. In addition to adding the Reader View icon to the address bar, which allows users to toggle between the original website and its Reader View, Firefox also provides a settings panel that let users adjust the text size, font, line spacing, and contrast in Reader View, or have a webpage read out loud. Users

[2] chrome://global/skin/aboutReader.css

are able to choose either a Serif or Sans Serif font, choose either a light, dark, or sepia background color, and adjust the font size from 12px (pixel unit) to 28px, the max content width from 440px to 1420px, and the line height from 22px to 57.2px. Note that in CSS the pixel unit does not correspond to a physical screen pixel; instead, it is the angular distance of a hypothetical pixel on a 96dpi screen at a distance of 28 inches (71 cm), which means that any style decisions made by Reader View render similarly on every screen [43]. As a result of these style decisions, Reader View pages are always one column, with images and text blocks stacked vertically on top of each other (Figure 6.1).

6.3.2 Analysis of Reader View Availability

To analyze how often Reader View finds webpages transformable, we randomly selected 100 websites from a dataset of website URLs obtained from the Alexa Top 500 Global Sites [97]. For each of the 100 websites, we selected 10 random child pages (one-level deep), for a total of 1100 page URLs. We then fed these URLs to *Readability.js* to evaluate if and how they might be transformed.

Results

Reader View was only available for 2 (2%) homepages[3] and 406 (41%) child pages from our sample. The low fraction of transformable homepages is because they often consist of little text and instead feature many visual elements, such as brand logos (individual images), navigation (lists), and calls to action (e.g., signup forms), without any sections containing a sufficient number of characters to trigger Reader View. We found that transformable webpages are mostly blogs, news, terms and conditions, or FAQs; in contrast, login pages, e-commerce sites, and pages with embedded content were unlikely to trigger Reader View. Interestingly, one of the homepages in our sample that Reader View did transform (`https://www.vertbaudet.ch/`) is an e-commerce site containing mostly images, which was transformed because it contained a sufficient number of characters in the elements at the bottom of the page.

[3] `http://bagishared.com/` and `https://www.vertbaudet.ch/`

6.3.3 Quantitative Comparison of Visual Changes Between Standard Webpages and their Reader View Versions

The previous section discussed when Reader View triggers changes, but did not yet describe concretely how it changes the visual appearance of websites. This section aims to quantify the changes between webpages and their Reader View.

To enable such statistical comparison between website designs, we analyzed the same 408 websites (2 homepages and 406 child pages) that we found could be transformed into Reader View in the section above. We first took screenshots of these webpages and their Reader View counterparts with a 1024×1280 resolution. We removed 94 webpage pairs where at least one page was less than 1280 pixels in height to ensure the same page dimensions for comparison. This resulted in 280 pairs of webpages for this analysis.

To quantify the visual design of standard webpages and the Reader View, we computed six image metrics (Table 6.2) for each screenshot using the algorithms provided by the open source project VizWeb [145]. We focused on image metrics that have been found to contribute to users' perceived visual appeal of a website as listed in [187] so as to characterize an important dimension of user experience. To analyze the difference between webpages and their Reader Views, we conducted paired t-tests and adjusted for multiple hypotheses testing using the Benjamini-Hochberg method [22].

Results

Our comparison shows that there are significant differences in the visual design of standard webpages and their Reader View counterparts (Table 6.2). Reader View pages are less colorful and less visually complex as shown in Figure 6.1's examples and supported by the statistics in Table 6.2.

Design features that prominently contribute to the perception of visual complexity are the number of image and text areas [187]. On average, Reader Views have only two image areas (reduced from an average of seven) and eight text groups (down from 21). Reader View therefore reduces both images and text groups by around a third.

Correspondingly, Reader View websites are less colorful (per the computational model of

colorfulness). This is mostly a result of two metrics, saturation and the number of quadtree leaves, which heavily influence perception of website colorfulness [187]. Indeed, Reader View websites have a significantly lower saturation value (reduced from an average of 34.55 to 7.83). The reduction in saturation is due to fewer images (i.e., fewer saturated pixels) and a uniform, white background.

Reader View webpages also avoid transitions between regions of different colors. This can be seen from the significant reduction of quadtree leaves from 73.73 to 37.84, suggesting that Reader View almost doubles the use of uniformly colored areas, including white space.

In summary, Reader View significantly reduces the number of images, text groups, the website's overall saturation, and any transitions between regions of different colors. As a result, Reader View pages are less visually complex and colorful than standard webpages. In combination, this might reduce distractions and support users in better focusing on the text — a hypothesis that we test in the next section.

6.4 Online Experiment

We conducted an online study with the aim of evaluating how Firefox's Reader View impacts reading performance, perceived readability, and user experience (RQ2). The study also investigates how people with dyslexia benefit from the Reader View compared to those without (RQ3).

6.4.1 Method

The online experiment was developed as a 10-minute within-subjects study (to account for individual differences in reading skills [188]) with two conditions, Standard Webpage vs. Reader View. The study was launched on the volunteer-based online experiment platform LabintheWild and advertised with the slogan "Test your reading speed!" on the site itself as well as on social media. After completing the study, participants received feedback on their reading speed in comparison to others. Providing this feedback rather than financially compensating participants was meant to attract intrinsically motivated volunteers who have been shown to exert more effort and provide more truthful responses than participants recruited through paid crowd platforms like Mechanical Turk [241].

The study was approved by our IRB, including a waiver of parental consent for minors participating in the study.

Reading Materials

To avoid learning effects from having seen the content of a webpage before, we constructed pairs of similar webpages, which had:

1. the same website category (e.g., news, e-commerce, etc.) with similar topics;

2. a similar reading difficulty as measured by the Flesch-Kincaid Grade Level [118] (a score indicating the U.S. grade level of education required to understand a given piece of text), with a difference less than 1;

3. a comparable length measured by the word count in the title and main content (less than 20% difference);

4. a similar visual design, as measured by having the same number of paragraphs and images.

Webpages were randomly selected from the Alexa Top 500 Global Sites dataset [97], and had to have a Reader View equivalent and use English text. To increase the likelihood that webpages were previously unknown to participants, we excluded those ranked 1-100 (i.e., the most popular ones). Because we had to compare participants' performance with standard webpages to their performance when using the Reader View of a different, but comparable, page (to avoid learning effects), we could not use the same sample from the quantitative analysis in Section 3 where the Reader View equivalents were not guaranteed to exist. Therefore, we manually explored the eligible websites and their child pages, created pairs, and modified the content of the pages by reducing the word count to keep the pairs of webpages comparable in length. We mitigated bias by randomly selecting all pages; those webpage pairs that were not comparable (according to the criteria listed above) were replaced by randomly selecting new webpages until a match was found.

The final stimuli were six webpages (comprising three pairs), of which we took screenshots of their standard view and their Reader View, resulting in 12 webpage screenshots (four examples are shown in Figure 6.1).

Metrics

We used the following metrics to gauge reading performance and user experience:

Reading speed: Calculated as *word count* divided by *adjusted runtime* in words per minute (WPM). *Adjusted runtime* measures how long participants spent on each webpage minus scrolling duration.

Comprehension questions: To check whether participants understood the content of the website and did not merely skim the text, we created three multiple-choice questions per webpage (e.g., "Where is the location of the film production?"). All questions could be answered directly from having read the text; the answers for at least two of the three questions were located in the first and the last paragraph to ensure the need to read the entire article. We tested the questions in a preliminary study with six participants.

Perceived Readability: Calculated based on 7 readability questions on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree) that we adapted from [158]. Participants were asked to rate their agreement with statements such as "Major points were clearly stated" and "It was easy for me to lose my place while reading."

Aesthetics and User Experience: Participants were also asked to rate their level of agreement with 9 statements on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree) using Lavie and Tractinsky's aesthetics questionnaire [127], which subdivides an overall impression of aesthetics into classical aesthetics (e.g., "The webpage has a clean design.") and expressive aesthetics (e.g., "The webpage has a creative design"). The questionnaire is commonly used to evaluate the user experience of websites [144]. We removed one question, "The webpage uses special effects", from the expressive aesthetics scale because it is unlikely to be informative for a static webpage.

Relative subjective duration (RSD): Measures participants' perception of how long it took to read the webpage [48]. Participants were asked "Can you estimate how long you

have spent on reading the webpage (shown below) in minutes?" and had choices in half-minute increments between 0.5 and 10 minutes. Previous work has shown that RSD not only predicts task engagement and task difficulty, but also aesthetic differences: the duration of difficult tasks or tasks with poor aesthetic qualities will be overestimated by participants while the duration of easy, aesthetically pleasing tasks tends to be underestimated [48,126]. We hypothesize that if Reader View improves readability, then participants will underestimate the duration of reading in Reader View relative to the duration of reading the standard webpages.

Demographics: Participants were asked to self-report their age, gender, native language, level of education, whether they have dyslexia (using three options: "no", "yes, I have been diagnosed by a professional," and "yes, but I have not been formally diagnosed" as in [68,158]).

The 12 webpage screenshots and the list of associated comprehension questions and answers for each are provided as supplementary materials.

Procedure

The experiment began with a brief overview of the study, an informed consent form, and a voluntary demographic questionnaire, followed by the task's instructions.

The experiment was split into two parts, one for each condition (Reader View and Standard Webpage). The order of the two conditions was randomized across participants. Participants were shown 3 webpages per condition, for 6 trials in total. Participants were not told about these conditions. The order of webpages was randomized within each condition. Participants were asked to read each webpage word by word and to answer three required comprehension questions immediately after reading each webpage. At the end of each condition, participants were asked to answer the 7 readability questions, 9 user experience questions and 1 RSD question for the last webpage viewed in that condition. A screenshot of the webpage was provided as a reminder. Participants were then given the opportunity to report on any technical difficulties, and to provide any other general comments or questions. The final page showed their personalized reading performance in comparison to others. The

Table 6.3: Average subjective Likert scale measures on a 7-point scale by page condition (Standard Webpage vs. Reader View) and by dyslexia status (self-diagnosed and formally diagnosed dyslexics were grouped together because their ratings did not significantly differ). Mann-Whitney U Tests were conducted to test whether Standard Webpage and Reader View received significantly different ratings, and whether participants with and without dyslexia provided significantly different ratings. Significant scales ($p < .05$) are bolded.

	Standard Website			**Reader View**			
Likert Scale	non-dyslexic μ (M)	dyslexic μ (M)	Cronbach's alpha	non-dyslexic μ (M)	dyslexic μ (M)	Cronbach's alpha	Significance (p value)
Readability	**4.63 (5)**	**3.75 (4)**	.75	**5.27 (6)**	**4.53 (5)**	.77	$< .001$
Classical Aesthetics	**4.11 (4)**	**3.50 (4)**	.91	**5.00 (5)**	**4.56 (5)**	.84	$< .001$
Expressive Aesthetics	2.93 (3)	2.63 (2)	.85	2.83 (3)	2.91 (3)	.86	$=.39$

entire study took 10-12 minutes to complete.

Participants

The experiment was deployed online for 4 months and completed 428 times. We excluded 37 participants who self-reported participating more than once. Our analysis therefore reports on the data of 391 participants.

Participants were between 11-72 years old (M=29.8, SD=12.5) and 55% were female. 69 (18%) participants reported to have dyslexia, of which 42 (61%) had been diagnosed by a professional, and 27 (39%) had not been formally diagnosed. 286 (73%) participants were English native speakers, while the native languages of other participants included 22 other languages. The plurality of participants (41.9%) reported having completed college, 20.2% completed graduate school, and 22.0% high school. The remaining participants were enrolled in professional schools (6.4%), pre-high school (2.6%), or had finished a Ph.D. education (4.6%).

6.4.2 Analysis

Reading Speed

We excluded 98 (of 2346) trials that were completed extremely fast (i.e., the reading speed was higher than $median + 3\times$ the Interquartile Range (IQR) (928 WPM)), indicating that participants might have only been skimming the text or not reading it at all. We also excluded one trial with a reading speed of less than 10 WPM. We further removed 363 trials where participants completed the study on a smartphone because people read and scroll differently on mobile devices than on a PC. To ensure that participants did not solely skim the text, we additionally discarded 111 trials where participants answered less than 2 out of 3 comprehension questions correctly.

We ran a series of linear mixed-effects regressioFn models with (log-transformed) reading speed as the dependent variable and participant as a random variable. Fixed effects were page condition, dyslexia status (non-dyslexic, self-diagnosed, and diagnosed participants), the interaction between page condition and dyslexia status, as well as the control variables screenshot width, word count, age, and native English (i.e., whether a participant reported to be a native English speaker) as fixed-effect variables (Table 6.4). Variables were included based on a comparison of models using the Akaike information criterion (AIC). T-tests (p-values) were calculated using Satterthwaite approximations for degrees of freedom.

Subjective Ratings

For all Likert scale items, we tested internal consistency using Cronbach's alpha [46]. All scales showed high reliability with $\alpha >= .75$ (Table 6.3). We therefore used the averages of participants' responses for each scale.

The subjective ratings for readability and aesthetics were not normally distributed according to both visual inspection of histograms and Kolmogorov-Smirnov tests; hence, we conducted non-parametric Mann-Whitney U tests for analysis. Mann-Whitney U tests also showed that diagnosed and self-diagnosed dyslexic participants did not provide significantly different ratings, leading us to group the two populations for the analysis of subjective questions.

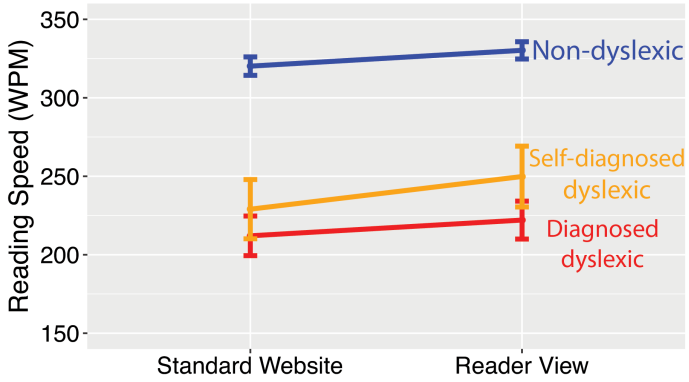


Figure 6.2: Average reading speed across non-dyslexic, diagnosed dyslexic, and self-diagnosed dyslexic participants in Words per Minute for Standard Webpages and their Reader Views. Error bars show standard error.

Relative Subjective Duration (RSD)

To analyze participants' perceptions of their reading duration, we first calculated the difference between participants' estimated and actual reading duration. We conducted an ANOVA comparing the duration difference between the two conditions. Page condition was modeled as a within-subject factor, dyslexia (non-dyslexic, self-diagnosed, and diagnosed participants) as a between-subject factor, and a dyslexia by page condition interaction.

6.4.3 Results

Reading Speed

Our results show that, across all participants and for those trials where participants answered at least two of the three comprehension questions correctly, Reader View significantly increases reading speed by 5% compared to the standard webpages (see estimates in Table 6.4). People who reported having been formally diagnosed with dyslexia read significantly slower than non-dyslexic participants by 43.7% and they are also significantly slower than those who reported having dyslexia but who haven't been formally diagnosed.

We did not find a significant interaction effect between Reader View and Dyslexia,

suggesting that all participants benefit from the Reader View at a similar rate (i.e., the slope of improvement between the two conditions is similar, albeit slightly steeper for self-diagnosed dyslexics). Non-dyslexic participants increased their reading speed from 320 to 330 WPM (college-educated adults have been previously found to have a reading speed of about 244 to 460 WPM when reading on screen [61]), while self-diagnosed dyslexics and diagnosed dyslexics increased it from 229 to 250 WPM and 212 to 222 WPM, respectively (Figure 6.2).

Age, word count, and native language also significantly impacted reading speed. With every year of age, reading speed decreases by 1%, confirming the findings of prior work [179]. Every additional word increases the reading speed by 0.1%; and being a native English speaker significantly increases reading speed by 23%. None of these factors interacted with Reader View, suggesting that Reader View does not differentially impact people of various ages or with different language backgrounds. Screenshot width also did not impact the reading speed.

Participants' comprehension scores did not differ between page conditions; instead, we found a speed-accuracy trade-off with participants who read faster answered less questions correctly. This likely removes the effect of page condition on comprehension scores.

