LabintheWild: Conducting Large-Scale Online Experiments With Uncompensated Samples

Katharina Reinecke
University of Michigan
Ann Arbor, MI 48109, USA
reinecke@umich.edu

Krzysztof Z. Gajos
Harvard School of Engineering and Applied Sciences
33 Oxford St., Cambridge, MA, USA
kgajos@eecs.harvard.edu

ABSTRACT

Web-based experimentation with uncompensated and unsupervised samples has the potential to support the replication, verification, extension and generation of new results with larger and more diverse sample populations than previously seen. We introduce the experimental online platform LabintheWild, which provides participants with personalized feedback in exchange for participation in behavioral studies. In comparison to conventional in-lab studies, LabintheWild enables the recruitment of participants at larger scale and from more diverse demographic and geographic backgrounds. We analyze Google Analytics data, participants' comments, and tweets to discuss how participants hear about the platform, and why they might choose to participate. Analyzing three example experiments, we additionally show that these experiments replicate previous in-lab study results with comparable data quality.

Author Keywords

Online experimentation; crowdsourcing; replication; social comparison; uncompensated samples, WEIRD

ACM Classification Keywords

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

INTRODUCTION

An estimated 95% of research findings in psychology, behavioral economics, and related fields are based on studies with student samples from Western and industrialized countries [1]. As a consequence, some of the knowledge derived from these usually small and locally-recruited student sample populations has been found to be non-generalizable [1, 17] — a finding that led researchers to call these participant samples "WEIRD", an acronym for Western, Educated, Industrialized, Rich, and Democratic [17].

The fact that WEIRD participants are not always representative of the broader population also affects the generalizability of findings in human-computer interaction (HCI). For instance, research has found that age and education level influence visual preferences for websites [32], and that groups in Western societies, such as in the US, are less likely to reach consensus in online scheduling systems than groups in Eastern societies, such as in China and Japan [33].

CSCW '15, March 14 - 18 2015, Vancouver, BC, Canada

http://dx.doi.org/10.1145/2675133.2675246

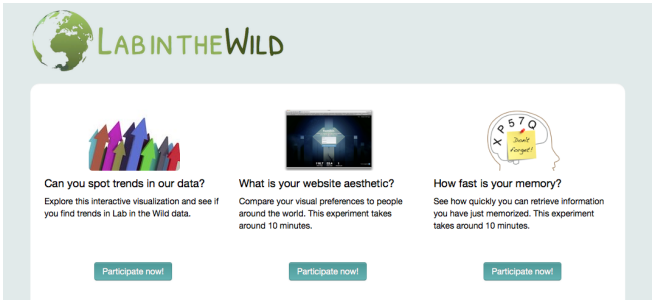

Figure 1. LabintheWild enables participants to compare themselves to others in exchange for study participation. By making experiments short, intrinsically motivating, and easy to access, LabintheWild has attracted nearly 750,000 participants from more than 200 countries.

To expand the range of available research methods, we show that online experiments with unsupervised and uncompensated samples (i.e., participants who complete experiments without a direct contact with the researcher and who do not receive financial compensation) have the potential to attract larger and more diverse participant pools than previously feasible in laboratory studies without compromising the data quality. More specifically, this paper introduces our online experiment platform LabintheWild, which enables participants to compare themselves to others in exchange for study participation. Synthesizing the major design decisions we made in the course of deploying LabintheWild, we reflect

on their impact on recruitment, participant behavior, and data quality. We show our progress toward three core objectives that guided the design of LabintheWild:

1. Larger scale: LabintheWild reaches more participants than what is feasible in typical laboratory studies. In less than two years of its existence, LabintheWild was visited more than 2 million times, and nearly 750,000 visitors completed an experiment (an average of about 1,000 participants/day). Using Google Analytics data, participants' comments, and tweets, we discuss our main sources of traffic, how people hear about LabintheWild, as well as their motivation and perceived benefit of participating.

2. Less WEIRD: In contrast to laboratory studies, LabintheWild participants report a larger age range, more diverse educational backgrounds, and — coming from more than 200 countries and six continents — a wider geographic dispersion. This data shows that LabintheWild enables comparative studies between a large number of geographic regions and a variety of demographic groups.

3. Equal data quality: We report on the results of three prior laboratory studies that we replicated on LabintheWild. Our results match those obtained in laboratory settings despite the different incentive structure on LabintheWild and despite the fact that our participants complete the tests without direct supervision. We discuss participants' role in reporting distractions, identifying untruthfully provided information, and sharing useful context.

We continue this paper with a brief overview of previous innovations for conducting experiments, before providing details about LabintheWild. The article is structured following our three core objectives as stated above, starting with an analysis of the scale, followed by an overview of how LabintheWild participants are less WEIRD, and concluding with an evaluation of the data quality in three replication experiments. We finish with an overall discussion and an overview of future work.

RELATED WORK

The importance of conducting behavioral experiments across a variety of demographic groups was perhaps most prominently raised by Henrich and colleagues, who concluded that WEIRD samples "are among the least representative populations one could find for generalizing about humans" [17, p. 1]. For certain kinds of research, online experiments provide a feasible mechanism for studying more representative populations. Mason et al. [26], for instance, suggested that crowdsourcing experiments using Amazon's online market platform Mechanical Turk can broaden the subject population and increase generalizability. Workers on Mechanical Turk, the so-called "Turkers", have been found to be more diverse and more representative of the US population than conventional student samples [19, 3].

Mechanical Turk also provides the infrastructure for recruiting large numbers of subjects at low cost (Turkers receive an average of $4.80 per hour [20]). The site has therefore become increasingly popular in various fields such as political science [3], economics [18], and psychology [4]. However, the financial incentive might also mean that some Turkers attempt to maximize their return while investing minimal effort and time into the tasks [10]. In particular, researchers worry that Turkers might not sufficiently read and focus on study instructions and materials, which could potentially compromise the data quality [28]. Much research has since shown that the data quality is indeed affected by some Turkers investing minimal effort [22, 28, 21, 10, 15]. It was also found that when including precautions in the design and analysis of experiments, such as mechanisms to detect satisficing, effective training procedures, outlier removal techniques, as well as incorporating contextual factors (e.g., hardware, demographics) in the statistical analyses, the data quality and statistical power can be comparable to that of in-lab studies [16, 18, 23, 28, 30, 31, 38]. Research that requires collection of truthful subjective responses, however, remains challenging on Mechanical Turk [22, 28, 34].

