Volunteer-Based Online Studies With Older Adults and People with Disabilities

Qisheng Li
Paul G. Allen School of Computer Science & Engineering
University of Washington
liqs@cs.washington.edu

Krzysztof Z. Gajos
School of Engineering and Applied Sciences
Harvard University
kgajos@eecs.harvard.edu

Katharina Reinecke
Paul G. Allen School of Computer Science & Engineering
University of Washington
reinecke@cs.washington.edu

ABSTRACT
There are few large-scale empirical studies with people with disabilities or older adults, mainly because recruiting participants with specific characteristics is even harder than recruiting young and/or non-disabled populations. Analyzing four online experiments on LabintheWild with a total of 355,656 participants, we show that volunteer-based online experiments that provide personalized feedback attract large numbers of participants with diverse disabilities and ages and allow robust studies with these populations that replicate and extend the findings of prior laboratory studies. To find out what motivates people with disabilities to take part, we additionally analyzed participants' feedback and forum entries that discuss LabintheWild experiments. The results show that participants use the studies to diagnose themselves, compare their abilities to others, quantify potential impairments, self-experiment, and share their own stories – findings that we use to inform design guidelines for online experiment platforms that adequately support and engage people with disabilities.

ACM Classification Keywords
H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

Author Keywords
Online Experimentation; Volunteers; Elderly People; People With Disabilities

INTRODUCTION
Soliciting information from people with disabilities and older adults is important for research and industry projects alike. Yet many researchers have struggled to recruit users that meet particular characteristics in sufficiently large numbers [12, 37]. Traditional recruiting methods through gatekeepers, such as local organizations or advocacy groups [3], and/or establishing local participant pools are time-intensive and often expensive. They also risk generalizability and data quality if the same small number of participants repeatedly take part in similar studies [40].

To study older adults and people with disabilities, some researchers have therefore turned to the online labor markets, such as Amazon Mechanical Turk (MTurk) [44, 11, 41], which support the efficient recruitment of participants at low cost. Online experimentation can be beneficial for people with disabilities; aside from receiving a financial compensation, participating online is often more convenient and feasible than having to travel to a laboratory [49, 8, 14]. However, online experiments can also provide challenges for disabled and elderly users. For example, researchers have found that usability problems and limitations of MTurk can make it inaccessible for people with disabilities [10, 43, 49], who often struggle to find tasks that match their abilities [49].

The goal of this paper 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 [39]. 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 first replicated four laboratory studies on LabintheWild, all of which offered tasks that were known to be impacted by various disabilities or age-related

ASSETS '18, October 22–24, 2018, Galway, Ireland

 ISBN 978-1-4503-5650-3/18/10…$15.00
DOI: https://doi.org/10.1145/3234695.3236360

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.

RELATED WORK

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 [12, 37], 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 [3], frequently establishing a local participant pool that can be used over time. For example, Johansson et al. [25] recruited participants with mental and cognitive disabilities through a local member-driven organization using snowball sampling. Similarly, the SiDE user pool [12, 13] 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 [13].

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 evaluated alternative ways to recruit and study participants. For example, researchers increasingly recruit participants through online labor markets, such as Amazon Mechanical Turk (MTurk) [4, 21, 9]. Compared to traditional laboratory experiments, online studies offer faster and more effortless participant recruitment, as well as larger and more diverse samples [18, 24, 4, 34]. 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 [18, 16, 20, 32, 39].

Researchers have also used MTurk to conduct studies with people with disabilities and other specific populations. Tenenbaum et al. [44], for example, recruited 153 individuals with physical disabilities and studied how various impairment-related factors influence vocational self-efficacies. Carr [11] 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. [41] 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 [49, 8]. People with disabilities will also often receive a sense of self-worth, self-efficacy and autonomy when participating in such studies [14]. They are motivated to contribute to scientific research [12] and to cognitively benefit from doing a task [8].

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 [49, 10]. 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 [43, 10]. 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 [49]. As a result of such barriers, the diversity of people on Mechanical Turk is still limited, both in terms of people with disabilities [10], and in terms of age range [8]. The following section introduces an alternative methodology for studying such diverse populations.