Subjective Ratings

As shown in Table 6.3 and Figure 6.3, participants rated the readability of Reader View pages significantly higher than standard websites (Figure 6.3a). They also felt that the Reader View had superior classical aesthetics (e.g., "clean," "pleasant") compared to standard webpages (Figure 6.3b), which suggests that the Reader View follows design rules that are thought to improve usability [127]. Ratings on expressive aesthetics (e.g., "fascinating," "creative") did not significantly differ between the conditions (Figure 6.3c). In contrast, we would have expected Reader View pages to receive lower ratings on expressive aesthetics than standard websites, given that the concept measures the creativity and originality of a design [127].

As Figure 6.3a shows, people with dyslexia rated webpages as significantly less readable

Table 6.4: The results of a linear mixed-effect model predicting log reading speed.

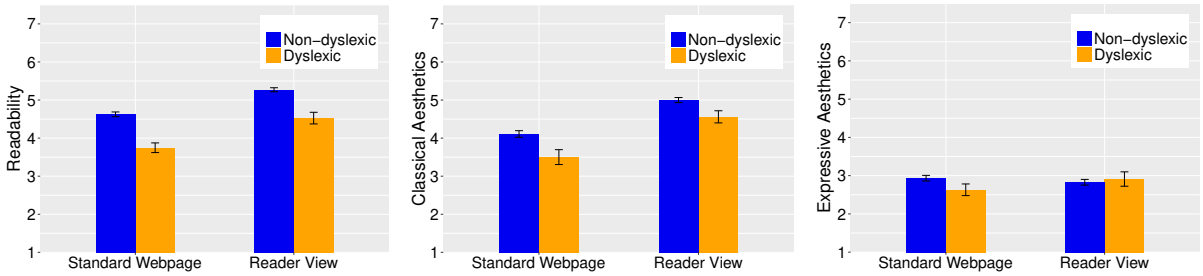
Variable	**Est.**	**SE**	**t-value**	$Pr(>\|t\|)$
(Intercept)	5.593	0.17	32.85	$< .001$ ***
Reader View [yes]	0.048	0.01	3.60	$< .001$ ***
Dyslexia [diagnosed]	-0.437	0.07	-6.04	$< .001$ ***
Dyslexia [self-diagnosed]	-0.371	0.10	-3.86	$< .001$ ***
RV × Dys [diagnosed]	-0.014	0.04	-0.34	=.74 (n.s.)
RV × Dys [self-diagnosed]	0.073	0.06	1.31	=.19 (n.s.)
Screenshot width	0.00	0.00	-0.56	=.58 (n.s.)
Word count	0.001	0.00	24.96	$< .001$ ***
Age	-0.010	0.00	-6.05	$< .001$ ***
Native English [yes]	0.226	0.05	4.88	$< .001$ ***

compared to those without dyslexia in both conditions. They also provided lower ratings on classical aesthetics (Figure 6.3b), indicating that they perceive both standard and Reader View webpages as significantly less aesthetically pleasing (and, thus, usable [127]) than non-dyslexics.

Relative Subjective Duration (RSD)

Participants overestimated their reading time by an average of 67s (143%) in the Reader View and 64s (142%) for Standard Webpages (the difference is not statistically significant, $F_{455} = 0.1, p = 0.75$).

Dyslexia had a significant main effect on the perceived reading duration ($F_{455} = 4.65, p < 0.05$). A Welch's two sample t-test suggests that the standard webpages led people diagnosed with dyslexia and those without to similar over-estimations of the reading duration (non-dyslexics mean = 63s, sd = 85s, dyslexics mean = 107s, sd = 92s, $t_{21} = 2.01$, $p = .06$). However, people formally diagnosed with dyslexia over-estimated the duration of the reading



(a) Average perceived readability. (b) Average classical aesthetics. (c) Average expressive aesthetics.

Figure 6.3: Average ratings of perceived readability, classical and expressive aesthetics for standard and Reader View webpages by non-dyslexics and people with dyslexia (self-diagnosed and formally diagnosed dyslexics were grouped together because their ratings did not significantly differ). Error bars represent standard error.

(m = 104s, sd = 100s) more than people without (m = 59s, sd = 73s, $t_{29} = 2.20$, $p < .05$) for the Reader View. Self-diagnosed participants did not differ from either non-dyslexics or people who had been formally diagnosed (Figure 6.4). There was no page condition by dyslexia interaction effect on RSD, meaning that participants' perceived task duration changed at similar rate between Standard Webpages and Reader Views.

6.5 Discussion

Our work is the first to demonstrate that Firefox's Reader View converts websites into pages that significantly improve reading speed, perceived readability, and perceived classical aesthetics (suggesting that the design is perceived as cleaner and more in line with recommendations for usable designs [127]). In particular, our participants improved their reading speed by 5% compared to standard websites on average. This is a modest, albeit significant improvement in reading speed: It is less than what speed reading training could achieve, for example, but is unlikely to impact a speed-accuracy trade-off that can result from speed reading [180]. However, we found that people with dyslexia did not improve their reading speed more from the Reader View than those without, and their overall reading

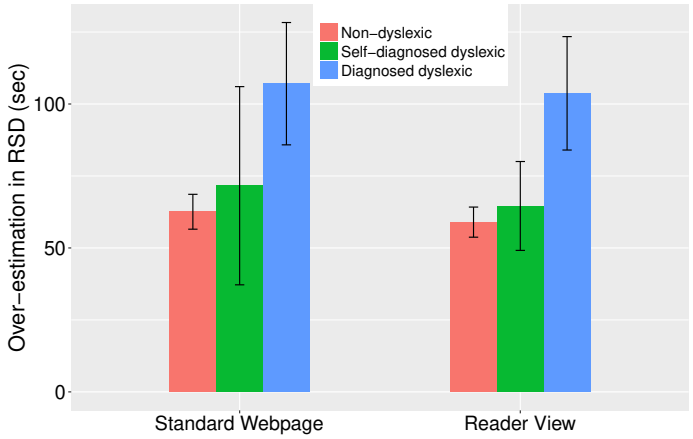


Figure 6.4: Over-estimation of participants' perception of reading duration (RSD) relative to the actual time spent reading a webpage. Error bars represent the standard errors.

speed remained well below that of non-dyslexic participants.

Our findings also showed that participants rated Reader View as significantly higher on classical aesthetics than standard webpages, while ratings on expressive aesthetics (a concept that measures the creativity and originality of a design [127]) did not significantly differ between Reader View and standard webpages. This suggests that Reader View conforms to usability-related design guidelines, such as being clean, clear, or symmetric [127], but its page design is not perceived as more sophisticated, fascinating, or creative than standard webpages. Hence, creatively designed websites might be valuable for an overall user experience, but detrimental for focused reading.

6.5.1 Design Changes that Improve Reading

Our work suggests that text-heavy websites ought to be designed with lower visual complexity than currently the case. Our analysis of the Reader View source code and quantitative comparison of visual changes that the Reader View triggers revealed what might cause the improvements in reading that our study showed: For one, Reader View reduces the number of text groups and images by a third. Its algorithm strips websites of any content deemed unnecessary for focused reading, including advertisements, menus, and logos. Images and

text blocks are stacked on top of each other in a one-column layout rather than distributed across multiple columns as in many websites. Reader View also doubles areas that are uniformly colored, mostly by converting any background color to white space. As a result, Reader View pages are significantly less colorful and less visually complex than standard websites.

Previous studies have produced inconclusive results as to whether a one-column layout improves reading speed compared to a two- or three-column layout (see, e.g., [45] vs. [63]). However, most prior work suggests that longer line lengths, as measured by the number of characters per line, result in faster reading [62, 193]. This suggests that Reader View's decision to use a larger content width than most standard websites plays a role in the reading speed improvement we have seen. It is also likely that Reader View's reduction of images contributes to an improved reading speed, since static images and ads distract reading [212]. Reader View's decision to display images and text in the same column means that users rarely see more than one image at once, which might have resulted in our participants being better able to avoid fixation on images [212].

6.5.2 Should Reader View be the Standard View?

We were surprised to find that only 2% of homepages and 41% of child pages (from a randomly selected sample of 100 website URLs) were available in Reader View. Our analysis of Reader View's source code showed that this is because homepages often do not contain a sufficient amount of text in a given section to trigger the Reader View. Websites that most often get converted into a Reader View are blogs and news websites, which usually have high word counts in the main content. However, due to the low availability of websites that do get transformed into Reader View, only few people benefit from the Reader View.

Given that we saw considerable improvements in reading speed, perceived readability, and classical aesthetics, one could argue that webpages with high word counts, such as blogs and news websites, should be presented in their Reader View by default, or even designed following the Reader View style sheet. This would mean that users are less likely to be exposed to advertisements (a change that would impact the revenue model of many websites) and less likely to be subjected to elements of branding. The latter could negatively

affect a company's brand memorability, but also the users' orientation on the web since logos, color schemes, and other forms of branding can serve as navigational cues ("Am I still on the same website?"). Hence, defaulting to the Reader View (or equivalent design choices) might neither be feasible from a company's perspective, nor particularly beneficial to the user – unless they are in fact intending to read the content.

One design suggestion is that search engines could either indicate which pages offer Reader View (e.g., by including the icon next to the page in the result list), or allow a way for filtering for "readerable" pages. Users could then more easily benefit from Reader View pages; in addition, explicitly signaling Reader View availability could encourage developers to make their websites Reader View compatible.

6.5.3 Designing for Reader View

Our results provide insight into what triggers Reader View and what prevents a webpage from being transformed. This can help web designers and developers to know how to design more "readerable" websites. In addition, we found that Firefox's Reader View occasionally removes content that was essential for an article, or fails to remove content that is irrelevant. Our inspection of Firefox's Reader View source code revealed the tags that are used to determine what to include and exclude, which can support developers in creating websites that will trigger the Reader View (or avoid it). However, other major web browsers do not openly publish their source code and do not provide more than a light description of their Reader View features. This leaves web developers to guess what might or might not trigger the various implementations of the Reader View features in different browsers. While we hope that developers of the Reader Views in Edge, Safari, and other web browsers will provide better guidelines for developers to know what tags to include or exclude, we realize that openly revealing this information can also result in adversarial behavior. For example, developers might deliberately change tags to avoid advertisements from getting stripped out of the Reader View. To prevent this, companies could only reveal specific algorithmic rules, or they could provide tools that give developers feedback on their code without explicitly revealing any such rules.

6.5.4 Additional Support for People with Dyslexia

Our study showed that the Reader View improved the reading speed and perceived readability for participants with dyslexia. Perhaps most importantly, they perceived the readability of Reader View pages as equally high as non-dyslexic participants perceived the readability of standard websites. This indicates that the Reader View can be a step toward equalizing the reading experience between users with varying reading abilities, at least to some extent. However, Reader View does not bring people with dyslexia to the same level of performance as people without dyslexia. One finding that has been suggested by prior research as beneficial for people with dyslexia but that has not yet been leveraged in the Reader View is adding additional spacing between the letters within words [108]. This is helpful for those dyslexics who experience crowding problems—for them, the letters within words will appear jumbled. For them, adding additional spacing between the letters has led to an increase in reading speed and accuracy [108].

6.6 Limitations and Future Work

Our study was not designed to disentangle the effect of specific design decisions on reading, but putting our results in context with prior work sheds light on which design changes that Reader View makes led to an improvement in reading speed. Future work could systematically explore which of the design changes result in the most noticeable improvement.

Another limitation of our study is that we only looked at websites as viewed on tablets, laptop or PC screens, but excluded mobile views and thus, smaller screen sizes. Websites that use a responsive layout often resemble the Reader View on a mobile screen, which is an interesting comparison for future studies.

Finally, our sample was not large and diverse enough to closely analyze the effects of demographics on perceived aesthetics and user experience that previous work has shown [185]. Therefore, the simple design inspired by the Reader View should be adopted with caution. We are excited to explore Reader View alternatives that improve both perceived aesthetics and reading speed over standard websites for people from various backgrounds.

6.7 Conclusion

This paper explored how Firefox's Reader View impacts reading fluency and user experience compared to standard websites. Our analysis of the source code and quantitative comparisons of Reader View's page designs with standard webpages showed that Reader View reduces the colorfulness and visual complexity of webpages, removing a third of images and text groups on average. An online study comparing Reader View with standard webpages showed that Reader View's design changes result in significant improvements in reading speed, perceived readability and aesthetics for people with and without dyslexia. However, we found that only 2% of homepages and 42% of child pages could be transformed into Reader View, suggesting that few websites provide this benefit. Our work is the first to systematically characterize how Reader View works, what improvements it achieves, and how this differs between people with varying reading abilities.

6.8 Datasets

We make available the dataset used for quantifying visual differences between standard and Reader View websites, the dataset from our online study, and the R-code for analysis at `https://github.com/QishengLi/CHI2019_Reader_View`.

Chapter 7

THE VIRTUAL CHINREST

7.1 Introduction

Psychophysical methodologies have been extensively applied to study human perception and performance in healthy adults, and to study individual differences across participants and in relation to a variety of clinical conditions. Yet most psychophysical studies are constrained to the laboratory because of the need to rigorously control visual stimulus presentation with the help of a physical chinrest. Given the difficulty of bringing participants into a lab, these studies generally rely on small samples and can risk generalizability to the larger population.

To conduct studies with larger and more diverse samples, researchers have developed and evaluated alternative ways to recruit participants, such as through the online labor market Amazon Mechanical Turk (MTurk) (e.g., [47, 89]) or through volunteer-based online experiment platforms such as LabintheWild [184]. Compared to traditional laboratory experiments, such online studies offer faster and more effortless participant recruitment [23, 80, 101, 148] and have resulted in large-scale studies comparing multiple demographic groups, ages, languages, and countries [85, 86, 135, 184, 185]. A growing body of literature has explored methodologies for conducting a broad range of experiments, and shown that online experiments yield results comparable to those obtained in conventional laboratory settings [15, 47, 52, 77, 135, 181, 182, 183, 186, 211, 243].

For instance, online experiments have been shown to accurately replicate the findings from behavioral experiments that rely on reaction time measurement [15, 47, 181, 182, 211, 243], rapid stimulus presentation [47, 77, 186] and learning tasks with complex instructions [47]. De Leeuw and Motz conducted a visual search experiment with interleaved trials implemented in both the Psychophysics Toolbox (in lab) and JavaScript (online) and showed that both software packages were equally sensitive to changes in response times [52]. Similarly, Reimers and Stewart demonstrated that two major ways of running experiments

online, using Adobe Flash or JavaScript, can both be used to accurately detect differences in response times despite differences in browser types and system hardware (machines) [183]. Researchers have also investigated if web-based within-subjects experiments studying visual perception can accurately replicate prior laboratory results [89, 140]. These online experiments replicated prior laboratory results despite not being able to control for participants' viewing distance and angle.

To the best of our knowledge, no prior work has investigated whether laboratory results of studies using a physical chinrest can be replicated online for between-subjects experiments in which metrics are being compared across participants, and therefore require tight control of a participants' viewing distance. To fill this gap, we developed the Virtual Chinrest, a novel method to accurately measure a person's viewing distance through the web browser. To estimate an individual's viewing distance, we measure the eccentricity of their blind spot location. We show that our method enables remote, web-based psychophysical experiments of human visual perception by making it possible to automatically adjust stimulus size and location to a participant's individual viewing distance.

7.2 The Virtual Chinrest

Our approach includes two tasks, first estimating an individual's screen resolution followed by their viewing distance:

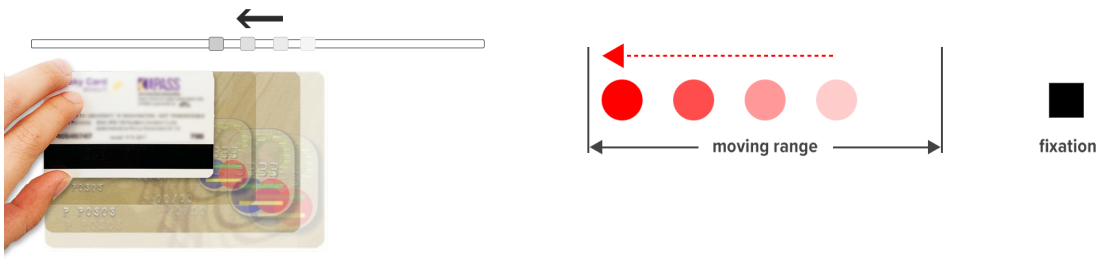
Screen Resolution

One challenge for conducting psychophysical experiments in the web browser is that the resolution and size of the display are unknown, prohibiting control of the size and location of stimuli presented to participants. To estimate the screen resolution, we calculate the *logical pixel density (LPD)* (in pixels per mm) of a display using a card task. We adopted a method that is already commonly used on the internet to help people measure items on the screen: As shown in Fig. 7.1a, we ask participants to place a real-world object (in our case a credit card or a card of equal size, which are standardized in size and widely available) on a specific place on the screen. Participants can adjust a slider until the size of an image of the object on the screen matches the real-world object. We then calculate the ratio between

the card image width in pixels and the physical card width in mm to obtain the LPD in pixel per mm: $LPD = cardImageWidth/85.60$ where $cardImageWidth$ is the width of the card image in the web browser in pixels after the participant adjusted the slider and 85.60 mm is the width of the card in the real world. Knowing the LPD, we can present online participants with stimuli of a precise size in pixels (on-screen distance) independent of their individual display sizes and resolutions where:

$$\text{LPD (px/mm)} = \frac{\text{On-screen Distance (px)}}{\text{Physical Distance (mm)}} \tag{7.1}$$

We will use this ratio (LPD) to convert between the on-screen distance and physical distance in the following calculation of the viewing distance.