While Mechanical Turk has a rapidly growing user population, its user base is insufficiently diverse to conduct studies spanning many countries (the plurality of Turkers come from the United States and India [30, 11, 11, 26]). Eriksson and Simpson reported that of 984 Turkers who participated in their study, no other country beside the US and India reached more than 5% of the overall sample [11].

Mechanical Turk's subject pool is additionally restricted to those who have signed up as Turkers, creating a barrier to participation. Removing this sign-up barrier, Germine et al. [13] developed TestMyBrain.org to make several cognitive and perceptual studies available online for anyone to participate in exchange for personalized performance feedback. Their analysis comparing data from online and in-lab participants in four distinct experiments showed no significant differences between results obtained in the two settings [13]. Our results corroborate these findings and extend them to a broader range of types of experiments.

ABOUT LABINTHEWILD

LabintheWild (www.labinthewild.org) is an online experiment platform for conducting behavioral research studies with self-selected, uncompensated web samples. When developing LabintheWild, we had three goals in mind:

1. Larger scale: The platform had to attract sufficiently large numbers of participants to enable rapid iteration of design prototypes at low cost, and to detect effects that cannot be reliably observed in small-scale experiments. This meant that we had to rethink recruitment strategies and participant compensation.

2. Less WEIRD: Our goal was to make it possible to reach participants from diverse geographic and demographic backgrounds in sufficient numbers to reason about the generalizability of findings and compare specific populations.

3. Equal data quality: We wanted the quality of the data collected on LabintheWild to match that of conventional methods. This required mechanisms that ensured participants' engagement and truthfulness, as well as mechanisms to detect if the quality of the data was compromised.

Major Design Decisions

Unrestricted Online Access

Developing LabintheWild as an online platform was central to meeting the goals of engaging larger and less WEIRD populations as easy online access lowers the barrier to participation compared to studies conducted in physical labs. LabintheWild studies are conducted without experimenter supervision, which provides the benefit of allowing participation in large numbers and around the clock independent of location and time zones.

LabintheWild experiments are open to anyone to participate. This has several consequences: First, even if the scientific objective of an experiment is to study a specific population (e.g., people who speak two or more languages fluently), we design studies such that all can have a rewarding experience participating in a study, which sometimes requires designing multiple tracks through an experiment. Second, we carefully design studies to be appropriate for minors and other vulnerable populations.

Finally, to make participation as easy as possible, we decided that people should not be required to sign up. This decision was further reinforced by a request from our Institutional Review Boards that we do not track our participants in any way. While these decisions lower the barrier to participation and protect participants' privacy, the current design of LabintheWild precludes longitudinal studies, combining a participant's results from multiple studies, or automatically detecting participants taking the same test more than once. This last limitation is particularly important because many experiments have prominent learning effects. To maintain integrity of our data, the first question we ask in all our questionnaires is whether a participant has taken the test before. Participants who answer this question in the affirmative are typically excluded from the analysis.

Participant Incentives and Compensation

One of the major design decisions that we made was that LabintheWild should work without providing financial compensation for study participation. Paying participants would limit both the size and the diversity of the participant pool because our financial resources are limited and because only some participants have the ability to receive online payments or are interested in spending their money online (e.g., Mechanical Turk workers can receive their payments only in U.S. dollars, Indian Rupees, or as Amazon.com store credit [26]). The necessary registration would also pose a hurdle. Instead, we leverage the human urge to learn about themselves and to compare themselves to others [12]: After participating in an experiment, participants are shown a personalized results page, which explains how they did and how their performance or preferences compare to others. For each experiment, we produce a short slogan that advertises the type of feedback participants would receive (e.g., "Can we guess your age?", "Are you more Eastern or Western?", "Test your social intelligence!"). We use these slogans to advertise each experiment on LabintheWild's front page and for sharing the experiments via social media.

The personalized feedback is designed to be either value-neutral or performance-based, but always compares the participant to others. In most cases, the feedback corresponds to our own research questions. For example, in a study that compares aesthetic preferences across cultures, participants' website preferences are shown in comparison to people from the same or other countries. However, if research questions and the corresponding experiments are not inherently interesting to participants, we find an aspect of the results that might be of higher interest. For example, for a Fitts' Law experiment that required participants to click on a number of dots on a screen for a duration of around five minutes—a task that can be extremely tedious and boring—we implemented a regression model to predict a person's age from their mouse movements. This experiment's slogan was "Can we guess your age?" (rather than the more obvious "What is you pointing ability?"). Our primary research interest was clearly revealed in the informed consent form and again on the personalized results page. But participants' main incentive was to see whether we could, indeed, guess their age.

Recruitment

Designing experiments to be intrinsically motivating and providing personalized feedback plays a major role in the recruitment of participants. LabintheWild supports a self-perpetuating recruitment mechanism by enabling sharing through online social networks. LabintheWild's pages, including the first and last page of each experiment, are equipped with social sharing buttons via Facebook, Twitter, Google+, Tumblr, and email.

When we launched LabintheWild in July 2012, we first advertised three studies through our personal social networks on Facebook, Twitter, and Google+. As more and more participants took part in our experiments, we observed an increase in mentions of LabintheWild in social shares, on blog posts, and in online newspaper articles: Participants and other interested people write about and link to LabintheWild, which attracts others to the site. To further leverage the word-of-mouth and to enable users to hear about new experiments and research results, we added a Facebook page and the possibility to "like" LabintheWild in April 2013, as well as an option to sign up for email notifications in July 2013.

When an experiment is ready to go online, we initially advertise it on our own online social networks. This limited deployment gives us a chance to uncover any remaining implementation bugs and design flaws. Next, we advertise the experiment on our Facebook page and add it to the homepage of LabintheWild. We do not otherwise actively advertise or recruit participants.