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 [22]. Second, it exposes participants to scientific concepts and increases their interest in research and scientific findings [35]. 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 [16, 18, 39]. 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 [39].

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 [45]. 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., [16, 39, 19] and Table 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 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 [38]; (2) A study comparing human listening rates between sighted, low-vision, and blind people [7], 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.

STUDYING PEOPLE WITH DISABILITIES AND OLDER ADULTS ON LABINTHEWILD

We replicated four studies on LabintheWild, chosen to represent a broad range of tasks (see Table 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.

Study 1: Weather Prediction Study

Our first study is a modification of Knowlton et al.'s study from 1994 [31], 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 [28, 33, 27, 15]. In particular, Gabay et al. [15] showed that adults with dyslexia performed better in the Weather 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 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 [15]) 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 [15] 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 [15], 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

Table 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 [31]	How quickly do you learn?	Amnesia [31], Dyslexia [15], Tourette syndrome [28, 33]	22	3,786	328 (8.66%)	52.76	5-99 (m=25, sd=11.5)
Memory [42]	How fast is your memory?	Cognitive decline in elderly people [47]	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 [1]	Test your social intelligence!	High-functioning Autism, Asperger's Syndrome [1]	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 [26, 48, 29]	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 [7]	How fast can you process words?	Vision Impairment [2]	2	453	143 (32%)	57.83	8-80 (m=34, sd=15)
Colorblindness [38]	Can we guess your color age?	Color vision deficiencies [6]	12	31,248	1,831 (3.85%)	70.6	5-94 (m=30, sd=15.2)

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 [31, 15]. 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 [15]. However, in contrast to Gabay et al.s finding [15], 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.

Study 2: Memory Study

Our second study is a replication of Sternberg's experiment [42], 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 [47]. 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 [36].

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 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.

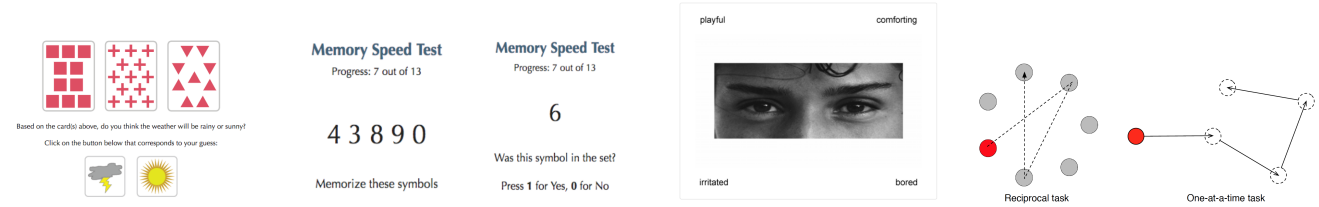
Participants

Over the course of around 40 months, 18,026 participants completed the study (see also Table 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 + $3 \times$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



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

Figure 1: Overview of the stimuli used in four of our LabintheWild experiments.

Table 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	***

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 [47]).

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 [42]. In addition, we found a significant interaction effect between age and set size (Table 2), demonstrating the complexity effect [36]. This confirms the results of [47] 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.

Study 3: Social Intelligence Study

Originally developed by Baron-Cohen et al. as the "Reading the Mind in the Eyes" test [1], 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 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 [1]. 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. [1], who had found means of 22 (sd=6.6) for ASD participants (N=15) and 26.2 (sd=3.6) for non-disabled controls.

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 [23]. 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 [26], they have longer verification times (the time interval between the end of a movement inside a target and the beginning of the click) [48], they make more pauses of 100ms [26] and their normalized jerk (fluctuations in the speed profile of the movement) is higher [29]. 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 [23]), 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 1d). Target sizes (10,

Table 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 [26].	$\beta = -0.0026, F_{1,209913} = 15370, p < 0.0001$	Yes
Older adults have longer verification times than young adults [48].	$\beta = 0.0422, F_{1,204509} = 46634, p < 0.0001$	Yes
Older adults make more pauses of 100ms than young adults [26].	$\beta = 0.0016, F_{1,204512} = 11116, p < 0.0001$	Yes
Normalized jerk is higher for older adults than for young adults [29].	$\beta = 0.0039, F_{1,204419} = 14349, p < 0.0001$	Yes

15, 25, 40, and 60 pixels) 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 [26, 48, 29]. Table 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.