(a) Card Task: Participants are asked to place a credit card or a card of equal size on the screen, and adjust the slider until the size of the image of the card on the screen matches the real-world card. We can therefore calculate the logical pixel density (LPD) of the display in pixels per inch to estimate distance s in Fig. 2.

(b) Blind Spot Task: Participants are asked to fixate on the static black square with their right eye closed while the red dot repeatedly sweeps from right to left; they are asked to press the spacebar when they perceive the red dot as disappearing. We then calculate the distance between the center of the black square and the center of the red dot when it disappears from the eye sight.

Figure 7.1: Card Task and Blind Spot Task procedures that are used to calculate the viewing distance using the Virtual Chinrest.

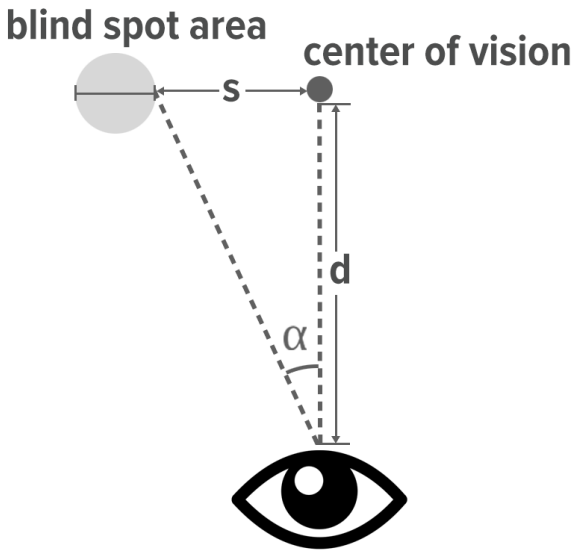


Figure 7.2: Trigonometric calculation of a participant's viewing distance using the human eye's blind spot. Knowing the distance between the center of display and the entry point of the blind spot area (s), and given that α is always around 13.5°, we can calculate the viewing distance (d).

Viewing Distance

The most critical issue for web-based online psychophysical experiments is how to control stimulus geometry given unknown viewing distance. To tackle this issue, we devised a method in which we leverage the fact that the entry point of the optic nerve on the retina produces a blind spot where the human eye is insensitive to light. The center of the blind spot is located at a relatively consistent angle of $\alpha = 15°$ horizontally ($14.33° \pm 1.3°$ in Wang et al. [234], $15.5° \pm 1.1°$ in Rohrschneider [199], $15.48° \pm 0.95°$ in Safran et al. [200], and $15.52° \pm 0.57°$ in Ehinger et al. [64]). Given this, we can calculate an individual's viewing distance from simple trigonometry, as shown in Fig. 7.2. More precisely, the LPD obtained from the card task lets us calculate the physical distance s by equation 7.1. Once we have detected the blind spot area, we can then calculate the viewing distance d.

Inspired by different educational blind spot animations existing online (e.g. [41]), we

designed and developed a browser-based blind spot test to estimate the physical distance between one's blind spot area and the center of display. Participants are asked to fixate on a static black square with their right eye closed while a red dot moves away from fixation (Fig. 7.1b). The red dot repeatedly sweeps from right to left. At a certain point on the display, the participant will perceive the dot as if it were disappearing. The participant is instructed to press a button when the dot disappears. We then calculate the distance between the center of the black square and the center of the red dot (when it disappears from the eye sight). Instead of using $\alpha = 15°$ as the average horizontal blind spot location as found in previous work [64, 199, 200, 234], we use 13.5° as the average blind spot angle because 1) our method captures the entry point of blind spots (of which the angle should be smaller) instead of the blind spot center, and 2) we calibrated our method by conducting preliminary experiments with a few participants and found that using 13.5° provided us with the most accurate results. The complete formula to calculate the individual viewing distance is:

$$\text{Viewing Distance } (d) = \frac{\text{Physical Distance } (s)}{\tan(\alpha)} \tag{7.2}$$

7.3 Methods

Lab study

In Exp.1, 19 participants completed the Virtual Chinrest experiment (consisting of the card task and blind spot test) using a physical chinrest in a psychophysical experiment room. Each participant completed the experiment once. In Exp. 2, 12 new participants performed the same experiment, but without a physical chinrest. The 2×3 within-subject experiments used two different-sized screens (13" and 23") and three seating distances: 43, 53, and 66 cm (17, 21, and 26 inch). We chose 53 cm because it was the distance used in the original laboratory study, and we chose 43 cm and 66 cm because the International Organization for Standardization (ISO) guidelines suggests distances between 40 cm and 75 cm are reasonable choices [102]. The 6 conditions were counterbalanced across participants.

Setup

Participants completed the experiment in a room with controlled artificial lighting. Exp. 1 was conducted using a 24" monitor (Model: LG 24GM77-B) with a resolution of 1920×1080. The viewing distance was set to 53 cm (21 inch). Exp. 2 was conducted with two monitors: a 13" Macbook Pro with a resolution of 2560×1600 pixels and a 23" monitor (Model: HP Compaq LA2306x) with a resolution of 1920×1080 pixels. To perform the card task, participants were provided a card of size 85.60 x 53.98 mm (a standard credit card size) in both experiments. The setup remained the same throughout the entire experiments.

Procedure

Both experiments asked participants at the beginning to assume a comfortable position and to keep this position throughout the experiment. The experiment started with an informed consent form, a demographic questionnaire, followed by the Virtual Chinrest experiment consisting of two tasks. During the blind spot task (Fig. 7.1b), participants were instructed to press the spacebar as soon as the red dot disappears from their left eyesight and repeat this process 5 times so that later we calculate the viewing distance by taking the average of the results. The participants in Exp. 1 (with chinrest) only completed the tasks once while participants in Exp. 2 (without chinrest) completed the tasks in all 6 conditions. Completion of the experiment in each condition took approximately 4 minutes. All experimental sessions were approved by the University of Washington Institutional Review Board and performed in accordance with the relevant guidelines and regulations.

Participants

A total of 19 participants and another 12 distinct participants completed the experiments in Exp.1 and Exp. 2, respectively. All of the participants were recruited from a local university, and all self-reported having normal or corrected-to-normal vision. Written informed consent was obtained from all participants.

Analysis

For the analysis of Exp. 2, we removed one participant who did not successfully complete the entire experiment. We also removed one data point of another participant who did not correctly complete the card task in the condition of [66 cm, 13"].

Online experiment

The online experiment was launched on the volunteer-based online experiment platform LabintheWild and advertised with the slogan "How accurate is your peripheral vision?" on the site itself as well as on social media.

Experimental Design

During each experimental session, we first presented the Virtual Chinrest experiment and used the results to calculate individual's viewing distance and to calibrate the stimuli's size and locations. Instead of creating stimuli (demonstrated in Fig. 7.5) using MATLAB, we created the stimuli as SVG on HTMLs and manipulated the stimuli using JavaScript. All the elements were created in a container with width of 900 pixels on the webpage. In the blind spot test, the dot was drawn in red with a diameter of 30 pixels, and the fixation square was drawn in black with a side length of 30 pixels (Fig. 7.1b). Replicating the original crowding study [108] in the unit of visual degrees, stimuli comprised four flankers — open circles with 1° diameter and a target — an open circle with a gap (target; an arc with reflex central angle of 330°). All stimuli were black and displayed on a white background (Fig. 7.5). Two conditions of target eccentricity (the center-to-center distance between the fixation mark at the center of the webpage and the target) were 4° and 6°. In each crowding experiment session, each participant was randomly assigned one target eccentricity, and the target eccentricity was fixed with the starting target-flanker distance being set as 1.3 times greater than half the eccentricity (3.9° for 6° eccentricity; 2.6° for 4° eccentricity).

During each crowding experiment session, the subsequent target-flanker distances (25 trials/steps in total) was controlled by the 1-up 3-down staircase procedure implemented in JavaScript [https://github.com/hadrienj/StaircaseJS]. On a given trial, the fixation mark

was displayed first and remained on the webpage for the entire session. After 500 ms of fixation onset, the stimuli were displayed either to the left or the right of the fixation for 150 ms. Only the fixation remained on the webpage until the participant submitted a response by using the arrow keys on the keyboard to indicate the direction (up or down) of the target gap. No feedback was provided during the experiment. There was a 500 ms blank between a participant's response and the beginning of the next trial.

The visual crowding, defined as the minimal center-to-center distance between a target and the flankers (in degrees), was used to quantify the crowding effects when participants could report the target identity at certain accuracy. Since we are using a 1-up 3-down staircase procedure, participants should be able to correctly report the target identity 79.4% of times.

Procedures

The experiment began with a brief overview of the study, an informed consent form approved by the University of Washington Institutional Review Board, and a voluntary demographic questionnaire, followed by the card task and the blind spot test with 5 trials to calculate participants' viewing distances. Participants were then presented the instruction of the crowding tasks and a practice session with 5 trials.

The main experiment was split into two blocks (two independent staircases, 25 trials each), and each was followed by another blind spot task with 3 trials. After the last blind spot test, participants were then given the opportunity to report on any technical difficulties, and to provide any other general comments or questions. The final page showed their personalized "crowding effect" in comparison to others. The entire study took 10-12 minutes to complete. All experimental sessions were approved by the University of Washington Institutional Review Board and performed in accordance with the relevant guidelines and regulations.

Participants

The experiment was deployed online for 15 months and completed 1198 times. We excluded 45 participants who self-reported participating more than once. Our analysis therefore reports on the data of 1153 participants. Informed consent was obtained from all participants.

Participants were between 7-71 years old (mean = 26.3, sd = 12.4) and 50.2% were female. 229 participants reported to have cognitive impairments, including dyslexia, learning disability, reading difficulties and Attention Deficit Disorder (ADD). 69 (6.0%) of all participants reported to have dyslexia. The plurality of participants (32.9%) reported having completed college, 21.3% completed graduate school, and 19.8% completed high school. The remaining participants were enrolled in professional schools, pre-high school, or unspecified.

Analysis

We deployed the online study in two stages, where we added more granular data log at the second stage, such as the percentage correctness of the experiment and the results of each individual trial. Therefore, the analysis of visual crowding effects (Fig. 7.7 a, b) was performed on the data of 793 participants from the second stage, the results in Table 7.2 was based on a subset of 570 participants who have explicitly reported whether they have dyslexia and/or other related impairments, while the results of the viewing distances from the three blind spot tests (Fig. 7.7 c, d, e) were reported from all 1153 participants.

We checked for data normality by both the visual inspection of histograms and the Shapiro-Wilk normality tests before each analysis. We then conducted parametric (e.g. the Welch two sample t-test) and non-parametric (e.g. Mann-Whitney U test) analysis accordingly. In the linear mixed-effects regression models, t-tests (p-values) were calculated using Satterthwaite approximations for the degrees of freedom.

The data analysis of all the experiments was performed in R, with the help of multiple packages [19, 125, 194, 238].

7.4 Results

Validation Experiments in the Lab

We conducted two controlled lab experiments to verify that our Virtual Chinrest method is valid and accurate.

Exp. 1. Validation of the Virtual Chinrest with a Physical Chinrest

The aim of our first experiment was to compare the accuracy of the viewing distance calculated with our Virtual Chinrest method to the viewing distance defined by a physical chinrest. Nineteen participants took part in the experiment with a physical chinrest, fixing their viewing distances at 53.0 cm. The experiment was implemented in JavaScript and run in the web browser; the two tasks of the experiment are schematized in Fig. 7.1a and 7.1b.

To our surprise, despite unavoidable sources of error such as variability of the blind spot location and of the response when the dot disappears, the viewing distance estimates were 53.0 $\pm$ 0.69 cm (mean $\pm$ standard error of the mean (sem)), which is very accurate given the physical viewing distance of 53.0 cm. The average absolute error was 2.36 cm.

Exp. 2. Distance Calculation with Different Display Sizes & Viewing Distances and No Physical Chinrest

In Exp. 2, we tested the accuracy of the Virtual Chinrest method when systematically changing the display size and participants' viewing distances. Participants did not use a physical chinrest in this experiment; instead, we controlled for participants' seating distances (defined by the distance between the center of the chair and the center of the display), but not for the exact viewing distances or potential head and upper body movements. This allowed us to validate the Virtual Chinrest in a more natural setting, with participants sitting in front of the computer as they would at home.

Twelve participants took part. We adopted a within-subject experimental design with the seating distance and the display size as two factors. The seating distance had three levels, 43, 53, and 66 cm, and the display size had two levels of 13" and 23". Participants were instructed to complete the same experimental procedure as Exp. 1 in 6 (3×2) conditions.

Actual Distance (cm)	**Estimated Distance Mean (Avg. Abs. Err)**		
	13”	*23”*	*Average*
43	47.2 (4.6)	44.3 (2.4)	45.8 (3.5)
53	55.7 (4.2)	53.6 (1.6)	54.7 (2.9)
66	63.7 (2.4)	61.7 (4.4)	62.6 (3.4)

Table 7.1: Calculated viewing distances of each condition using Virtual Chinrest in Exp. 2, a 3×2 within-subject lab study.

Our results show that the Virtual Chinrest detects users' seating distance (as a proxy for viewing distance) with an average absolute error of 3.25 cm (sd = 2.40 cm). Table 7.1 and Fig 7.3 present the results of the different conditions: among the 3 distances, the viewing distance of 53 cm was predicted most accurately with an average absolute error of 2.88 cm (mean $\pm$ sem = 54.7 $\pm$ 0.76 cm). The viewing distance of 43 cm was predicted least accurately with an average absolute error of 3.46 cm (mean $\pm$ sem = 45.8 $\pm$ 0.74 cm). We found that the viewing distances were over-estimated by 1.4 cm when the larger display (23”) was used and underestimated by 0.86 cm when the smaller display (13”) was used. A paired t-test confirmed this difference is statistically significant ($t_{(31)} = -4.56, p < .001$). However, despite the small amounts of bias introduced by these different conditions, the overall accuracy was still very high.

Horizontal Blind Spot Location Estimation

Both Exp. 1 and Exp. 2 in which we controlled for viewing distances or seating distances allowed us to calculate participants' horizontal blind spot locations based on the data from

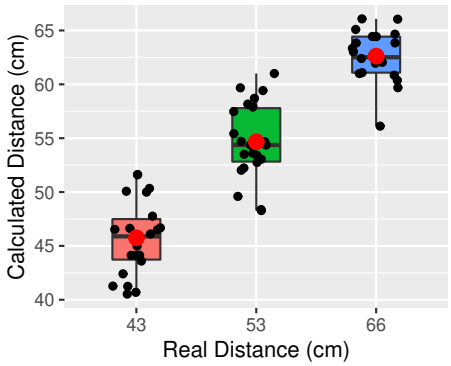


Figure 7.3: The box plot and the 12 individual viewing distances calculated using Virtual Chinrest in three distance conditions (43, 53, and 66 cm or 17, 21, and 26 inch) in Exp. 2. The red dots represent the calculated mean distance in each condition. The average absolute error is 3.25 cm (sd = 2.40 cm) across all three conditions.

the blind spot tasks. Combining the data from both experiments, the mean horizontal blind spot entry point location is 13.59° (min = 11.53°, max = 16.01°) with a SD of 0.96°. The distribution of the estimated blind spot locations is plotted in Fig. 7.4. Since the mean blind spot diameter is around 4.5° [64, 234], the center of the blind spots from our results is then $15.84° \pm 0.96°$, comparable to prior work, in which the blind spot locations are ranged between 14.33° and 15.52° [64, 199, 234].