Because some experiments are naturally more intrinsically motivating than others, we employ a special recruitment strategy among participants who just completed a study. Below their personalized results, we suggest two other LabintheWild experiments. Consequently, our most popular experiments end up generating traffic for those that are less popular.

In our recruitment, we focused on encouraging participants to advertise LabintheWild to their friends. We also offer participants the option of following LabintheWild's Facebook page

or signing up for email announcements, but we made these opportunities less prominent. Our decision was motivated by the fact that new studies appear on LabintheWild every two months on average, which provides little opportunity for nurturing frequent return visitors.

Study design

To ensure that participation does not become tedious or exhausting, we design our experiments to take 5–15 minutes. Although it might be feasible to conduct longer studies, a duration of around 15 minutes has proven to successfully engage participants on TestMyBrain.org [13]. We also clearly communicate progress through the experiment, use friendly, casual language, and occasionally include humorous interludes.

To ensure honest answers to our demographic questions, we chose to make the questions optional wherever possible. We also provide easy and non-judgmental mechanisms for reporting situations that might have compromised the data quality at the end of each experiment. Specific questions depend on the experiment, but we frequently ask participants if they experienced technical difficulties or distractions and whether they cheated in any way. As we discuss later, participants often provide an informative explanation of how they were distracted or how they cheated.

Choice of studies

Our decision to limit the length of our studies imposes constraints on the experiment design. Studies that require showing large numbers of stimuli are shortened by presenting participants with different randomized subsamples of the full set of stimuli, and we account for the resulting differences in sample frequencies in the analyses.

The unsupervised online environment and the social recruitment mechanism additionally limit the type of studies that can be feasibly run on LabintheWild. For example, we have not attempted to run studies that require participants to be in specific environments or use specific devices. Studies that require deception (i.e., to avoid that demand characteristics [29] influence participants' behavior) could also impact the results in the longer term, because the debriefing content at the end of the study might be shared with others.

LARGE SCALE

Conducting experiments with uncompensated and self-selected online samples at large scale naturally requires a self-sustaining recruitment mechanism. This section first reports on the current scale of LabintheWild, that is, its visitor and participant numbers, before describing how people hear about LabintheWild and why they participate. Information about visits to LabintheWild is based on data from Google Analytics, while the participation data is based on data logged on LabintheWild.

Visitors and Participant Numbers on LabintheWild

Between launching LabintheWild in July 2012 and April 2014, the platform has been visited 2,072,384 times, 11.6% of which were return visits. Visitors came from 219 countries and regions, with the plurality of them coming from the United States (36.66%, see Table 1 for a list of countries).

Country	# visitors	% of total	Country	# visitors	% of total
United States	759,663	36.66	Israel	9,459	0.46
United Kingdom	312,135	15.06	Italy	9,026	0.44
Hungary	159,431	7.69	Spain	7,902	0.38
Canada	96,446	4.65	Switzerland	7,461	0.36
Romania	87,074	4.20	Mexico	7,337	0.35
Lithuania	78,349	3.78	Denmark	6,696	0.32
Australia	45,107	2.18	Poland	6,364	0.31
Germany	43,358	2.09	Greece	6,236	0.30
Norway	36,168	1.75	Malaysia	6,090	0.29
Singapore	28,940	1.40	Philippines	4,753	0.23
China	28,938	1.40	Hong Kong	4,436	0.21
Netherlands	28,424	1.37	Bosnia and Herzegovina	4,154	0.20
Macedonia (FYROM)	25,072	1.21	Russia	3,591	0.17
Finland	23,944	1.16	Turkey	3,430	0.17
Chile	19,082	0.92	Argentina	3,373	0.16
India	16,108	0.78	South Africa	3,239	0.16
Austria	14,544	0.70	South Korea	3,234	0.16
France	14,466	0.70	Slovakia	3,107	0.15
Sweden	13,387	0.65	Bulgaria	2,897	0.14
Ireland	13,282	0.64	Pakistan	2,742	0.13
Japan	12,279	0.59	Portugal	2,681	0.13
New Zealand	11,544	0.56	United Arab Emirates	2,580	0.12
Belgium	11,533	0.56	Thailand	2,576	0.12
Serbia	11,293	0.54	Taiwan	2,545	0.12
Brazil	10,132	0.49	Vietnam	2,317	0.11

Table 1. The top 50 countries (of more than 200) with the largest numbers of visitors on LabintheWild between June 2012 and April 2014.

Referral source	# visitors	% of total	Referral source	# visitors	% of total
facebook.com	408,162	19.70	mokslas.delfi.lt	8,364	0.40
m.facebook.com	142,343	6.87	gizmodo.co.uk	8,349	0.40
huffingtonpost.co.uk	114,424	5.52	krone.at	7,367	0.36
dailymail.co.uk	83,460	4.03	neogaf.com	6,556	0.32
eduline.hu	77,120	3.72	komando.com	5,602	0.27
t.co	57,555	2.78	yoda.ro	5,019	0.24
realitatea.net	45,089	2.18	rtv.net	4,838	0.23
tumblr.com	24,347	1.17	wowbiz.ro	4,649	0.22
hvg.hu	20,467	0.99	plus.url.google.com	4,470	0.22
origo.hu	19,223	0.93	boards.4chan.org	4,210	0.20
reddit.com	18,792	0.91	laikas.lt	3,632	0.18
alfa.lt	15,123	0.73	updateordie.com	3,491	0.17
szeretlekmagyarorszag.hu	14,800	0.71	taringa.net	3,482	0.17
forum.bodybuilding.com	14,110	0.68	telegraf.rs	3,376	0.16
technologijos.lt	13,782	0.67	h2w.iask.cn	3,137	0.15
weibo.com	13,007	0.63	i-am-bored.com	3,136	0.15
crnobelo.com	11,011	0.53	tigerdroppings.com	3,129	0.15
side3.no	10,932	0.53	smallbusiness.yahoo.com	2,995	0.14
faculteti.mk	10,390	0.50	quo.mx	2,269	0.11
sociedad.biobiochile.cl	9,552	0.46	pcwelt.de	2,048	0.10

Table 2. List of predominant referral sources to the site between June 2012 and April 2014.