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.

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 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 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 [17].

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).

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 [1], 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)

Table 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

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 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!)*

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 [30]. 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.

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:

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.

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.

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.

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 [14], or co-authorships offered for involvement in the larger research cycle, similar to Stanford's crowd research project [46].

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 [35], 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. [35] 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 [5]). 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.

REFERENCES

1. Simon Baron-Cohen, Sally Wheelwright, Jacqueline Hill, Yogini Raste, and Ian Plumb. 2001. 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 (2001), 241–251.
2. Marialena Barouti, Konstantinos Papadopoulos, and Georgios Kouroupetroglou. 2013. Synthetic and natural speech intelligibility in individuals with visual impairments: Effects of experience and presentation rate. In *European AAATE Conference, Portugal*. 695–699.
3. Heather Becker, Greg Roberts, Janet Morrison, and Julie Silver. 2004. Recruiting people with disabilities as research participants: Challenges and strategies to address them. *Mental Retardation* 42, 6 (2004), 471–475.
4. Adam J Berinsky, Gregory A Huber, and Gabriel S Lenz. 2012. Using Mechanical Turk as a subject recruitment tool for experimental research. *Political Analysis* 20 (2012), 351–68.
5. Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and others. 2010. VizWiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. ACM, 333–342.
6. J. Birch. 2001. *Diagnosis of Defective Colour Vision*. Oxford: Butterworth-Heinemann.
7. Danielle Bragg, Cynthia Bennett, Katharina Reinecke, and Richard Ladner. 2018. A Large Inclusive Study of Human Listening Rates. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM.
8. Robin Brewer, Meredith Ringel Morris, and Anne Marie Piper. 2016. "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)*. ACM, New York, NY, USA, 2246–2257.
9. Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2011. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science* 6, 1 (2011), 3–5.
10. Rocío Calvo, Shaun K. Kane, and Amy Hurst. 2014. 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)*. ACM, New York, NY, USA, 257–258.
11. Alaina Carr. 2014. An exploration of Mechanical Turk as a feasible recruitment platform for cancer survivors. *Undergraduate Honors Theses* 59 (2014).
12. Marianne Dee and Vicki L. Hanson. 2014. 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)*. ACM, New York, NY, USA, 35–42.
13. Marianne Dee and Vicki L. Hanson. 2016. A Pool of Representative Users for Accessibility Research: Seeing Through the Eyes of the Users. *ACM Trans. Access. Comput.* 8, 1, Article 4 (Jan. 2016), 31 pages.
14. Xianghua Ding, Patrick C. Shih, and Ning Gu. 2017. 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)*. ACM, New York, NY, USA, 642–654.
15. Yafit Gabay, Eli Vakil, Rachel Schiff, and Lori L Holt. 2015. Probabilistic category learning in developmental dyslexia: Evidence from feedback and paired-associate weather prediction tasks. *Neuropsychology* 29, 6 (2015), 844.
16. Laura Germine, Ken Nakayama, Bradley C Duchaine, Christopher F Chabris, Garga Chatterjee, and Jeremy B Wilmer. 2012. Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic bulletin & review* 19, 5 (2012), 847–857.
17. Barney G Glaser and Anselm L Strauss. 2009. *Discovery of grounded theory: Strategies for qualitative research*. Transaction Publishers.
18. Samuel D Gosling, Simine Vazire, Sanjay Srivastava, and Oliver P John. 2004. Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American psychologist* 59, 2 (2004), 93.
19. Joshua K Hartshorne and Laura T Germine. 2015. When does cognitive functioning peak? The asynchronous rise and fall of different cognitive abilities across the life span. *Psychological science* (2015), 0956797614567339.
20. Jeffrey Heer and Michael Bostock. 2010. Crowdsourcing graphical perception: Using Mechanical Turk to assess visualization design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. 203–212.
21. J J Horton, D G Rand, and R J Zeckhauser. 2011. The online laboratory: Conducting experiments in a real labor market. *Experimental Economics* (2011).
22. Bernd Huber, Katharina Reinecke, and Krzysztof Z. Gajos. 2017. 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)*. ACM, New York, NY, USA, 1882–1886.