The discrepancy of the average blind spot locations from previous studies (e.g. [199, 234]) and our own finding that the horizontal blind spot locations ranged between 11.53° and 16.01° may suggest that any minor inaccuracies in our viewing distance estimation was caused by variations across individuals' blind spot locations.

Online Replication of a Laboratory Study on Visual Crowding Using the Virtual Chinrest

In Exp. 1 and Exp. 2 we have demonstrated that the Virtual Chinrest is highly accurate in measuring the viewing distances, even in relatively uncontrolled settings with variable viewing distances, display sizes, and potential movements of the head and upper torso. This

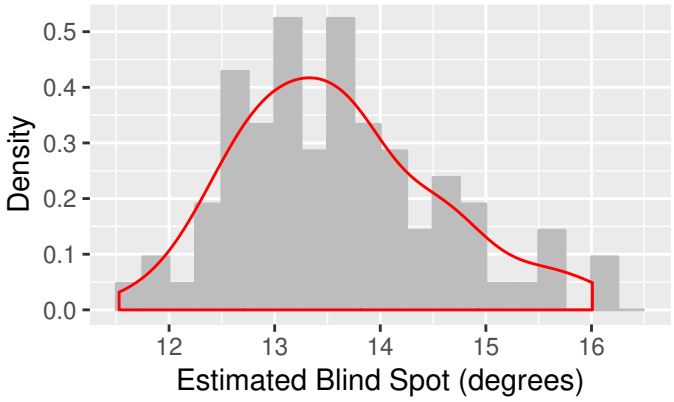


Figure 7.4: The distribution of the estimated horizontal blind spot entry point locations (mean = 13.59°, sd = 0.96°) of 30 participants, 85 experimental sessions from Exp. 1 and Exp. 2.

allows us to further examine whether we can use the Virtual Chinrest to conduct the type of studies that would typically rely on a physical chinrest, in an uncontrolled online environment. We aim to replicate classic findings from psychophysical experiments measuring visual crowding (e.g. [108,130,171,227]) — studies that require precise control over the retinal location of stimuli. Visual crowding is a phenomenon that occurs in peripheral vision where an observer's ability to identify a target is greatly reduced when the target is flanked by nearby objects. Using the visual crowding paradigm, we can measure individual differences of visual crowding effects, i.e., how much distance between the target and flankers one needs to correctly identify the target. These individual differences in low-level visual processing have been related to high-level cognitive function such as reading ability [30,147]. Measuring an individual's crowding effect requires being able to (a) present the target at the same eccentricity and (b) manipulate the distance between the target and flankers using the same units (i.e., in visual angle) across individuals. Thus, without knowing the viewing distance and the display size, it is impossible to measure an individual's crowding effect.

We developed a version of the visual crowding experiment (see stimuli in Fig. 7.5) as a 10-minute online test that began with setting up the Virtual Chinrest (by asking participants

to perform the card task and the blind spot test). Each participant was randomly assigned one target eccentricity, 4° or 6°. Participants were then presented instructions for the visual crowding experiment and asked to perform a practice session with 5 trials. The main experiment was split into two blocks and each block was followed by another blind spot task to determine whether participants have changed position.

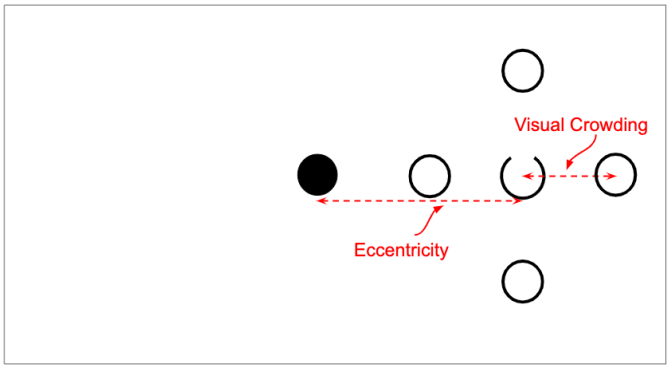


Figure 7.5: The main stimuli used in the visual crowding experiments (Exp. 3 and Exp. 4): After 500 msec of fixation on the central mark, crowding stimuli appeared at either the left or the right side of the display. The stimuli disappeared after 150 msec and participants reported the direction of the gap (up or down) using the keyboard.

Exp. 3. Validation of Browser-based Measurements of Crowding

Our first goal was to ensure that our browser-based implementation of the visual crowding experiment can gather high quality data with and without a Virtual Chinrest. To do so, we conducted a within-subjects laboratory study in which 19 participants took the experiment (with the same target eccentricity assigned to each of them) in two conditions: (1) using a physical chinrest set to a viewing distance of 53 cm and (2) using the Virtual Chinrest on a laptop in their desired position (i.e., on the lap or on a desk) and desired distance. The latter condition was intended to simulate an in-situ environment that participants might find themselves in when participating in an online experiment. We compared an individual's

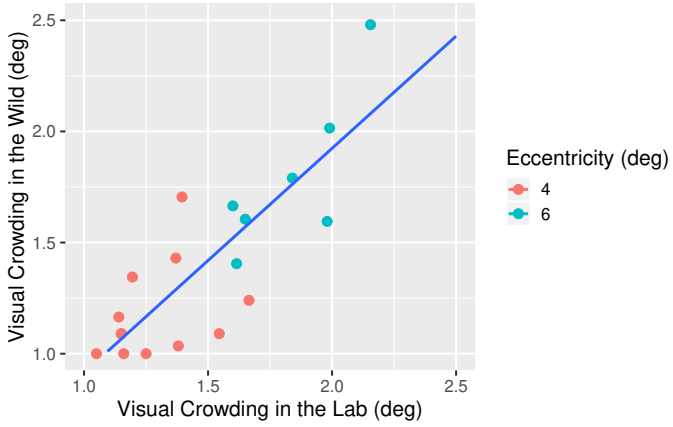


Figure 7.6: The visual crowding measures in Exp. 3 were significantly correlated (Pearson's $r = 0.86, p < 0.001, n = 18$) in the controlled and uncontrolled laboratory settings where 18 participants successfully completed the visual crowding experiment both in the lab with a physical chinrest and using the Virtual Chinrest on a laptop in their desired position and distance. Visual crowding effects increased as the eccentricity of the target increased (mean = 1.228° at 4° and mean = 1.811° at 6°, $t_{(9.08)} = -5.122, p < 0.001$ by Welch two sample t-test), confirming conventional eccentricity-dependent crowding effects.

crowding effect between the two conditions.

Results of 18 participants show that individuals' crowding effect measures are highly correlated in the controlled and uncontrolled laboratory setting (Pearson's $r = 0.86, p < 0.001, n = 18$), suggesting that individual differences can be precisely reproduced using the Virtual Chinrest when not controlling for viewing distance and angle (we removed the data of one participant who did not correctly follow the instruction). Fig. 7.6 presents each individual's crowding effect in the two conditions, grouped by eccentricity. The results are aligned with previous findings that the crowding effect is linearly dependent on eccentricity [28, 108, 129, 237]. A Welch two sample t-test showed that the average crowding effect is 1.228° when eccentricity is 4°, which is significantly different from the average crowding effect of 1.811° when eccentricity is 6° ($t_{(9.08)} = -5.122, p < 0.001$).

Exp. 4. Visual Crowding Experiment in the Wild

To evaluate whether we can replicate results from the visual crowding experiment in a truly uncontrolled environment, we conducted an online experiment on the volunteer-based experiment platform LabintheWild [186]. LabintheWild attracts participants from diverse demographic and geographic backgrounds [135,184,185,186]. Participants use a wide range of browsers, devices, and displays, and take experiments in a variety of situational lighting conditions and seating positions [184]. Our goal is to evaluate whether we can accurately replicate the visual crowding experiment despite this diversity.

Our experiment results, based on the data of 793 participants, replicate the previously found positive correlation between crowding effect and eccentricity [28,108,129,237]. More precisely, we compared the crowding effect between two target eccentricities. The results showed that participants' crowding effect increased as the eccentricity of the target increased from 4° (mean = 1.61°, sem = 0.02°) to 6° (mean = 2.66°, sem = 0.06°), and a non-parametric Mann-Whitney U test confirmed that the results are statistically significant ($W = 60502$, $p < .001$; Fig. 7.7a), confirming eccentricity-dependent crowding effects from previous studies [28,108,129,237].

Since we have a large and diverse sample, we further tested whether and how other covariates might be predictive of visual crowding: We ran a linear mixed-effects regression model with visual crowding as the dependent variable and participant as a random variable. We included *age* and *age_squared* (i.e., the square of the variable *age*) as fixed effects. Other fixed effects were *eccentricity* (4° or 6°), *dyslexia* (1 or 0) and *gender*. As shown in Table 7.2, the eccentricity-dependent crowding effects hold even when controlling for these other variables.

Our results show that people who self-reported having been diagnosed with dyslexia (N=59, excluding 10 who reported having additional impairments) have significantly higher visual crowding (e = 4°: 1.90°; e = 6°: 3.03°) than those without (e = 4°: 1.62°; e = 6°: 2.58°) in both target eccentricity conditions, consistent with the findings of prior work [108,147,216], although the relationship between dyslexia and visual crowding is highly debated [30, 58, 88, 142, 210] (Fig. 7.7a). In addition, we found that visual crowding is

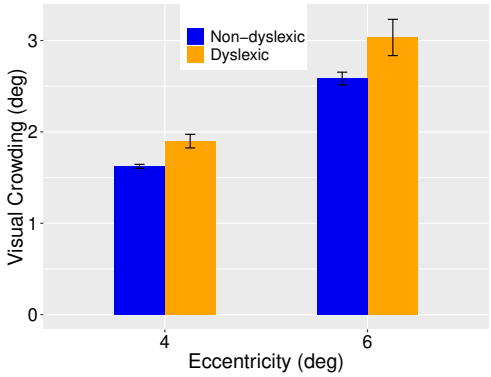
roughly half of the eccentricity: the ratio of the crowding to the eccentricity is 0.40 (4°) to 0.44 (6°), following the Bouma's law [29,172] also conformed by other studies [147,173]. Age also significantly impacted visual crowding, confirming previous findings demonstrating increased visual crowding in aging populations [150,169,205]. Moreover, we find increased crowding in young children compared to adults. Thus, there is a quadratic relationship between crowding and age, and individuals with dyslexia (on average) display increased crowding across the sampled age range.

Variable	**Est.**	**SE**	**t-value**	$Pr(>\|t\|)$
(Intercept)	-0.38	-0.20	-1.88	= .06 .
Eccentricity [6°]	0.53	0.02	21.39	< .001 ***
Dyslexia [yes]	0.26	0.10	2.70	< .005 **
Age_squared	0.004	0.001	2.68	< .005 **
Age	-0.02	0.01	-1.77	= .07 .
Gender	0.04	0.06	0.74	0.46 (n.s.)

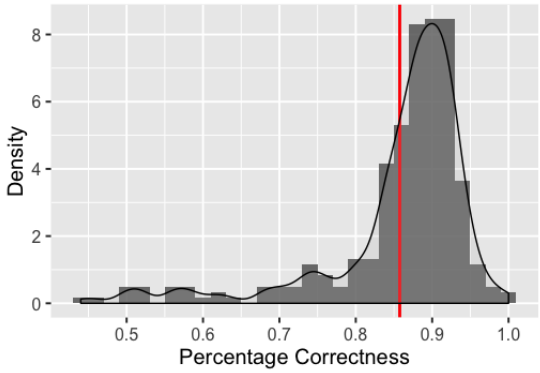
Table 7.2: The results of a quadratic mixed-effect model predicting visual crowding.

The average accuracy of the crowding experiment (50 trials) across all participants was 85.56% (median = 88%, max = 100%, min = 42%; Fig. 7.7b). This is aligned with the 79.4% correction rate of the 1-up 3-down staircase procedure, which demonstrates that we obtained reliable and accurate data the same as observed in our laboratory study or in prior work [131].

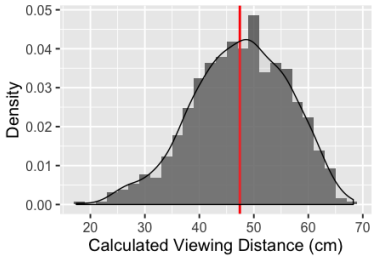
Our experiment included three blind spot tests at the beginning, in the middle, and at the end of the study to evaluate whether and how much participants move and whether it is sufficient to only assess their viewing distance once for a 10-minute online study. In our online experiment, participants varied in their viewing distance between 17.4 cm and 68.3 cm (mean = 47.4 cm, sd = 8.9 cm, see Fig. 7.7c). As shown in Fig. 7.7d, the average within-subjects standard deviation of estimated viewing distances (across the three blind spot tests) is 3.9 cm (min = 0.003 cm, max = 22.7 cm). Estimated by a one-way random effects model with absolute agreement, the intra-class correlation of a participant's estimated viewing



(a) The average visual crowding effects were significantly different between target eccentricity of 4° and 6°, and between participants with and without dyslexia. Error bars represent standard error.

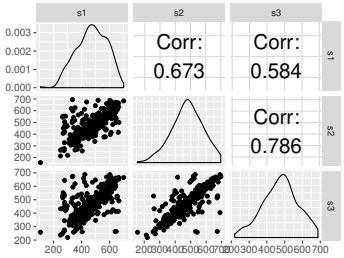


(b) The distribution of the percentage correctness of the crowding experiment across all participants. The average (indicated by the red vertical line) is 85.56%.



(c) The distribution of the viewing distances across all participants calculated by Virtual Chinrest. Our participants' viewing distances were between 17.4 cm and 68.3 cm with mean = 47.3 cm and sd = 8.9 cm.

(d) The distribution of the within-subjects standard deviation (SD) of the viewing distances across all participants: the average is 3.9 cm (min = 0.003 cm, max = 22.7 cm).



(e) The pairwise correlation of calculated viewing distances among three blind spot tasks at the beginning (s1), in the middle (s2) and at the end (s3) of the crowding experiment.

Figure 7.7: The results of Exp. 4 where 793 participants completed the visual crowding experiment implemented using Virtual Chinrest on LabintheWild.

distances before, during and after the crowding experiment is $\rho = 0.88$ (see Fig. 7.7e for pairwise correlations). This suggests that different participants vary substantially in their viewing distance (underlining the need for a Virtual Chinrest), but participants do not move much over the course of a 10-minute online experiment. We found no substantial difference in visual crowding between people who moved more and less, and therefore, assessing the viewing distance once at the beginning of an experiment may be sufficient for most participants.

7.5 Discussion

This paper introduced the Virtual Chinrest, a novel method that allows estimating participants' viewing distances, and thus, calibrating the size and location of stimuli in online experiments. We validated our method in two laboratory studies in which we varied the viewing distance and display size, showing that the Virtual Chinrest estimates participants' viewing distances with an average absolute error of 3.25 cm – a negligible error given an average viewing distance of 53 cm. Using the Virtual Chinrest in an online environment with 1153 participants, we were able to replicate and extend the results of a laboratory study on visual crowding, which requires particularly tight control of viewing distance and angle. More specifically, we replicated three prior findings: (1) the positive correlation between the crowding effect and eccentricity in [28, 108, 129, 237], (2) the finding that participants with dyslexia experience higher visual crowding than those without dyslexia [108, 147, 216], and (3) the increase in visual crowding that occurs with aging [150, 169, 205]. Moreover, we extended these results by showing that there is a quadratic relationship between age and visual crowding. Our findings pave the way for laboratory studies requiring a physical chinrest to be conducted online, enabling psychophysical studies with larger and more diverse participant samples than previously possible.

The Virtual Chinrest is not necessary for all types of psychophysical online experiments. For example, prior work has successfully replicated visual perception experiments on proportional judgments of spatial encodings and luminance contrast, and investigated the effects of chart size and gridline spacing for optimized parameters for web-based display via online experiments, without controlling for viewing distance, display size or resolution [89, 140].

These prior experiments followed a within-subjects design, did not require cross-device comparisons, and the results are not sensitive to changes in visual degrees.