During the July 2012 – April 2014 period eleven distinct experiments were available on LabintheWild at some point. Visitors completed 744,739 experimental sessions. This number only includes those participants who have finished the whole length of an experiment and did not report to have taken the same experiment before. Because we do not track participants across experiments, we do not know how many unique participants completed experiments on the site. An analysis of visitor flow suggests that very few participants completed more than two experiments in a single visit.

How People Hear About LabintheWild

Table 2 lists the predominant referral sources to LabintheWild over the past two years. Although Facebook is the primary source of traffic, more than 5,000 websites currently mention LabintheWild or one of its experiments and lead visitors to the site.[1] The fact that people link to the site is also no-

[1]This data reports on statistics provided by Google Analytics and reports on visitors, not participants.

ticeable in that even after removal of an experiment from the LabintheWild front page, many people still access the experiment directly. The diversity of referral sources additionally conveys that LabintheWild attracts people with a variety of interests.

To find out why people invite others to participate, we analyzed the 200 most recent tweets that included LabintheWild's URL or referred to the site name. Following the thematic analysis method [14], we first added codes to the tweets that described the reasoning behind encouraging others to visit LabintheWild. These codes were then iteratively clustered into themes. We saw two patterns emerge: First, people appear to share their interest in science for altruistic reasons, such as exemplified with these tweets:

> *Participate in these #studies because #science is cool! And it's an interesting insight into yourself.*
>
> *Intriguing research trying to quantify difficult-to-quantify things. Add to their data.*
>
> *LabintheWild.org Help advance the design of all sorts of cool things. Older people especially welcome.*

The last tweet referred to a request that we added at the end of an experiment to tell not just "your friends" to participate, but especially people over 50. After adding this sentence, the average age of participants noticeably increased.

A second reason for sharing the experiments is to communicate the discovery of something that could potentially be exciting for others:

> *Savvy web users - do you think you can spot an untrustworthy website? Take this test*
>
> *take these tests and learn that you're older and more japanese than you ever knew: LabintheWild.org*
>
> *Check out LabintheWild.org for some pretty neat cultural / perception experiments [...]*
>
> *These tests have been amusing: LabintheWild.org , particularly the social intelligence one. Let me know if you try any of them!*
>
> *take some fun tests! one of them will guess your age by how you click on the red dot.*

The tweets demonstrate the role that online social media users play in recruiting others and spreading the word.

Why People Come to LabintheWild

To explore the reasons for participation, we again analyzed tweets that contained references to LabintheWild. In addition, we also included comments that participants provided at the end of four experiments that represented a diverse set of study and feedback designs. The resulting set contained nearly 30,000 comments. We first excluded one-word comments (e.g., "interesting", "fascinating", "boring", "no") that did not contain any reasoning. We then thematically analyzed the data [14] by labeling the remaining comments and tweets with keywords, and clustering them until no more additional categories emerged.

The analysis indicates that people come to LabintheWild for diverse reasons, but that the main categories are (1) an urge to compare themselves to others, (2) a general curiosity about themselves, (3) a fascination with the idea that their own characteristics could be predicted, and (4) the desire for improving particular skills. In the following, comments (but not tweets) will be reported along with the test and participant number that they refer to.

Consistent with the social comparison theory [12], many tweets sent after participation showed that people enjoyed the comparative feedback that LabintheWild experiments provide. For example, some Twitter users wrote

> *I got 31/36 (above average).... not bad for a confirmed social nincompoop! RT Test your social intelligence*
>
> *Uh oh. Only 3% of people (roughly 0 / 10) from United Kingdom share your visual preferences!*
>
> *I challenge you all to beat my score of 35 out of 36! Post your results here.*

Sharing results online seems to be a way of differentiating oneself from others and showing off these differences.

Other tweets suggest that participants appreciate discovering something new about themselves:

> *Fascinating. Apparently I'm better at perceiving the background. Psychology tests are awesome.*
>
> *LabintheWild.org I love this ..according to this I'm more Japanese than American..& I'm 30!! Yipee!!*

Similarly, participants' comments describe that the studies made them aware of their own preferences:

> *love the test. It makes you think about design trends, simplicity in web design and color usage. I rated some websites lower just because I didn't like the color combination!* [P42367, aesthetics test]
>
> *Roughly around the break it started to become clear to me that I like quite simple pages. At least when it comes to colors. But even more important than that seemed to be the layout of the page. If the main thing was big and clear, maybe accompanied with a calm picture, it appealed to me more than the over simplistic ones or the ones that were just simply too full of <beep>.* [P459, aesthetics test]

Some tweets show that participants were surprised about the accuracy of results, and fascinated that the studies could produce unexpected predictions about themselves.

> *So, er, LabintheWild.org guessed my age exactly! Science is cool.*
>
> *My score: Low Colorfulness, High Complexity, Medium Color Sat.: Visual Preferences Test - yes, I like neutral colors.*
>
> *okay this is sort of spooky. It makes me a little more afraid of smart people :) It guessed me at 29 (i'm 30)...*

Comments additionally suggest that for some participants improving particular skills is the main motivation for participating:

> *Very interesting! It would be nice to have the results after the test, I'd be curious to know if I rated the designs consistently, and also to take a better look at which ones I've rated high or low* [aesthetics test, P43024]

> *It would be great [...] to see which eyes we guessed wrong. I personally took the test hoping to practice or improve my skills; but as it stands I really don't take anything out of it.* [P872, social intelligence test]

> *Thanks! I hope to use this to improve my business skills.* [P15464, social intelligence test]

In some cases, revealing the correct answers so that participants can see what exactly they got wrong risks the validity of the experiment if the answers get shared or discussed between participants. Hence, we only offer a summary of an individual's results, but provide the full answer key whenever participants ask us for it via email.