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

24. P. Ipeirotis. 2010. Demographics of mechanical turk. (March 2010). NYU Working Paper No. CEDER-10-01.

25. Stefan Johansson, Jan Gulliksen, and Ann Lantz. 2015. User Participation When Users Have Mental and Cognitive Disabilities. In *Proceedings of the 17th international ACM SIGACCESS conference on Computers and Accessibility (ASSETS '15)*. ACM, New York, NY, USA, 69–76.

26. Simeon Keates and Shari Trewin. 2005. 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')*. ACM, 68–75.

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

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

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

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

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

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

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

34. Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods* 44, 1 (2012), 1–23.

35. Nigini Oliveira, Eunice Jun, and Katharina Reinecke. 2017. Citizen Science Opportunities in Volunteer-Based Online Experiments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, 6800–6812.

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

37. Helen Petrie, Fraser Hamilton, Neil King, and Pete Pavan. 2006. Remote Usability Evaluations With Disabled People. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*. ACM, New York, NY, USA, 1133–1141.

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

39. Katharina Reinecke and Krzysztof Z. Gajos. 2015. 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*. 1364–1378.

40. Andrew Sears and Vicki L. Hanson. 2012. Representing Users in Accessibility Research. *ACM Trans. Access. Comput.* 4, 2, Article 7 (March 2012), 6 pages.

41. Nicholas A. Smith, Isaac E. Sabat, Larry R. Martinez, Kayla Weaver, and Shi Xu. 2015. A Convenient Solution: Using MTurk To Sample From Hard-To-Reach Populations. *Industrial and Organizational Psychology* 8, 2 (2015), 220–228.

42. Saul Sternberg. 1966. High-speed scanning in human memory. *Science* 153, 3736 (1966), 652–654.

43. Saiganesh Swaminathan, Kotaro Hara, and Jeffrey P. Bigham. 2017. The Crowd Work Accessibility Problem. In *Proceedings of the 14th Web for All Conference on The Future of Accessible Work (W4A '17)*. ACM, New York, NY, USA, Article 6, 4 pages.

44. Rachel Z Tenenbaum, Conor J Byrne, and Jason J Dahling. 2014. Interactive effects of physical disability severity and age of disability onset on RIASEC self-efficacies. *Journal of Career Assessment* 22, 2 (2014), 274–289.

45. International Telecommunication Union. 2015. ICT Facts & Figures: The World in 2015. (2015).

46. Rajan Vaish, Snehalkumar Neil S Gaikwad, Geza Kovacs, Andreas Veit, Ranjay Krishna, Imanol Arrieta Ibarra, Camelia Simoiu, Michael Wilber, Serge Belongie, Sharad Goel, and others. 2017. Crowd Research: Open and Scalable University Laboratories. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*. ACM, 829–843.

47. Pascal WM Van Gerven, Fred Paas, Jeroen JG Van Merriënboer, and Henk G Schmidt. 2004. Memory load and the cognitive pupillary response in aging. *Psychophysiology* 41, 2 (2004), 167–174.

48. N. Walker, D. A. Philbin, and A. D. Fisk. 1997. Age-related differences in movement control: adjusting submovement structure to optimize performance. *J Gerontol B Psychol Sci Soc Sci* 52, 1 (January 1997).

49. Kathryn Zyskowski, Meredith Ringel Morris, Jeffrey P. Bigham, Mary L. Gray, and Shaun K. Kane. 2015. 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)*. ACM, New York, NY, USA, 1682–1693.