However, there are two main types of visual perception studies that are unlikely to replicate if conducted without controlling for participants' viewing distance: (1) between-subjects studies that compare specific metrics across participants, because they require a consistent measurement across various environments and devices, and (2) any study that requires visual stimuli sizes to be the same across participants. Visual crowding is a prime example, where thresholds are expressed in units of visual angle. For these types of study designs, each participant's viewing distance derived from the Virtual Chinrest can be used for adapting the size of the stimuli and/or as a control variable in the analysis.

Our results show that after instructing online participants to keep their position throughout the experiment, they indeed move very little – on average, participants viewing distance changed by 3.9 cm. This suggests that for 10-minute experiments, asking participants to take the 30 seconds to set up the Virtual Chinrest once at the beginning of the experiment may be enough. For longer experiments, we suggest assigning the Virtual Chinrest tasks multiple times throughout the experiments to adjust the stimuli correspondingly.

In summary, we developed the Virtual Chinrest to measure a person's viewing distance through the web browser, enabling large-scale psychophysical experiments of visual perception to be conducted online. Our method makes it possible to automatically adjust the stimulus configurations according to each participant's viewing distance, producing reliable results for visual perception studies that are sensitive to display parameters. We hope that our method will enable researchers to leverage the power of studies with large and diverse online samples, which often have greater external validity, can detect smaller effects, and have a higher probability of finding similarities and differences between populations than traditional laboratory studies.

Chapter 8

DISCUSSION

In this dissertation, I have demonstrated the viability of conducting volunteer-based online experiments with people with cognitive disabilities through a series of quantitative and qualitative studies. By replicating and extending four experiments on LabintheWild, I show that we are able to attract sufficiently large numbers of participants with cognitive disabilities to robustly replicate and extend prior laboratory study results on dyslexia, autism, age-related memory decline, and age-related motor impairments. By conducting semi-structured interviews with test participants, I have learned that online tests act as an important resource that addresses the shortcomings in support systems for people with professionally diagnosed or suspected cognitive disabilities before and after receiving a professional diagnosis. In particular, online tests can lower barriers to a professional diagnosis, provide valuable information about the nuances of a disability, and support people in forming a disability identity – an invaluable step toward a positive acceptance of oneself. An interview study with six healthcare professionals further confirms that well-designed online tests can be a good starting point for people to explore their cognitive conditions, but they could definitely not replace a professional diagnosis because of various challenges. Finally, by designing and conducting two studies on LabintheWild, I demonstrate novel techniques for controlling aspects, such as a participant's viewing distance, in online experiments. I also demonstrate using online experiments as a method to generate novel knowledge and inform design guidelines of technologies for cognitive disabilities, in particular, dyslexia.

Following this section, I discuss how this work is connected to existing research (to Disability Studies in particular), when it is suitable to conduct volunteer-based online experiments, and what should such studies look like in the future.

8.1 Where Are Volunteer-based Online Studies Situated in Disability Studies?

8.1.1 Forming Disability Identity

In Chapter 4, I describe that participants with (suspected) cognitive disabilities use online tests predominantly before a professional diagnosis to validate their own suspicions and justify the need for a professional diagnosis, because they often face barriers when seeking professional help due to factors, such as cost, access issues and resistance within their families. As such, these often relatively informal and anonymous tests play a unique role in the support systems of people with disabilities: a way of slowly and informally introducing people to their disability without the potential risks perceived by an official, inescapable professional diagnosis.

This underlines the important role of online tests in forming a disability identity, which includes an acceptance of one's disability, developing a positive view of oneself, and feeling connected to others with similar experiences [60]. Establishing a disability identity has been shown to support individuals in coming to terms with their disability, and to lower stress levels and the risk of mental health effects [109]. In Chapter 4, our interviewees report that online tests slowly help them accept their disability, while also providing a reason for connecting with others. For example, participants frequently post their results of online tests in online communities (as shown in [135] and [137]), facilitating *communal attachment* [60]. This allows for a valuable additional pathway towards forming a disability identity.

8.1.2 Emancipatory Research vs. Volunteer Studies

As demonstrated by prior work, in contrast to financial compensation, our volunteer participants receive personalized feedback which (1) enables self-reflection and social comparison [94], (2) exposes participants to scientific concepts and increases their interest in research and scientific findings [167], (3) ensures data quality as participants are intrinsically motivated to provide honest answers and exert themselves [77, 80, 186], and (4) serves as a word-of-mouth recruitment tool because participants share their results on social networking sites or other web pages [186]. Therefore, an important question is if people with disabilities

in particular benefit from the personalized feedback and if this reward is proportional to their levels of participation.

This is relevant to the discussion of emancipatory research, a research perspective of producing knowledge that can be of benefit to disadvantaged people – one which recognizes the power imbalance in traditional research processes and aims to empower respondents through research. In particular, there has been increasing interest in the role of research in relation to the empowerment and thus the inclusion of disabled people [18]. I reflect on how our findings on the design and deployment of online experiments with people with disabilities should be aligned with the six core principles of the emancipatory research paradigm shown in Table 8.1, particularly #1, #3, #4, #6.

First, researchers should follow the social model of disability, which views the origins of disability as the mental attitudes and physical structures of society, rather than the medical model of disability which frames an individual's impairment as the cause for their inability to participate fully in society. This suggests that the online tests should avoid any claims about how one's "impairment" could be the potential cause of underperformance on the tests. Instead, test developers should help people interpret the test results by taking away the burden of disability on the individual, but rather, reflecting on what is 'wrong' with the situation, context, or environment. Next, Principle #3 suggests that any research production should provide disabled participants with self-empowerment and the removal of disabling social and physical barriers. Moreover, the nature of the engagement should be determined by disabled people [219]. Findings from this dissertation have demonstrated that the personalized results are providing knowledge and support for people with cognitive disabilities at different stages of their lives. As both the participants and the doctors suggest, additional resources should be included, facilitating the removal of barriers for people to receive proper professional help as the next step after taking the online tests. Both Principle #4 and Principle #6 suggest that people with disabilities should involve in the research production, rather than solely being the "participants", fulfilling the changing needs of these populations. Our interview study with the participants with cognitive disabilities in Chapter 4 is taking the first step to involving them in research by probing their opinions about the online tests and creating design implications for designing future tests. In the

Table 8.1: Six core principles of the emancipatory research paradigm [219]

#	Core Principle
1	The adoption of a social model of disablement as the epistemological basis for research production
2	The surrender of claims to objectivity through overt political commitment to the struggles of disabled people for self-emancipation
3	The willingness only to undertake research where it will be of practical benefit to the self-empowerment of disabled people and/or the removal of disabling barriers
4	The evolution of control over research production to ensure full accountability to disabled people and their organizations
5	Giving voice to the personal as political whilst endeavouring to collectivize the political commonality of individual experiences
6	The willingness to adopt a plurality of methods for data collection and analysis in response to the changing needs of disabled people

future, researchers should provide additional opportunities for participants to involve in the development of such online tests, for instance, by conducting participatory design sessions with participants with disabilities. This way, it also paves the way for researchers to make themselves "available" to disabled people [242].

8.1.3 Online Tests Should Align With the Affirmative Model of Disability

As one of the design implications, I briefly described in Chapter 4 that online tests should align with the affirmative model of disability [221] by highlighting a test participant's strengths and providing additional resources that describe positive examples. The affirmation model finds its root in the social model of disability but the focus here is on self-acceptance rather than on policies and external environments. It aims to classify disability as non-tragic, instead focusing on positive experiences and both individual and collective

identities that disabilities can provide [221].

When it comes to the application of this model of disability to online tests, it is important to note that CDS and the identity model do not change the diagnosis itself, but can instead help a person to recognize their disability and accept it as part of their identity. Building upon the CDS notion that language is inherently political and resists objectivity through subtle ideological expressions [92], tests could be more conscious of the language they use and avoid a negative tone. While it is natural for concepts, such as disability, that society views in a negative light to be associated with limited and negative language, it is important to move away from this negativity as it promotes a tragedy-based view of disability [92, 221]. Tests should instead encourage participants to think of mental or cognitive disabilities through an affirmative lens: to imagine strengths they have as a result of their disability, to not see abnormality or interdependence as negative, and to view disability-related limitations not as innate to them, but a part of an ableist society [221]. This can be accomplished through positive, thought-provoking language as well as additional resources for the affirmative understanding of disability and political action.

8.2 Would Online Tests Be Suitable For Studying Other Types of Disabilities?

In this dissertation, I focus my research on examining online studies that target cognitive disabilities. Thus, an important question is whether volunteer-based online tests are also suitable for studying other types of disabilities. Prior work has demonstrated that it is possible to conduct *large-scale* (paid or volunteer-based) online tests to study *a variety* of disabilities. For instance, Findlater and Zhang conducted an online study [1] that captures input performance with a mouse and/or touchscreen from over 700 participants – the analysis has demonstrated the continuous relationship between age and input performance for people with and without motor impairment [67]. Another example is that Reinecke et al. developed an online color differentiation test on LabintheWild that collected data from around 30,000 participants, analyzing people's varying color differentiation abilities linked to a combination of situational lighting conditions, age, gender, and Color Vision

[1]Participants are recruited from Amazon Mechanical Turk and Cint

Deficiency [184]. Similarly, Bragg et al. conducted a LabintheWild study with 453 participants (including 143 who are visually impaired) that demonstrates the various listening rates that people with different demographics and abilities could achieve [32]. Last but not least, in [73] where 95 ataxia, 46 parkinsonism, and 29 control participants and 230k online participants completed a rapid web-based computer mouse test, Gajos et al. were able to develop models that measure ataxia progression with high sensitivity and distinguish ataxia or parkinsonism from healthy controls with high sensitivity and specificity.

However, the constraints of web browsers might limit what studies can be conducted online. For example, it would be very difficult to conduct the studies online if it requires external devices, such as eye-tracking equipment and chin rests. Therefore, more work will need to be done to establish validated methodologies that are suitable for the online environment, similar to the Virtual Chinrest, to enable conducting additional types of studies reliably online. In addition, one immediate first step we can take is to make all the online tests themselves accessible, for example, for screen reader users, people who have limited motor abilities, and those with cognitive impairment.

8.3 Is Volunteer-based Online Experiment Always the Way to Go?

8.3.1 From Participants' Point of View

In both Chapter 3 and Chapter 4, I show that people with (suspected) cognitive disabilities benefit from taking online tests to better understand how their (suspected) cognitive disabilities affect their lives and to receive support from groups of people with similar experiences. However, it needs to be emphasized that online tests should not be seen as superior to, or a replacement for, a professional diagnosis. In fact, prior work showed that web search of medical information, especially when it is employed as a diagnostic procedure, increases the anxieties of people who have little or no medical training [236]. Instead, I hope to showcase that, with all the barriers to receiving professional healthcare and the stigma associated with being labeled as having cognitive or mental disabilities, getting a formal diagnosis is not always possible and desirable; in those situations, taking online tests provides great benefits and can be a first step for people to better understand themselves and prepare

them to seek support from professionals.

8.3.2 From Researchers' Point of View

The primary goal of researchers is to collect reliable data from participants with various backgrounds. While a larger number and more diverse participants could be reached via online studies, some questions should be asked by the researchers before diving into developing the online tests. For example, have the studies been validated in an unsupervised, uncontrolled environment? How long do you have for the recruitment; will using alternative recruitment methods, such as Prolific or Amazon Mechanical Turk, warrant faster responses in a shorter period of time? And finally, can we provide accurate and appropriate personalized feedback in return for participation? These questions might serve as a useful starting point to help researchers assess whether volunteer-based online studies are the best course of action. It is also shown in this dissertation that a variety of studies can attract large numbers of volunteer participants and can be completed rigorously online, including experiments that collect subjective rating scales on various tasks, and objective measurements such as reaction time, implicit and explicit memory during the learning process, reading speed, and pointing performance using a mouse.

8.3.3 From Healthcare Professionals' Point of View

In Chapter 5, I find that healthcare providers make a diagnosis significantly depending on their experience seeing and treating other people and it is a highly adaptive process based on the individual's demographics and developmental background. The doctors also mentioned that there does not exist one test that is suitable for everyone: there are many aspects of the same condition – some symptoms of the same condition might be easy to see in one person, and virtually invisible in another. Therefore, if well-developed and validated, online tests could be a useful starting point for people to start exploring themselves, but again they should never be the replacement of an official diagnosis by healthcare professionals.

8.4 What Should the Online Studies Look Like for People With (Cognitive) Disabilities in the Future?

Here, I synthesize the lessons learned and summarize the four most significant design implications derived from this dissertation. I also discuss how I envision the way that online studies will fit in the overall support systems for people with disabilities in the future.

8.4.1 Design Implications

Develop and Provide Validated Tests. Because of various barriers to seek for professional help, our participants expressed a desire to use online experiments to diagnose themselves and interpret the results related to their (suspected) disabilities. To address participants' need for diagnosing themselves and for testing the severity of their disability, it is essential to develop and provide validated tests. This can be done in two ways. First, adopt existing studies or instruments that have proven to be reliable for assessing specific disabilities. This is also what psychiatrists recommend looking for to verify the validity of a test. Second, develop new studies to address participants' need for specific tests. This is more difficult and time-consuming, and requires the involvement of domain experts in those disabilities as well as rigorous validations before publishing the tests.

Provide participants with nuances of their conditions and with comparison to a specific group of people. Our interviews with participants revealed a desire to receive personalized feedback that allows specific comparison to others with a similar disability. A similar calibration based on other patients with similar conditions is also commonly used by psychiatrists when making a clinical diagnosis. Therefore, providing participants with a choice of a comparison group will not only fulfill participants' needs of comparing with similar people, motivating people to recruit others with similar disabilities through their social networks, but also help people interpret the results more accurately. Achieving this requires bootstrapping the data with sufficient results from a specific group, which may or may not be available from prior literature. One solution would be an integrated model, in which participants who want to compare themselves to a specific group can recruit others, and results are then communicated back to anyone who signed up to receive specific results

about this group with a time delay. Another possible solution is to bring in the domain expertise and experience of the psychiatrists and integrate it into the personalized results. Of course, balancing the trade-off between providing *timely* feedback based only on the participant's performance and using other patients' results while preserving their *privacy* would be important for future research to be further investigated.

Support participants in sharing and discussing results. Online experimentation platforms, such as LabintheWild, should support participants in sharing and discussing results with others. This includes supporting participants to discuss results with other participants, either through some internal forums within the platform, but a potentially better idea is to direct participants to existing external forums where the discussion of the same tests is already going on. This also includes solving people's needs of sharing and discussing results with researchers who can provide answers, support, and debriefing information to people who may need it. For people with disabilities, some questions might have to be answered by an expert with specific expertise, who may or may not be the researchers themselves. So how to train and motivate such expert groups to provide this support will be exciting new research in crowdsourcing and citizen science.

The personalized results should align with the affirmative model of disability and provide additional resources that help participants move forward. Last but not least, as I already discussed in Chapter 8.1.1, online tests sometimes serve a role in helping people form a disability identity and develop a positive view of themselves. Thus, the test results should focus on providing positive interpretations of people's performance and focusing on their strengths rather than personal tragedies. In addition, both the participants and the psychiatrists emphasize the importance of including additional resources that guide people on what the next step is. To address this need, the result page could include pointers to resources such as how to find an adequate healthcare professional for a formal diagnosis. Partnering with hotlines, hospitals, and other services available may also be a way of providing online test participants with immediate, in-person support if needed. This integration of online tests into the healthcare systems, however, is beyond the scope of this dissertation and requires significant future research to examine its viability. I will briefly describe what I envision the future of online tests could achieve in the next section.

8.4.2 The Future of Online Tests as Part of the Support Systems

This dissertation is motivated by the fact that as researchers, we often struggle to find large numbers of participants with diverse backgrounds and abilities for studying certain disabilities. As I examine the viability of using volunteer-based online studies for fulfilling this goal, I discover that participants use the online studies in a way that is far beyond their intended purpose: people use online tests to make self-diagnosis and sometimes forgo an official diagnosis because of them. Although it is certainly discouraged to use online tests like this right now, incorporating online tests into the official healthcare systems is not impossible in the future. Note that this is more pertinent to tests that are explicitly advertised to test a disability in one way or another, rather than tests that people interpret to be about certain disabilities but that are actually about testing people's abilities or behaviors in general. I imagine that these volunteer-based online test platforms could be modified and integrated into the healthcare systems as pre-screening tools and an initial channel that supports the bi-directional communication between the participants and the healthcare professionals. Together with the existing instruments, online tests could be part of the diagnostic procedures in the future. With solid clinical validations and appropriate regulations, online tests or ones of similar formats could even serve as diagnostic tools by themselves, making self-diagnosis viable and significantly lowering the barriers to obtaining a diagnosis.