Discussion

The previous analyses demonstrate the feasibility of conducting uncompensated online experiments with large numbers of participants (an average of a thousand participants a day), allowing us to study more participants at lower cost than possible in lab. People's comments and tweets show that LabintheWild participants play an essential role in enabling this scale by recruiting others.

A noteworthy aspect of LabintheWild is that 88.4% of its visitors are new to the site. Hence, in contrast to Mechanical Turk, LabintheWild largely taps into new participant pools rather than a stable base of returning participants. There are two reasons for this: First, LabintheWild does not require visitors to sign up before engaging with the content, and thus, lowers the barrier for testing it out. The second, and more important, reason is that LabintheWild does not offer new content at the same frequency as Mechanical Turk. Because it takes us two months on average to add a new experiment to LabintheWild, we made little effort to encourage participants to return other than giving them the opportunity to follow the LabintheWild Facebook page where we announce new experiments. However, we do encourage the recruitment of others by including social network sharing buttons in the results pages of our experiments.

LESS WEIRD

In contrast to the convenience samples of many in-lab experiments (most often North American undergraduate students [17]), one of LabintheWild's core goals is to conduct experiments with less WEIRD participants. In particular, our aim is to reach a broader range of age groups, countries, and education levels.

Table 3 shows participants' self-reported demographics for nine of our previous experiments. These experiments were on the LabintheWild front page for varying lengths in time (usually, LabintheWild features five experiments at the same time), but most of them remained accessible through the study link and had ongoing, albeit lower, traffic afterwards. The number of participants includes only those who did not report to have participated in the study before. The percentage of US participants represents people who are currently living in the US, but are not necessarily originally from the US.

Across experiments, approximately 49% of all participants are female. This is slightly more balanced than the 55% females found in Mechanical Turk samples [26], and shows that females as much as males are attracted to participate in LabintheWild experiments.

LabintheWild participants have a mean age of 29 years (median 26, range 5-99), which is older than the average age in laboratory studies, but younger than the mean age of 32 years (median 30) that has been reported for Turkers (see e.g., [26]). LabintheWild reaches a larger age range of people than Mechanical Turk does ([26] report an age range of 18 to approximately 75), mainly due to the fact that Turkers are required to be at least 18 years before being able to sign up.

About 73% of LabintheWild participants report to be currently enrolled in college or have a college degree or higher, which suggests that our participant samples are not representative of the general population. The higher education level is consistent with that of the Mechanical Turk population [30]. In fact, Paoloacci et al. suggest that the higher education level of Turkers could be typical among early adopters of technology [30]. Importantly, roughly 27% of our participants have a lower education level than the traditional student samples recruited for in-lab experiments.

Participants self-reported to be from more than 200 countries and regions on six continents, which is in line with the Google Analytics report for LabintheWild visitors (see also Table 1). Moreover, approximately 66% of participants across all experiments (the majority in all except one of our previous studies) comes from countries other than the US. In contrast, Mechanical Turk surveys suggest that Turkers come from 190 countries, but 81% of them are from India and the US [30]. This makes the overall LabintheWild sample less WEIRD than the typical Western convenience samples in lab experiments, and it also demonstrates that LabintheWild enables us to reach more geographically diverse participants in higher numbers than currently possible using Mechanical Turk. We expect this diversity of LabintheWild to increase after the site and experiments are offered in languages other than English.

Discussion

The previous analysis of demographics demonstrates that LabintheWild participants are arguably less WEIRD than common convenience samples of laboratory studies (see, e.g., [3, 17]). LabintheWild is also more diverse than Mechanical Turk in terms of a larger age range, and a larger number of countries that contribute substantial numbers of participants. This opens up new possibilities for cross-country comparisons and cross-cultural research. Likewise, more diverse populations within countries enable comparisons between several demographic groups. In the past, the diversity among LabintheWild participants has allowed us to empiri-

Experiment	Abbreviation	# months on front page	# participants	% female	age range	mean age (median)	stdev age	# countries	% US participants	% college or above	# of native languages	% native English speakers
What is your website aesthetic? (Experiment 1)	aesthetics	21	42,171	52.7	6-99	32 (29)	12.83	190	41.4	73.5	38	62.6
How fast is your memory? (Experiment 2)	memory	4	1,121	N/A	13-99	26 (23)	11.53	81	28.8	62.2	36	49.6
Test your social intelligence! (Experiment 3)	social intelligence	10	125,570	48.5	12-98	30 (26)	12.24	227	46.9	N/A	N/A	N/A
What do you perceive as colorful?	colorfulness	3	8,600	56.9	7-99	29 (26)	12.37	140	40.6	72.5	38	57.4
Are you more Eastern or Western?	frame-line	21	7,623	57.7	6-99	27 (24)	11.58	148	47.3	74.1	37	70.7
How do you predict changes in future trends?	graph prediction	5	1,216	45.3	6-99	26 (23)	12.95	86	40.8	73.1	37	59.3
What do you perceive as complex?	complexity	1	176	46.6	12-70	33 (30)	12.19	32	65.9	90.3	23	61.9
Trust us. You will love this test!	trust	7	1,944	56.9	6-99	27 (23)	12.63	97	43.4	69.5	35	66.4
Can we guess your age?	age guessing	4	556,330	39	5-99	32 (29)	12.09	139	30.7	N/A	N/A	66.4

Table 3. Demographic composition of participant samples from nine LabintheWild experiments. Two further experiments are not shown here, because they were only online for brief periods of time. The first three experiments are presented in this paper as part of the data quality analyses.

cally demonstrate differences between the visual preferences of people from more than 40 countries and various demographic groups [32].

The finding that 73% of LabintheWild participants have received or are currently pursuing a college degree or higher is not surprising given our recruitment methods with our own social networks as a starting point. We believe that a key to increasing the diversity will be to further explore how exactly different segments of the population hear about LabintheWild, and what motivates them to take part.

EQUAL DATA QUALITY

The third and perhaps most important goal of LabintheWild was to maintain the data quality on par with traditional in-lab experiments with controlled conditions and experimenter supervision. To evaluate this, the following sections report on LabintheWild replications of three experiments from the literature, and show how participants play an important role in detecting issues that could compromise the data quality.

Experiment Replications

We implemented online versions of the following three experiments from the literature:

1. An experiment on subjective ratings on appeal [25].
2. A study of working memory processing speed [35].
3. A test assessing social intelligence, originally called "Reading the mind in the eyes" [2].

In the first experiment, participants are asked to provide subjective ratings of the aesthetic appeal of websites. They rate each website twice, which enables an analysis of reliability. We include this test because soliciting truthful and reliable subjective responses is still considered challenging in unsupervised online studies [22, 28, 34] making this test a particularly sensitive probe of the truthfulness of data reported by LabintheWild participants: are they reporting their actual subjective reactions at least as reliably as in-lab participants or are they prone to "spamming" as some participants on Amazon Mechanical Turk? The successful completion of this study also requires careful attention as stimuli are presented for only 500ms each, giving us insights into whether LabintheWild participants get distracted substantially more than in-lab participants.

The second experiment on working memory processing speed involves working under considerable cognitive load. Hence, this test provides an insight into whether or not LabintheWild participants are willing to put in substantial mental effort into completing experiments.

In some experiments, as in the test measuring the working memory processing speed, the overall phenomenon is reproducible across laboratories, but the exact values of the measurements differ substantially across experimental settings. We included a third test, the "Reading the mind in the eyes" study, because its measurements should be exactly replicable *independent* of the experimental setting. The scores obtained in this experiment depend on both actual ability and effort. If LabintheWild participants achieved scores similar to those obtained by supervised in-lab participants, it would provide evidence that they exerted as much effort.

Experiment 1: Website Aesthetics

The first experiment that we replicated was originally presented by Lindgaard and colleagues in 2006 [25]. The authors established that people are able to make reliable judgments on the visual appeal of websites after a short stimulus exposure time of 500ms.

The experiment was divided into two blocks, in which participants were shown the same set of websites (in different randomized order) for 500ms each. After each website, they were asked to rate the visual appeal of the site on a single-item 9-point Likert scale. The results showed that the ratings of the same websites in the two phases are significantly correlated, suggesting a high intra-participant reliability.

In our replication, we used the intra-participant reliability to verify the effort and truthfulness of participants' subjective visual preference ratings. If participants give the study their undivided attention and provide ratings that truly correspond to their visual preferences, the ratings between the two phases should be highly correlated.

Main Results From Prior Studies

In the original study [25], the authors tested the intra-participant reliability of 20 participants using Pearson correlations for each participant's rating in the first and the second phase. All correlations were significant at a $\alpha < .05$ level and all correlation coefficients were above r=.60 (see Table 4).

We closely followed the design of the third experiment presented in [25], replicating the condition in which participants were shown website stimuli with a 500ms exposure time. Participants were first presented with an informed consent form, a demographics questionnaire, as well as a screen containing

instructions on the task. The instructions also emphasized the short stimulus exposure time, which would require extra attention. Participants were then asked to rate websites on visual appeal on a single-item 9-point Likert scale. Website stimuli were pre-loaded in the background to ensure that the exposure time was unaffected by varying Internet connection speeds. The experiment was divided into three parts, starting with a practice session, and two main experiment parts, which included the same website screenshots in (differently) randomized order.

To keep the experiment under 10 minutes, we used 5 practice website screenshots (as opposed to 20 in the original study), and 30 websites per test phase (as opposed to 50). We added text and a picture between the two test phases to divert participants' attention from the main purpose of the test and encourage them to take a break. The study ended with personalized feedback about a participant's visual preferences in comparison to other people from their country of current residence.

Analysis

Following the procedure described in [25], we conducted separate Pearson product-moment correlations for each participant. Lindgaard and colleagues only provided an aggregation of their correlation coefficients, and so we did not perform a comparative statistical analysis of the distributions.

Participants

For an overview of participant demographics see Table 3. Because the original laboratory study was conducted at a North American university and required participants to speak English as their first language, we only included 10,976 LabintheWild participants who had always lived in Canada or the US, and who reported English as their native language. The demographic composition changed only slightly: Participants had a mean age of 33 (stdev = 13.24, median 29), and 59% were female. The education level increased to 79% who reported pursuing college or above.

Adjustments of Data

We excluded 166 participants for whom our system reported that the stimulus was not displayed for 500ms, and 75 participants who reported having experienced technical difficulties, and/or having cheated. Together this accounted for 2.2% of the 10,976 participants, resulting in a data set of 10,735 participants.

Correlation coefficient	Laboratory study (N=20)		LabintheWild, 18-24 years old (N=199)		LabintheWild, all (N=10,735)	
	N	%	N	%	N	%
.0-.09					3	0.03
.10-.19					9	0.08
.20-.29					21	0.20
.30-.39			1	0.5	67	0.62
.40.49			1	0.5	212	1.97
.50-.59			4	2.01	539	5.02
.60-.69	1	5	6	3.02	1338	12.46
.70-.79	4	20	16	8.04	3055	28.46
.80-.89	15	75	97	48.74	4469	41.63
.90-.99			74	37.19	1022	9.52

Table 4. Overview of the distribution of correlation coefficients comparing the results of the laboratory study to the one conducted on LabintheWild. LabintheWild participants were only included if they were from North-America and reported being a native English speaker.

Main Findings

Table 4 provides an overview of the correlation coefficients from the in-lab study and those from LabintheWild. Because participants in the original laboratory study were students at a Canadian university (participants' age was not reported in [25]), we first analyzed how the results of LabintheWild participants aged 18-24 years (N=199) compared against the in-lab results. All correlations on LabintheWild were significant at a $\alpha < .05$ level. In comparison to Lindgaard et al.'s study, a higher percentage of participants (85.93% versus 75% in the laboratory study) achieved correlation coefficients above $r = .80$. This suggests that LabintheWild participants at comparable age to the in-lab study participants were at least as consistent in their responses than those in the lab study.