Chapter 9

CONCLUSION

In this dissertation, I completed a series of work to demonstrate the viability of volunteer-based online experiments in studying people with disabilities, and in particular, cognitive disabilities, at scale. By replicating and extending four experiments on LabintheWild, I showed that sufficiently large numbers of participants with various disabilities are self-motivated to take part in these studies that robustly replicate and extend prior laboratory study results. By conducting interviews, I then identified how volunteer-based online studies related to cognitive disabilities are perceived by different stakeholders, i.e. test participants and healthcare professionals. Finally, by designing and conducting two studies on LabintheWild, I demonstrated the feasibility of using online experiments as a method to generate novel knowledge and inform design guidelines of technologies for cognitive disabilities, in particular, dyslexia.

Together, the work demonstrates my thesis statement:

Volunteer-based online studies can be conducted with people with cognitive disabilities in a way that is large-scale, self-motivating, helpful for participants, and enables rigorous experiments.

Summary of Contributions

This dissertation has made interdisciplinary contributions across multiple research areas: in addition to extending research in accessibility and crowdsourcing within HCI, it also makes empirical contributions to dyslexia research and vision science, as well as artifact contributions to the broader research community for conducting online experiments. The specific contributions of this dissertation include:

Empirical contributions to **HCI**, **accessibility** and **crowdsourcing**:

- Four experiments with a total of 356k participants on LabintheWild that not only

replicated but also extended prior laboratory studies about dyslexia, autism, age-related memory decline, and age-related motor impairment [135] (Chapter 3).

- Two interview studies with 16 test participants and 6 healthcare professionals, showing how volunteer-based online experiments are currently perceived and how they can be improved in the future from different perspectives [137] (Chapter 4 and 5).

- An LabintheWild study with 391 participants with and without dyslexia, showing how Reader View (a feature provided by many web browsers) improves people's digital reading experience and how it can be improved to better support those with varying reading skills [138] (Chapter 6).

- Three in-laboratory studies and one large-scale LabintheWild study with 1153 participants that validated the Virtual Chinrest [136] (Chapter 7).

Empirical contributions to **dyslexia research** and **vision science**:

- Confirming previous findings about dyslexia in two LabintheWild studies with significantly larger numbers of participants [136, 138] (Chapter 6 and 7).

- Replicating prior in-laboratory results on visual crowding with more than 1k participants in an online experiment, demonstrating visual crowding increases with eccentricity, dyslexia, and age, respectively; and extending these results by showing that there is a quadratic relationship between age and visual crowding [136] (Chapter 7).

Artifact contributions to the broader **interdisciplinary research communities**:

- A novel method, the Virtual Chinrest, that measures a participant's viewing distance in the web browser which makes it possible to automatically adjust stimulus configurations in online experiments based on viewing distances [136] (Chapter 7).

- An open-sourced JavaScript library and a jsPsych plugin for anyone to integrate the Virtual Chinrest in their online experiments [136] (Chapter 7).

BIBLIOGRAPHY

[1] Accessible writing guide. http://www.sigaccess.org/welcome-to-sigaccess/resources/accessible-writing-guide/. Accessed: 2020-08-13.

[2] National institute of mental health. https://www.nimh.nih.gov/health/statistics/index.shtml. Accessed: 2020-05-20.

[3] Pros and cons of diagnosis. https://reneehoekstra.com/pros-and-cons-of-diagnosis/. Accessed: 2020-08-13.

[4] Aaiz Ahmed and S Samuel. Self-diagnosis in psychology students. *The International Journal of Indian Psychology*, 5(1):148–164, 2017.

[5] Renato D Alarcón. Culture, cultural factors and psychiatric diagnosis: review and projections. *World psychiatry*, 8(3):131, 2009.

[6] Apple. Working with safari reader, 2018. Accessed 15-September-2018.

[7] American Psychiatric Association et al. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.

[8] International Dyslexia Association. Dyslexia basics, 2017. Accessed 31-August-2018.

[9] International Dyslexia Association et al. Definition of dyslexia. *Retrieved from dyslexiaida. org*, 2002.

[10] Fakhrul Anuar Aziz and Husniza Husni. Interaction design for dyslexic children reading application: a guideline. 2012.

[11] Sairam Balani and Munmun De Choudhury. Detecting and characterizing mental health related self-disclosure in social media. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1373–1378, 2015.

[12] Priscilla Balestrucci, Katrin Angerbauer, Cristina Morariu, Robin Welsch, Lewis L Chuang, Daniel Weiskopf, Marc O Ernst, and Michael Sedlmair. Pipelines bent, pipelines broken: Interdisciplinary self-reflection on the impact of covid-19 on current and future research (position paper). In *2020 IEEE Workshop on Evaluation and Beyond-Methodological Approaches to Visualization (BELIV)*, pages 11–18. IEEE, 2020.

[13] Russell A Barkley and Kevin R Murphy. Identifying new symptoms for diagnosing adhd in adulthood. *The ADHD Report*, 14(4):7–11, 2006.

[14] Lisa J Barney, Kathleen M Griffiths, and Michelle A Banfield. Explicit and implicit information needs of people with depression: a qualitative investigation of problems reported on an online depression support forum. *BMC psychiatry*, 11(1):88, 2011.

[15] Jonathan S Barnhoorn, Erwin Haasnoot, Bruno R Bocanegra, and Henk van Steenbergen. Qrtengine: An easy solution for running online reaction time experiments using qualtrics. *Behavior research methods*, 47(4):918–929, 2015.

[16] Simon Baron-Cohen, Sally Wheelwright, Jacqueline Hill, Yogini Raste, and Ian Plumb. The "reading the mind in the eyes" test revised version: a study with normal adults, and adults with asperger syndrome or high-functioning autism. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, 42(2):241–251, 2001.

[17] Marialena Barouti, Konstantinos Papadopoulos, and Georgios Kouroupetroglou. Synthetic and natural speech intelligibility in individuals with visual impairments: Effects of experience and presentation rate. In *European AAATE Conference, Portugal*, pages 695–699, 2013.

[18] Len Barton. Emancipatory research and disabled people: Some observations and questions. *Educational review*, 57(3):317–327, 2005.

[19] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

[20] Heather Becker, Greg Roberts, Janet Morrison, and Julie Silver. Recruiting people with disabilities as research participants: Challenges and strategies to address them. *Mental Retardation*, 42(6):471–475, 2004.

[21] Gillinder Bedi, Facundo Carrillo, Guillermo A Cecchi, Diego Fernández Slezak, Mariano Sigman, Natália B Mota, Sidarta Ribeiro, Daniel C Javitt, Mauro Copelli, and Cheryl M Corcoran. Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia*, 1:15030, 2015.

[22] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[23] Adam J Berinsky, Gregory A Huber, and Gabriel S Lenz. Using Mechanical Turk as a subject recruitment tool for experimental research. *Political Analysis*, 20:351–68, 2012.

[24] Eta S Berner. *Clinical decision support systems*, volume 233. Springer, 2007.

[25] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342. ACM, 2010.

[26] J. Birch. *Diagnosis of Defective Colour Vision*. Oxford: Butterworth-Heinemann, 2001.

[27] Richard Blumenthal* and Jean Endicott. Barriers to seeking treatment for major depression. *Depression and anxiety*, 4(6):273–278, 1996.

[28] Herman Bouma. Interaction effects in parafoveal letter recognition. *Nature*, 226(5241):177, 1970.

[29] Herman Bouma. Visual interference in the parafoveal recognition of initial and final letters of words. *Vision research*, 13(4):767–782, 1973.

[30] Herman Bouma and Ch P Legein. Foveal and parafoveal recognition of letters and words by dyslexics and by average readers. *Neuropsychologia*, 15(1):69–80, 1977.

[31] Patricia Greig Bowers and Maryanne Wolf. Theoretical links among naming speed, precise timing mechanisms and orthographic skill in dyslexia. *Reading and Writing*, 5(1):69–85, 1993.

[32] Danielle Bragg, Cynthia Bennett, Katharina Reinecke, and Richard Ladner. A Large Inclusive Study of Human Listening Rates. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '18. ACM, 2018.

[33] Robin Brewer, Meredith Ringel Morris, and Anne Marie Piper. "why would anybody do this?": Understanding older adults' motivations and challenges in crowd work. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '16, pages 2246–2257, New York, NY, USA, 2016. ACM.

[34] Renee Brimstone, Jill E Thistlethwaite, and Frances Quirk. Behaviour of medical students in seeking mental and physical health care: exploration and comparison with psychology students. *Medical education*, 41(1):74–83, 2007.

[35] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1):3–5, 2011.

[36] Brian Butterworth, Sashank Varma, and Diana Laurillard. Dyscalculia: from brain to education. *science*, 332(6033):1049–1053, 2011.

[37] Rocío Calvo, Shaun K. Kane, and Amy Hurst. Evaluating the accessibility of crowdsourcing tasks on amazon's mechanical turk. In *Proceedings of the 16th international ACM SIGACCESS conference on Computers and Accessibility*, ASSETS '14, pages 257–258, New York, NY, USA, 2014. ACM.

[38] Alastair G Cardno, Frühling V Rijsdijk, Pak C Sham, Robin M Murray, and Peter McGuffin. A twin study of genetic relationships between psychotic symptoms. *American Journal of Psychiatry*, 159(4):539–545, 2002.

[39] Alaina Carr. An exploration of mechanical turk as a feasible recruitment platform for cancer survivors. *Undergraduate Honors Theses*, 59, 2014.

[40] Kathy Charmaz. *Constructing grounded theory: A practical guide through qualitative analysis*. sage, 2006.

[41] Eric H. Chudler. Sight (vision).

[42] World Wide Web Consortium. Web content accessibility guidelines (wcag) 2.0, 2008. Accessed 10-September-2018.

[43] World Wide Web Consortium. Css values and units module level 4, 2018. Accessed 10-September-2018.

[44] Juliet Corbin and Anselm Strauss. *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage publications, 2014.

[45] Anthony Creed, Ian Dennis, and Stephen Newstead. Proof-reading on vdus. *Behaviour & Information Technology*, 6(1):3–13, 1987.

[46] Lee J Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334, 1951.

[47] Matthew JC Crump, John V McDonnell, and Todd M Gureckis. Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PloS one*, 8(3):e57410, 2013.

[48] Mary Czerwinski, Eric Horvitz, and Edward Cutrell. Subjective duration assessment: An implicit probe for software usability. In *Proceedings of IHM-HCI 2001 conference*, volume 2, pages 167–170, 2001.

[49] Meredyth Daneman and Patricia A Carpenter. Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, 19(4):450–466, 1980.

[50] Ronald D Davis and Eldon M Braun. *The gift of dyslexia: why some of the brightest people can't read and how they can learn*. Souvenir Press, 2011.

[51] Munmun De Choudhury, Sanket S. Sharma, Tomaz Logar, Wouter Eekhout, and René Clausen Nielsen. Gender and cross-cultural differences in social media disclosures of mental illness. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, page 353–369, New York, NY, USA, 2017. Association for Computing Machinery.

[52] Joshua R de Leeuw and Benjamin A Motz. Psychophysics in a web browser? comparing response times collected with javascript and psychophysics toolbox in a visual search task. *Behavior Research Methods*, 48(1):1–12, 2016.

[53] Marianne Dee and Vicki L. Hanson. A large user pool for accessibility research with representative users. In *Proceedings of the 16th international ACM SIGACCESS conference on Computers and Accessibility*, ASSETS '14, pages 35–42, New York, NY, USA, 2014. ACM.

[54] Marianne Dee and Vicki L. Hanson. A pool of representative users for accessibility research: Seeing through the eyes of the users. *ACM Trans. Access. Comput.*, 8(1):4:1–4:31, January 2016.

[55] Andrew Dillon. Reading from paper versus screens: A critical review of the empirical literature. *Ergonomics*, 35(10):1297–1326, 1992.

[56] Linda Dimeff and Marsha M Linehan. Dialectical behavior therapy in a nutshell. *The California Psychologist*, 34(3):10–13, 2001.

[57] Xianghua Ding, Patrick C. Shih, and Ning Gu. Socially embedded work: A study of wheelchair users performing online crowd work in china. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 642–654, New York, NY, USA, 2017. ACM.

[58] Adi Doron, Mauro Manassi, Michael H Herzog, and Merav Ahissar. Intact crowding and temporal masking in dyslexia. *Journal of Vision*, 15(14):13–13, 2015.

[59] Bradley Duchaine, Laura Germine, and Ken Nakayama. Family resemblance: Ten family members with prosopagnosia and within-class object agnosia. *Cognitive neuropsychology*, 24(4):419–430, 2007.

[60] Dana S Dunn and Shane Burcaw. Disability identity: Exploring narrative accounts of disability. *Rehabilitation Psychology*, 58(2):148, 2013.

[61] Mary Dyson and Mark Haselgrove. The effects of reading speed and reading patterns on the understanding of text read from screen. *Journal of Research in Reading*, 23(2):210–223, 2000.

[62] Mary C Dyson. How physical text layout affects reading from screen. *Behaviour & Information Technology*, 23(6):377–393, 2004.

[63] Mary C Dyson and Gary J Kipping. The legibility of screen formats: are three columns better than one? *Computers & Graphics*, 21(6):703–712, 1997.

[64] Benedikt V Ehinger, Katja Häusser, Jose P Ossandon, and Peter König. Humans treat unreliable filled-in percepts as more real than veridical ones. *Elife*, 6:e21761, 2017.

[65] Nermin Eissa, Mohammed Al-Houqani, Adel Sadeq, Shreesh K Ojha, Astrid Sasse, and Bassem Sadek. Current enlightenment about etiology and pharmacological treatment of autism spectrum disorder. *Frontiers in neuroscience*, 12:304, 2018.

[66] Horacio Fabrega Jr. Culture and history in psychiatric diagnosis and practice. *Psychiatric Clinics of North America*, 24(3):391–405, 2001.

[67] Leah Findlater and Lotus Zhang. Input accessibility: A large dataset and summary analysis of age, motor ability and input performance. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '20, New York, NY, USA, 2020. Association for Computing Machinery.

[68] Adam Fourney, Meredith Ringel Morris, Abdullah Ali, and Laura Vonessen. Assessing the readability of web search results for searchers with dyslexia. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '18, pages 1069–1072, New York, NY, USA, 2018. ACM.

[69] Jeanne C Fox, Michael Blank, Virginia G Rovnyak, and Rhoneise Y Barnett. Barriers to help seeking for mental disorders in a rural impoverished population. *Community mental health journal*, 37(5):421–436, 2001.

[70] Nicholas J Fox, Katie J Ward, and Alan J O'Rourke. The 'expert patient': empowerment or medical dominance? the case of weight loss, pharmaceutical drugs and the internet. *Social science & medicine*, 60(6):1299–1309, 2005.

[71] Allen Frances. A report card on the utility of psychiatric diagnosis. *World Psychiatry*, 15(1):32, 2016.

[72] Yafit Gabay, Eli Vakil, Rachel Schiff, and Lori L Holt. Probabilistic category learning in developmental dyslexia: Evidence from feedback and paired-associate weather prediction tasks. *Neuropsychology*, 29(6):844, 2015.

[73] Krzysztof Z Gajos, Katharina Reinecke, Mary Donovan, Christopher D Stephen, Albert Y Hung, Jeremy D Schmahmann, and Anoopum S Gupta. Computer mouse use captures ataxia and parkinsonism, enabling accurate measurement and detection. *Movement Disorders*, 35(2):354–358, 2020.

[74] Katharina Galuschka, Ruth Görgen, Julia Kalmar, Stefan Haberstroh, Xenia Schmalz, and Gerd Schulte-Körne. Effectiveness of spelling interventions for learners with dyslexia: A meta-analysis and systematic review. *Educational Psychologist*, 55(1):1–20, 2020.

[75] Paul J Gerber and Dale S Brown. *Learning Disabilities and Employment.* ERIC, 1997.

[76] Laura Germine, Nathan Cashdollar, Emrah Düzel, and Bradley Duchaine. A new selective developmental deficit: Impaired object recognition with normal face recognition. *Cortex*, 47(5):598–607, 2011.