When including all LabintheWild participants (i.e., those from the US and Canada who reported speaking English as their first language), we see a wider distribution of coefficients. We found that 84 correlations (0.78% of the total) were not significant at a $\alpha < .05$ level, indicating that a small percentage of LabintheWild participants might have been distracted or randomly rated the websites. The remaining 99.22% of correlations were significant, and 92.07% of correlations were above $r = .60$; the vast majority of LabintheWild participants were as or even more consistent in their responses as participants in lab.

We performed an additional analysis across the entire data set of 42,171 participants from 190 countries, again excluding participants with technical difficulties and/or those who reported having cheated. The excluded data accounted for 2.3%, resulting in 41,201 participants. We observed no substantial relationship between country of origin and the results, indicating that the phenomenon of people forming lasting impressions of aesthetic appeal based on brief exposure is universal across the sample population of our dataset.

In summary, our overall results match Lindgaard et al.'s finding that participants are able to form consistent impressions of a website's appeal after seeing it for only 500ms. Moreover, our results indicate that this phenomenon generalizes across a variety of demographic and geographic groups.

Experiment 2: Working Memory Processing Speed

In 1966, Saul Sternberg published results of an experiment demonstrating that the time needed to retrieve an item from working memory was linearly proportional to the number of items stored there [35].

Tasks

This study has been replicated and extended by numerous researchers (e.g., [6, 7, 8, 9, 36, 37]). All variants of this study have a common basic structure: In each experimental block, participants are first presented with a sequence of several (typically 1–6) symbols to memorize (Figure 2 left). After that, they are presented with one symbol at a time (a probe) and asked whether the symbol was present in the original set (Figure 2 right).

We based our main design decisions on the original study [35] (some of the details were not revealed in the original paper,

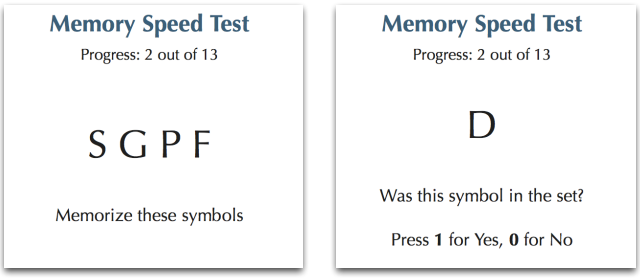

Figure 2. An example task in the Working Memory Processing Speed experiment.

but were made apparent in a later publication [37]). Each experimental block presented participants with a set of 1–6 randomly chosen symbols and 11 probes. Three of the probes (i.e., 27%) were positive (that is, contained a symbol from the original set) and 8 were negative. The order of the positive and negative probes was random and so was the selection of the specific symbols to display.

Unlike in the original study, we used the combined set of digits and uppercase English alphabet letters (rather than just digits). The choice of symbol does not affect the main outcome of the study [5]. We removed digit-letter pairs that could be visually confused for each other (i.e., 1-I, 0-O). The resulting set contained 32 distinct symbols.

Main Results From Prior Studies

The linear relationship between the size of the symbol set and the reaction time has been reliably confirmed and researchers repeatedly reported r^2 values of .99 and higher [35, 8]. This is the primary finding that we aimed to replicate.

Another prominent property of this task is that the marginal response time per item does not change with practice: That is, the absolute response time tends to improve as participants gain practice with the task, but the *slope* of the line of fit, which captures how much additional time is needed as the size of the set held in working memory increases by one item, does not change with practice [24, 27]. This is the second finding we aimed to replicate.

Detecting the relationship between set size and the response time rests on the assumption that participants actually make the effort to memorize the items and to respond correctly. All replications of this experiment exclude participants whose overall accuracy fails to meet an accuracy criterion (typically a number over 95%). A recent article provided the distribution of participant accuracies prior to any exclusions showing that lab-based participants accurately recalled items as often as 98% on average when there was only one item to be remembered and 90% on average when they had to memorize six items [8]. We compare these results to the accuracies of LabintheWild participants.

Procedure

The first three screens of the study presented participants with the basic information about the study, the informed consent form, and a short demographic questionnaire, which also asked whether participants had taken the study before. All questions were optional. Next, participants were presented with brief instructions followed by a single practice block of 11 trials. During the practice block participants were given feedback about the correctness of their responses immediately after each trial. No such feedback was provided during the subsequent blocks. The main experiment consisted of 12 blocks (2 for each set size between 1 and 6), each with 11 trials. Participants could take breaks between blocks.

The feedback page showed each participant their mean accuracy (what percentage of the probes they gave correct responses to) and their mean response time in comparison to previous study participants.

Measures and Analysis

We first conducted a least squares linear regression with *set size* [1–6 items] as the only factor and the *response time* as the response variable. As in the original study, we first averaged response times over all non-excluded blocks separately for each set size.

Next, we conducted an analysis of variance meant to uncover learning effects. This was a between-subjects analysis with the following factors and levels:

- *set size* [1–6 items] (modeled as a continuous variable)
- *prior exposure* {first time, repeat} (modeled as a categorical variable)
- *age bin* {10–14, 15–19, 20–24, 25–29, 30–39, 40–49, 50+} (modeled as an ordinal variable)

The *response time* measured participants' reaction time.

To analyze participant accuracies, we computed the mean percentage of correct responses for each of the six set sizes. Because the results in prior work were only reported graphically (i.e., they were summarized visually in a chart, but neither mean values nor variances were explicitly stated in the article), we did not attempt a statistical comparison.

Participants

The experiment was completed 1,319 times (198 repeat participants). The demographics of 1,121 participants who took this test for the first time can be seen in Table 3.

Adjustments of Data

We removed 100 experiment instances, in which participants reported having experienced technical difficulties, and/or cheated. These exclusions accounted for 7.6% of the 1,319 completed experiments leaving 1,219.