[77] Laura Germine, Ken Nakayama, Bradley C Duchaine, Christopher F Chabris, Garga Chatterjee, and Jeremy B Wilmer. Is the web as good as the lab? comparable performance from web and lab in cognitive/perceptual experiments. *Psychonomic bulletin & review*, 19(5):847–857, 2012.

[78] David C Giles and Julie Newbold. Self-and other-diagnosis in user-led mental health online communities. *Qualitative Health Research*, 21(3):419–428, 2011.

[79] Barney G Glaser and Anselm L Strauss. *The Discovery of Grounded Theory: Strategies for Qualitative Research.* Transaction Publishers, 2009.

[80] Samuel D Gosling, Simine Vazire, Sanjay Srivastava, and Oliver P John. Should we trust web-based studies? a comparative analysis of six preconceptions about internet questionnaires. *American psychologist*, 59(2):93, 2004.

[81] Peter Gregor, Anna Dickinson, Alison Macaffer, and Peter Andreasen. Seeword – a personal word processing environment for dyslexic computer users. *British Journal of Educational Technology*, 34(3):341–355, 2003.

[82] Amelia Gulliver, Kathleen M Griffiths, and Helen Christensen. Perceived barriers and facilitators to mental health help-seeking in young people: a systematic review. *BMC psychiatry*, 10(1):113, 2010.

[83] Pegah Hafiz, Kamilla Woznica Miskowiak, Lars Vedel Kessing, Andreas Elleby Jespersen, Kia Obenhausen, Lorant Gulyas, Katarzyna Żukowska, and Jakob Eyvind Bardram. The internet-based cognitive assessment tool: System design and feasibility study. *JMIR formative research*, 3(3):e13898, 2019.

[84] Ella Hancock-Johnson, Chris Griffiths, and Marco Picchioni. A focused systematic review of pharmacological treatment for borderline personality disorder. *CNS drugs*, 31(5):345–356, 2017.

[85] Joshua K Hartshorne and Laura T Germine. When does cognitive functioning peak? the asynchronous rise and fall of different cognitive abilities across the life span. *Psychological science*, 26(4):433–443, 2015.

[86] Joshua K Hartshorne, Joshua B Tenenbaum, and Steven Pinker. A critical period for second language acquisition: Evidence from 2/3 million english speakers. *Cognition*, 177:263–277, 2018.

[87] Stefan Hawelka and Heinz Wimmer. Impaired visual processing of multi-element arrays is associated with increased number of eye movements in dyslexic reading. *Vision Research*, 45(7):855–863, 2005.

[88] Stefan Hawelka and Heinz Wimmer. Visual target detection is not impaired in dyslexic readers. *Vision Research*, 48(6):850–852, 2008.

[89] Jeffrey Heer and Michael Bostock. Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 203–212, New York, NY, USA, 2010. ACM.

[90] Ellen J Hoffman and Sanjay J Mathew. Anxiety disorders: a comprehensive review of pharmacotherapies. *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine: A Journal of Translational and Personalized Medicine*, 75(3):248–262, 2008.

[91] John J Horton, David G Rand, and Richard J Zeckhauser. The online laboratory: Conducting experiments in a real labor market. *Experimental economics*, 14(3):399–425, 2011.

[92] David Hosking. The theory of critical disability theory. 2008.

[93] Hsiu-Fang Hsieh and Jing-Jy Wang. Effect of reminiscence therapy on depression in older adults: a systematic review. *International journal of nursing studies*, 40(4):335–345, 2003.

[94] Bernd Huber, Katharina Reinecke, and Krzysztof Z. Gajos. The effect of performance feedback on social media sharing at volunteer-based online experiment platforms. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '17, pages 1882–1886, New York, NY, USA, 2017. ACM.

[95] Charles Hulme and Margaret J Snowling. *Developmental disorders of language learning and cognition*. John Wiley & Sons, 2013.

[96] Steven E Hyman. The diagnosis of mental disorders: the problem of reification. *Annual review of clinical psychology*, 6:155–179, 2010.

[97] Alexa Internet Inc. The alexa top sites services, 2017. Accessed 15-August-2018.

[98] Adrienne Ingrum. High school dropout determinants: The effect of poverty and learning disabilities. *The Park Place Economist, XIV*, pages 73–79, 2006.

[99] International Organization for Standardization. 9241-9 Ergonomic requirements for office work with visual display terminals (VDTs)-Part 9: Requirements for non-keyboard input devices, 2000.

[100] P. Ipeirotis. Demographics of mechanical turk. NYU Working Paper No. CEDER-10-01, March 2010.

[101] Panagiotis G Ipeirotis. Demographics of mechanical turk. 2010.

[102] Ergonomics of Human-system Interaction — Part 303: Requirements for Electronic Visual Displays. Standard, International Organization for Standardization, Geneva, CH, November 2008.

[103] James E Jackson. Towards universally accessible typography: A review of research on dyslexia. 2014.

[104] Jeffrey L Jackson, Mark Passamonti, and Kurt Kroenke. Outcome and impact of mental disorders in primary care at 5 years. *Psychosomatic Medicine*, 69(3):270–276, 2007.

[105] Vidhi Jain and Prakhar Agarwal. Symptomatic diagnosis and prognosis of psychiatric disorders through personal gadgets. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 118–123, 2017.

[106] Joseph R Jenkins, Lynn S Fuchs, Paul Van Den Broek, Christine Espin, and Stanley L Deno. Sources of individual differences in reading comprehension and reading fluency. *Journal of Educational Psychology*, 95(4):719, 2003.

[107] Stefan Johansson, Jan Gulliksen, and Ann Lantz. User participation when users have mental and cognitive disabilities. In *Proceedings of the 17th international ACM SIGACCESS conference on Computers and Accessibility*, ASSETS '15, pages 69–76, New York, NY, USA, 2015. ACM.

[108] Sung Jun Joo, Alex L White, Douglas J Strodtman, and Jason D Yeatman. Optimizing text for an individual's visual system: The contribution of visual crowding to reading difficulties. *Cortex*, 103:291–301, 2018.

[109] Lisa D Kahan and D Dean Richards. The effects of context on referential communication strategies. *Child development*, pages 1130–1141, 1986.

[110] Simeon Keates and Shari Trewin. Effect of age and parkinson's disease on cursor positioning using a mouse. In *Proceedings of the 7th international ACM SIGACCESS conference on Computers and Accessibility*, ASSETS 05', pages 68–75. ACM, 2005.

[111] Martin B Keller, Philip W Lavori, Jean Endicott, William Coryell, and Gerald L Klerman. Double depression": two-year follow-up. *Am J Psychiatry*, 140(6):689–694, 1983.

[112] Robert Kendell and Assen Jablensky. Distinguishing between the validity and utility of psychiatric diagnoses. *American journal of psychiatry*, 160(1):4–12, 2003.

[113] Kenneth S Kendler, Michael C Neale, Ronald C Kessler, Andrew C Heath, and Lindon J Eaves. Major depression and generalized anxiety disorder: same genes,(partly) different environments? *Archives of general psychiatry*, 49(9):716–722, 1992.

[114] Szabolcs Kéri, O Kelemen, G Szekeres, N Bagoczky, R Erdelyi, A Antal, G Benedek, and Z Janka. Schizophrenics know more than they can tell: probabilistic classification learning in schizophrenia. *Psychological medicine*, 30(1):149–155, 2000.

[115] Szabolcs Kéri, Csaba Szlobodnyik, György Benedek, Zoltán Janka, and Júlia Gádoros. Probabilistic classification learning in tourette syndrome. *Neuropsychologia*, 40(8):1356–1362, 2002.

[116] Caroline J Ketcham, Rachael D Seidler, Arend W A Van Gemmert, and George E Stelmach. Age-related kinematic differences as influenced by task difficulty, target size, and movement amplitude. *J Gerontol B Psychol Sci Soc Sci*, 57(1):P54–64, January 2002.

[117] Arif Khan, James Faucett, Pesach Lichtenberg, Irving Kirsch, and Walter A Brown. A systematic review of comparative efficacy of treatments and controls for depression. *PloS one*, 7(7), 2012.

[118] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. 1975.

[119] Shinobu Kitayama, Sean Duffy, Tadashi Kawamura, and Jeff T Larsen. Perceiving an object and its context in different cultures: A cultural look at new look. *Psychological science*, 14(3):201–206, 2003.

[120] Barbara J Knowlton, Larry R Squire, and Mark A Gluck. Probabilistic classification learning in amnesia. *Learning & Memory*, 1(2):106–120, 1994.

[121] Steven Komarov, Katharina Reinecke, and Krzysztof Z. Gajos. Crowdsourcing performance evaluations of user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 207–216, 2013.

[122] Yiren Kong, Young Sik Seo, and Ling Zhai. Comparison of reading performance on screen and on paper: A meta-analysis. *Computers & Education*, 123:138–149, 2018.

[123] Robert F Krueger and Kristian E Markon. Reinterpreting comorbidity: A model-based approach to understanding and classifying psychopathology. *Annu. Rev. Clin. Psychol.*, 2:111–133, 2006.

[124] Sri Hastuti Kurniawan and Panayiotis Zaphiris. Reading online or on paper: Which is faster? 2001.

[125] Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13):1–26, 2017.

[126] Kevin Larson and Rosalind Picard. The aesthetics of reading. In *Appears in Human-Computer Interaction Consortium Conference, Snow Mountain Ranch, Fraser, Colorado*, 2005.

[127] Talia Lavie and Noam Tractinsky. Assessing dimensions of perceived visual aesthetics of web sites. *Int. J. Hum.-Comput. Stud.*, 60(3):269–298, March 2004.

[128] Seija Leinonen, Kurt Müller, Paavo HT Leppänen, Mikko Aro, Timo Ahonen, and Heikki Lyytinen. Heterogeneity in adult dyslexic readers: Relating processing skills to the speed and accuracy of oral text reading. *Reading and Writing*, 14(3-4):265–296, 2001.

[129] Dennis M Levi. Crowding—an essential bottleneck for object recognition: A mini-review. *Vision research*, 48(5):635–654, 2008.

[130] Dennis M Levi, Srividhya Hariharan, and Stanley A Klein. Suppressive and facilitatory spatial interactions in peripheral vision: Peripheral crowding is neither size invariant nor simple contrast masking. *Journal of vision*, 2(2):3–3, 2002.

[131] HCCH Levitt. Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical society of America*, 49(2B):467–477, 1971.

[132] Clayton Lewis. Hci for people with cognitive disabilities. *SIGACCESS Access. Comput.*, (83):12–17, September 2005.

[133] Laura Foran Lewis. Exploring the experience of self-diagnosis of autism spectrum disorder in adults. *Archives of psychiatric nursing*, 30(5):575–580, 2016.

[134] Guo Li, Xiaomu Zhou, Tun Lu, Jiang Yang, and Ning Gu. Sunforum: Understanding depression in a chinese online community. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 515–526, 2016.

[135] Qisheng Li, Krzysztof Z Gajos, and Katharina Reinecke. Volunteer-based online studies with older adults and people with disabilities. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 229–241. ACM, 2018.

[136] Qisheng Li, Sung Jun Joo, Jason D Yeatman, and Katharina Reinecke. Controlling for participants' viewing distance in large-scale, psychophysical online experiments using a virtual chinrest. *Scientific reports*, 10(1):1–11, 2020.

[137] Qisheng Li, Josephine Lee, Christina Zhang, and Katharina Reinecke. How online tests contribute to the support system for people with cognitive and mental disabilities. In *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–15, 2021.

[138] Qisheng Li, Meredith Ringel Morris, Adam Fourney, Kevin Larson, and Katharina Reinecke. The impact of web browser reader views on reading speed and user experience. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press, 2019.

[139] Bruce G Link and Jo C Phelan. Stigma and its public health implications. *The Lancet*, 367(9509):528–529, 2006.

[140] Yang Liu and Jeffrey Heer. Somewhere over the rainbow: An empirical assessment of quantitative colormaps. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 598:1–598:12. ACM, 2018.

[141] Lori A Lott, Marilyn E Schneck, Gunilla Haegerström-portnoy, John A Brabyn, Ginny L Gildengorin, Catherine G West, et al. Reading performance in older adults with good acuity. *Optometry and Vision Science*, 78(5):316–324, 2001.

[142] William J Lovegrove, Alison Bowling, D Badcock, and Mary Blackwood. Specific reading disability: differences in contrast sensitivity as a function of spatial frequency. *Science*, 210(4468):439–440, 1980.

[143] Manuel Madriaga. Enduring disablism: Students with dyslexia and their pathways into uk higher education and beyond. *Disability & Society*, 22(4):399–412, 2007.

[144] Sascha Mahlke. Visual aesthetics and the user experience. In Marc Hassenzahl, Gitte Lindgaard, Axel Platz, and Noam Tractinsky, editors, *The Study of Visual Aesthetics in Human-Computer Interaction*, number 08292 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany, 2008. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany.

[145] Rahmatri Mardiko. Project vizweb, 2018. Accessed 11-September-2018.

[146] Rachel Marsh, Gerianne M Alexander, Mark G Packard, Hongtu Zhu, Jeffrey C Wingard, Georgette Quackenbush, and Bradley S Peterson. Habit learning in tourette syndrome: a translational neuroscience approach to a developmental psychopathology. *Archives of general psychiatry*, 61(12):1259–1268, 2004.

[147] Marialuisa Martelli, Gloria Di Filippo, Donatella Spinelli, and Pierluigi Zoccolotti. Crowding, reading, and developmental dyslexia. *Journal of vision*, 9(4):14–14, 2009.

[148] Winter Mason and Siddharth Suri. Conducting behavioral research on amazon's mechanical turk. *Behavior research methods*, 44(1):1–23, 2012.

[149] Ignatius G Mattingly, J Kavanagh, and I Mattingly. Reading, the linguistic process, and linguistic awareness. 1972.

[150] Jason S McCarley, Yusuke Yamani, Arthur F Kramer, and Jeffrey RW Mounts. Age, clutter, and competitive selection. *Psychology and Aging*, 27(3):616, 2012.

[151] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019.

[152] Microsoft. Reading view, 2017. Accessed 15-September-2018.

[153] Microsoft. Learning tools, 2018. Accessed 10-September-2018.

[154] Alec L Miller, Jennifer J Muehlenkamp, and Colleen M Jacobson. Fact or fiction: Diagnosing borderline personality disorder in adolescents. *Clinical psychology review*, 28(6):969–981, 2008.

[155] Aliaksei Miniukovich, Antonella De Angeli, Simone Sulpizio, and Paola Venuti. Design guidelines for web readability. In *Proceedings of the 2017 Conference on Designing Interactive Systems*, DIS '17, pages 285–296, New York, NY, USA, 2017. ACM.

[156] Ramin Mojtabai. Unmet need for treatment of major depression in the united states. *Psychiatric Services*, 60(3):297–305, 2009.

[157] Thato Monei and Athena Pedro. A systematic review of interventions for children presenting with dyscalculia in primary schools. *Educational Psychology in Practice*, 33(3):277–293, 2017.

[158] Meredith Ringel Morris, Adam Fourney, Abdullah Ali, and Laura Vonessen. Understanding the needs of searchers with dyslexia. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 35:1–35:12, New York, NY, USA, 2018. ACM.

[159] Mozilla. Firefox reader view for clutter-free web pages, 2018. Accessed 23-June-2018.

[160] Kate Nation and Margaret J Snowling. Semantic processing and the development of word-recognition skills: Evidence from children with reading comprehension difficulties. *Journal of Memory and Language*, 39(1):85–101, 1998.

[161] Jason M Nelson and Noel Gregg. Depression and anxiety among transitioning adolescents and college students with ADHD, dyslexia, or comorbid ADHD/dyslexia. *Journal of Attention Disorders*, 16(3):244–254, 2012.

[162] Sarah Nettleton, Roger Burrows, and Lisa O'Malley. The mundane realities of the everyday lay use of the internet for health, and their consequences for media convergence. *Sociology of health & illness*, 27(7):972–992, 2005.

[163] Mark W Newman, Debra Lauterbach, Sean A Munson, Paul Resnick, and Margaret E Morris. It's not that i don't have problems, i'm just not putting them on facebook: challenges and opportunities in using online social networks for health. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pages 341–350, 2011.

[164] Redhwan Nour. Web searching by individuals with cognitive disabilities. *SIGACCESS Access. Comput.*, (111):19–25, July 2015.

[165] Jan M Noyes and Kate J Garland. Computer- vs. paper-based tasks: Are they equivalent? *Ergonomics*, 51(9):1352–1375, 2008.

[166] Cliodhna O'Connor, Irini Kadianaki, Kristen Maunder, and Fiona McNicholas. How does psychiatric diagnosis affect young people's self-concept and social identity? a systematic review and synthesis of the qualitative literature. *Social Science & Medicine*, 212:94–119, 2018.

[167] Nigini Oliveira, Eunice Jun, and Katharina Reinecke. Citizen science opportunities in volunteer-based online experiments. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 6800–6812, 2017.

[168] World Health Organization. *International statistical classification of diseases and related health problems*, volume 1. World Health Organization, 2004.

[169] Cynthia Owsley. Aging and vision. *Vision research*, 51(13):1610–1622, 2011.

[170] Kathleen O'Leary, Arpita Bhattacharya, Sean A. Munson, Jacob O. Wobbrock, and Wanda Pratt. Design opportunities for mental health peer support technologies. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, page 1470–1484, New York, NY, USA, 2017. Association for Computing Machinery.

[171] Denis G Pelli, Melanie Palomares, and Najib J Majaj. Crowding is unlike ordinary masking: Distinguishing feature integration from detection. *Journal of vision*, 4(12):12–12, 2004.

[172] Denis G Pelli and Katharine A Tillman. The uncrowded window of object recognition. *Nature neuroscience*, 11(10):1129, 2008.

[173] Denis G Pelli, Katharine A Tillman, Jeremy Freeman, Michael Su, Tracey D Berger, and Najib J Majaj. Crowding and eccentricity determine reading rate. *Journal of vision*, 7(2):20–20, 2007.

[174] Timothy J Perfect and Elizabeth A Maylor. *Rejecting the dull hypothesis: The relation between method and theory in cognitive aging research.* Oxford University Press, 2000.

[175] Jessica L Peters, Lauren De Losa, Edith L Bavin, and Sheila G Crewther. Efficacy of dynamic visuo-attentional interventions for reading in dyslexic and neurotypical children: A systematic review. *Neuroscience & Biobehavioral Reviews*, 2019.

[176] Helen Petrie, Fraser Hamilton, Neil King, and Pete Pavan. Remote usability evaluations with disabled people. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, pages 1133–1141, New York, NY, USA, 2006. ACM.

[177] Ria Poole, Daniel Smith, and Sharon Simpson. How patients contribute to an online psychoeducation forum for bipolar disorder: A virtual participant observation study. *JMIR mental health*, 2(3):e21, 2015.

[178] John Powell and Aileen Clarke. Investigating internet use by mental health service users: interview study. In *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems*, page 1112. IOS Press, 2007.

[179] Lorraine A Ramig. Effects of physiological aging on speaking and reading rates. *Journal of communication disorders*, 16(3):217–226, 1983.

[180] Keith Rayner, Elizabeth R Schotter, Michael EJ Masson, Mary C Potter, and Rebecca Treiman. So much to read, so little time: How do we read, and can speed reading help? *Psychological Science in the Public Interest*, 17(1):4–34, 2016.

[181] Stian Reimers and Elizabeth A Maylor. Task switching across the life span: effects of age on general and specific switch costs. *Developmental psychology*, 41(4):661, 2005.

[182] Stian Reimers and Neil Stewart. Adobe flash as a medium for online experimentation: A test of reaction time measurement capabilities. *Behavior Research Methods*, 39(3):365–370, 2007.

[183] Stian Reimers and Neil Stewart. Presentation and response timing accuracy in adobe flash and html5/javascript web experiments. *Behavior research methods*, 47(2):309–327, 2015.

[184] Katharina Reinecke, David R. Flatla, and Christopher Brooks. Enabling Designers to Foresee Which Colors Users Cannot See. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '16, pages 2693–2704, 2016.

[185] Katharina Reinecke and Krzysztof Z Gajos. Quantifying visual preferences around the world. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 11–20, 2014.

[186] Katharina Reinecke and Krzysztof Z. Gajos. Labinthewild: Conducting large-scale online experiments with uncompensated samples. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work Social Computing*, CSCW '15, page 1364–1378, New York, NY, USA, 2015. Association for Computing Machinery.

[187] Katharina Reinecke, Tom Yeh, Luke Miratrix, Rahmatri Mardiko, Yuechen Zhao, Jenny Liu, and Krzysztof Z. Gajos. Predicting users' first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 2049–2058, New York, NY, USA, 2013. ACM.

[188] Luz Rello. Dyslexia and web accessibility: Synergies and challenges. In *Proceedings of the 12th Web for All Conference*, W4A '15, pages 9:1–9:4, New York, NY, USA, 2015. ACM.

[189] Luz Rello and Ricardo Baeza-Yates. Good fonts for dyslexia. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '13, pages 14:1–14:8, New York, NY, USA, 2013. ACM.

[190] Luz Rello and Ricardo Baeza-Yates. The effect of font type on screen readability by people with dyslexia. *ACM Trans. Access. Comput.*, 8(4):15:1–15:33, May 2016.

[191] Luz Rello and Ricardo Baeza-Yates. How to present more readable text for people with dyslexia. *Univers. Access Inf. Soc.*, 16(1):29–49, March 2017.

[192] Luz Rello and Jeffrey P. Bigham. Good background colors for readers: A study of people with and without dyslexia. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '17, pages 72–80, New York, NY, USA, 2017. ACM.

[193] Luz Rello, Gaurang Kanvinde, and Ricardo Baeza-Yates. 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*, W4A '12, pages 36:1–36:9, New York, NY, USA, 2012. ACM.

[194] William Revelle. *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois, 2018. R package version 1.8.12.

[195] Kathryn E. Ringland, Jennifer Nicholas, Rachel Kornfield, Emily G. Lattie, David C. Mohr, and Madhu Reddy. Understanding mental ill-health as psychosocial disability: Implications for assistive technology. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, page 156–170, New York, NY, USA, 2019. Association for Computing Machinery.

[196] Melisa Robichaud, Naomi Koerner, and Michel J Dugas. *Cognitive behavioral treatment for generalized anxiety disorder: From science to practice*. Routledge, 2019.

[197] Thomas L Rodebaugh, Robert M Holaway, and Richard G Heimberg. The treatment of social anxiety disorder. *Clinical Psychology Review*, 24(7):883–908, 2004.

[198] Claudia Rodríguez-Aranda. Reduced writing and reading speed and age-related changes in verbal fluency tasks. *The Clinical Neuropsychologist*, 17(2):203–215, 2003.

[199] Klaus Rohrschneider. Determination of the location of the fovea on the fundus. *Investigative ophthalmology & visual science*, 45(9):3257–3258, 2004.

[200] Avinoam B Safran, Bernadette Mermillod, Christophe Mermoud, C De Weisse, and Dominique Desangles. Characteristic features of blind spot size and location, when evaluated with automated perimetry: Values obtained in normal subjects. *Neuro-ophthalmology*, 13(6):309–315, 1993.

[201] Pedro Sanches, Axel Janson, Pavel Karpashevich, Camille Nadal, Chengcheng Qu, Claudia Daudén Roquet, Muhammad Umair, Charles Windlin, Gavin Doherty, Kristina Höök, et al. Hci and affective health: Taking stock of a decade of studies and charting future research directions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2019.

[202] Shruti Sannon, Elizabeth L. Murnane, Natalya N. Bazarova, and Geri Gay. "i was really, really nervous posting it": Communicating about invisible chronic illnesses across social media platforms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA, 2019. Association for Computing Machinery.

[203] Marian Sauter, Dejan Draschkow, and Wolfgang Mack. Building, hosting and recruiting: A brief introduction to running behavioral experiments online. *Brain sciences*, 10(4):251, 2020.

[204] Enrico Schulz, Urs Maurer, Sanne van der Mark, Kerstin Bucher, Silvia Brem, Ernst Martin, and Daniel Brandeis. Impaired semantic processing during sentence reading in children with dyslexia: combined fmri and erp evidence. *Neuroimage*, 41(1):153–168, 2008.

[205] Charles T Scialfa, Sheila Cordazzo, Katherine Bubric, and John Lyon. Aging and visual crowding. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 68(4):522–528, 2012.

[206] Andrew Sears and Vicki L. Hanson. Representing users in accessibility research. *ACM Trans. Access. Comput.*, 4(2):7:1–7:6, March 2012.

[207] Ruth S Shalev. Developmental dyscalculia. *Journal of child neurology*, 19(10):765–771, 2004.

[208] Stewart A Shankman, Peter M Lewinsohn, Daniel N Klein, Jason W Small, John R Seeley, and Sarah E Altman. Subthreshold conditions as precursors for full syndrome disorders: a 15-year longitudinal study of multiple diagnostic classes. *Journal of Child Psychology and Psychiatry*, 50(12):1485–1494, 2009.

[209] Sally E Shaywitz. Dyslexia. *New England Journal of Medicine*, 338(5):307–312, 1998.

[210] Mark M Shovman and Merav Ahissar. Isolating the impact of visual perception on dyslexics' reading ability. *Vision research*, 46(20):3514–3525, 2006.

[211] Travis Simcox and Julie A Fiez. Collecting response times using amazon mechanical turk and adobe flash. *Behavior research methods*, 46(1):95–111, 2014.

[212] Jaana Simola, Jarmo Kuisma, Anssi Öörni, Liisa Uusitalo, and Jukka Hyönä. The impact of salient advertisements on reading and attention on web pages. *Journal of Experimental Psychology: Applied*, 17(2):174, 2011.

[213] Nicholas A. Smith, Isaac E. Sabat, Larry R. Martinez, Kayla Weaver, and Shi Xu. A convenient solution: Using mturk to sample from hard-to-reach populations. *Industrial and Organizational Psychology*, 8(2):220–228, 2015.

[214] Brisa Solé, Esther Jiménez, Carla Torrent, Maria Reinares, Caterina del Mar Bonnin, Imma Torres, Cristina Varo, Iria Grande, Elia Valls, Estela Salagre, et al. Cognitive impairment in bipolar disorder: treatment and prevention strategies. *International Journal of Neuropsychopharmacology*, 20(8):670–680, 2017.

[215] Tobias Sonne, Paul Marshall, Carsten Obel, Per Hove Thomsen, and Kaj Grønbæk. An assistive technology design framework for adhd. In *Proceedings of the 28th Australian Conference on Computer-Human Interaction*, OzCHI '16, page 60–70, New York, NY, USA, 2016. Association for Computing Machinery.

[216] Donatella Spinelli, Maria De Luca, Anna Judica, and Pierluigi Zoccolotti. Crowding effects on word identification in developmental dyslexia. *Cortex*, 38(2):179–200, 2002.

[217] Zachary Steel, Claire Marnane, Changiz Iranpour, Tien Chey, John W Jackson, Vikram Patel, and Derrick Silove. The global prevalence of common mental disorders: a systematic review and meta-analysis 1980–2013. *International journal of epidemiology*, 43(2):476–493, 2014.

[218] Saul Sternberg. High-speed scanning in human memory. *Science*, 153(3736):652–654, 1966.

[219] Emma Stone and Mark Priestley. Parasites, pawns and partners: disability research and the role of non-disabled researchers. *British journal of sociology*, pages 699–716, 1996.

[220] David Sun, Pablo Paredes, and John Canny. Moustress: detecting stress from mouse motion. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 61–70, 2014.

[221] John Swain and Sally French. Towards an affirmation model of disability. *Disability & society*, 15(4):569–582, 2000.

[222] Saiganesh Swaminathan, Kotaro Hara, and Jeffrey P. Bigham. The crowd work accessibility problem. In *Proceedings of the 14th Web for All Conference on The Future of Accessible Work*, W4A '17, pages 6:1–6:4, New York, NY, USA, 2017. ACM.

[223] Rachel Z Tenenbaum, Conor J Byrne, and Jason J Dahling. Interactive effects of physical disability severity and age of disability onset on riasec self-efficacies. *Journal of Career Assessment*, 22(2):274–289, 2014.

[224] International Telecommunication Union. ICT Facts & Figures: The World in 2015, 2015.

[225] Rajan Vaish, Snehalkumar Neil S Gaikwad, Geza Kovacs, Andreas Veit, Ranjay Krishna, Imanol Arrieta Ibarra, Camelia Simoiu, Michael Wilber, Serge Belongie, Sharad Goel, et al. Crowd research: Open and scalable university laboratories. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, pages 829–843. ACM, 2017.

[226] Jordi Vallverdú and David Casacuberta. Ethical and technical aspects of emotions to create empathy in medical machines. In *Machine Medical Ethics*, pages 341–362. Springer, 2015.

[227] Ronald Van den Berg, Jos BTM Roerdink, and Frans W Cornelissen. On the generality of crowding: Visual crowding in size, saturation, and hue compared to orientation. *Journal of Vision*, 7(2):14–14, 2007.

[228] Pascal WM Van Gerven, Fred Paas, Jeroen JG Van Merriënboer, and Henk G Schmidt. Memory load and the cognitive pupillary response in aging. *Psychophysiology*, 41(2):167–174, 2004.

[229] Yvonne Vezzoli, Asimina Vasalou, and Kaska Porayska-Pomsta. Dyslexia in sns: An exploratory study to investigate expressions of identity and multimodal literacies. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW), December 2017.

[230] Susanna N Visser, Melissa L Danielson, Rebecca H Bitsko, Joseph R Holbrook, Michael D Kogan, Reem M Ghandour, Ruth Perou, and Stephen J Blumberg. Trends in the parent-report of health care provider-diagnosed and medicated attention-deficit/hyperactivity disorder: United states, 2003–2011. *Journal of the American Academy of Child & Adolescent Psychiatry*, 53(1):34–46, 2014.

[231] Stephanie von Ammon Cavanaugh. Depression in the medically ill: Critical issues in diagnostic assessment. *Psychosomatics: Journal of Consultation and Liaison Psychiatry*, 1995.

[232] Greg Wadley, Reeva Lederman, John Gleeson, and Mario Alvarez-Jimenez. Participatory design of an online therapy for youth mental health. In *Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration*, OzCHI '13, page 517–526, New York, NY, USA, 2013. Association for Computing Machinery.

[233] N. Walker, D. A. Philbin, and A. D. Fisk. Age-related differences in movement control: adjusting submovement structure to optimize performance. *J Gerontol B Psychol Sci Soc Sci*, 52(1), January 1997.

[234] Mengyu Wang, Lucy Q Shen, Michael V Boland, Sarah R Wellik, Carlos Gustavo De Moraes, Jonathan S Myers, Peter J Bex, and Tobias Elze. Impact of natural blind spot location on perimetry. *Scientific reports*, 7(1):6143, 2017.

[235] Julia Shuppert West, Lynda Kayser, Paul Overton, and Robert Saltmarsh. Student perceptions that inhibit the initiation of counseling. *The School Counselor*, 39(2):77–83, 1991.

[236] Ryen W White and Eric Horvitz. Cyberchondria: studies of the escalation of medical concerns in web search. *ACM Transactions on Information Systems (TOIS)*, 27(4):1–37, 2009.

[237] David Whitney and Dennis M Levi. Visual crowding: A fundamental limit on conscious perception and object recognition. *Trends in cognitive sciences*, 15(4):160–168, 2011.

[238] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[239] Coralie Joy Wilson, Frank P Deane, Kellie L Marshall, and Andrew Dalley. Reducing adolescents' perceived barriers to treatment and increasing help-seeking intentions: effects of classroom presentations by general practitioners. *Journal of Youth and Adolescence*, 37(10):1257–1269, 2008.

[240] Wolf P Wolfensberger, Bengt Nirje, Simon Olshansky, Robert Perske, and Philip Roos. The principle of normalization in human services. 1972.

[241] Teng Ye, Katharina Reinecke, and Lionel P Robert Jr. Personalized feedback versus money: The effect on reliability of subjective data in online experimental platforms. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 343–346. ACM, 2017.

[242] Gerry Zarb. On the road to damascus: First steps towards changing the relations of disability research production. *Disability, Handicap & Society*, 7(2):125–138, 1992.

[243] Rolf A Zwaan and Diane Pecher. Revisiting mental simulation in language comprehension: Six replication attempts. *PloS one*, 7(12):e51382, 2012.

[244] Kathryn Zyskowski, Meredith Ringel Morris, Jeffrey P. Bigham, Mary L. Gray, and Shaun K. Kane. Accessible crowdwork?: Understanding the value in and challenge of microtask employment for people with disabilities. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 1682–1693, New York, NY, USA, 2015. ACM.