As in the original experiment, we removed experimental instances in which accuracy (i.e., the fraction of trials in which participants correctly recalled whether a probe was in the original set) did not meet the requirement. In the original study, three out of eleven participants (27%) were excluded and the mean accuracy of the participants retained for analysis was 98%. The accuracy threshold for exclusion was not reported. We excluded experimental sessions in which accuracy was lower than 94% (N=103, or 8.4%). We found that setting the threshold any higher did not impact the results. We

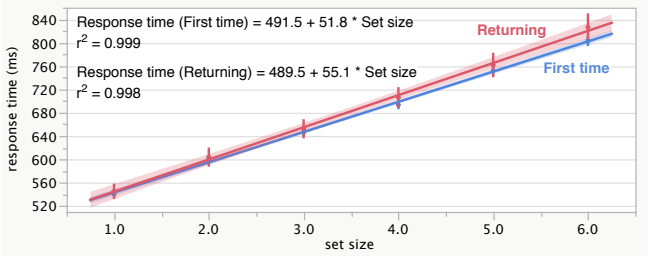

Figure 3. Main results for the Working Memory Processing Speed test: Results show a strong linear relationship between the number of items held in working memory and response time. Error bars show 95% confidence intervals. There is no significant difference in slope between first time and returning participants.

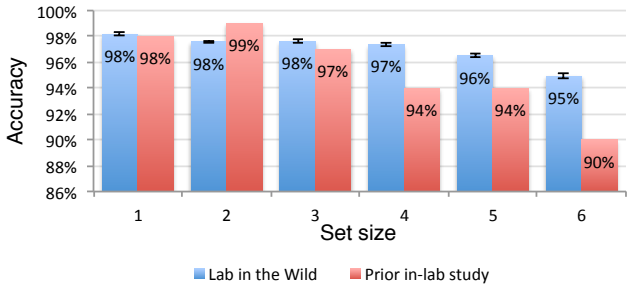

Figure 4. Accuracy comparison between LabintheWild and a prior laboratory study with 36 university students. Participants on LabintheWild were as likely to correctly recall what symbols were shown to them as participants in a traditional laboratory experiment. Error bars show the standard errors (available for LabintheWild participants only).

thus retained 1116 completed experimental sessions for analysis, 951 of which were completed by first time participants and 165 by returning participants.

In the original study, the first three trials in each block were excluded. We only removed the first trial because we found no substantial differences in mean or standard deviation in the response times for any of the subsequent probes.

To identify participants who might have gotten distracted during the test, we looked for extreme outliers in response time. Specifically, we computed the median and interquartile range (IQR) for the response times for set size of 6 (the condition with the slowest response times and largest variance). Similarly to prior studies with unsupervised online participants [23], we computed median + 3×IQR (2028ms) as the cut-off. 1.7% of the trials were thus identified as extreme outliers. Some took as long as 1.5 minutes indicating a major distraction. Because such a distraction was likely to have caused the participants to forget the symbols they were meant to memorize, we excluded entire blocks that contained at least one extreme outlier trial. This resulted in 15.2% of the blocks being excluded from analysis.

Main Findings

The main results are summarized in Figure 3. For first time participants, we observed a linear relationship between the number of items held in working memory and the response time: the response time increased by 51.8ms (±2.0ms, 95% confidence interval) for each additional item stored in working memory. This value is consistent with prior work and the model fit ($r^2 = 0.998$) matched or exceeded fits observed in prior work.

For people who reported having taken the test before, the results were similar: their response time increased by 55.2 (±2.2)ms for each additional item stored in working memory. The model fit is similar as for first time participants: $r^2 = 0.998$.

Comparing performance of first time participants to the returning ones, we observed no significant main effect of prior exposure to the experiment ($F_{1,9667} = 0.01, n.s.$). There was also no significant interaction between prior exposure and set size ($F_{1,9667} = 0.60, n.s.$). This last result indicates that — consistent with prior results — we observe no learning effect on the rate with which response time changes with the number of items held in working memory.

Accuracy

Figure 4 illustrates the performance of all first time LabintheWild participants (only excluding those who reported technical difficulties or cheating) compared to the accuracy of 36 students participating in a recent laboratory replication of this study [8]. Because of a lack of information about variance in the prior study, we did not test the significance of these differences, but the comparison does suggest that the unsupervised LabintheWild participants are attending to the task at least as well as in-lab participants.

In summary, we have reproduced two prominent results associated with this experimental paradigm: a strong linear relationship between set size and the response time and lack of significant learning effects. Our results also suggest that participants on LabintheWild are at least as accurate in recalling the items as participants from a conventional in-lab study.

Experiment 3: Reading the Mind in the Eyes

Our third experiment replicates the "Reading the mind in the eyes" test [2]. Participants were presented with 36 images (plus one practice image), each showing a portion of a person's face that included the eyes, but not the nose or the mouth. For each such image, participants were shown four words describing emotions and asked to choose one that was most likely being expressed by the person in the image. We used identical images in the same order as the original study [2].

The test was initially developed in the context of the study of Autism and the results demonstrated that, on average, people on the Autism spectrum are less likely to recognize a person's emotion just by looking at their eyes than controls drawn from the general population [2]. Later investigations in the area of collective intelligence demonstrated that the average performance of group members on this test was a better predictor of the group's success on a difficult problem solving task than the group's mean IQ [39]. Because of this association, we advertised it as a test of social intelligence on LabintheWild.

The original publication [2] included results for British students and for the general British population. For both popu-

	Original study						Lab in the Wild				
			Age (years)		Gender (%)			Age (years)		Gender (%)	
	N	age data available for	Mean	Stdev	Female	Male	N	Mean	Stdev	Female	Male
General British Population	122	88	46.5	16.9	55%	45%	1973	46.3	8.0	59%	41%
British student-aged participants	103	103	20.8	0.8	49%	51%	1519	20.4	1.1	30%	70%

Table 5. Reading the Mind in the Eyes study (social intelligence): Participants from the original study and a matching subset of participants from LabintheWild.

lations, results were reported separately for men and women. Arguably, the performance on this test is dependent on both ability and effort. Equal or higher performance for matching populations on LabintheWild would provide evidence that LabintheWild participants exerted as much effort as participants who performed the study in lab.

Procedure

The first three pages of the study presented participants with a brief description of the study and its duration, an informed consent page, and instructions on the task. They were given one practice trial for which they received feedback about the correctness of their response. Participants then performed 36 experimental trials. No feedback about the correctness of those responses was provided until the very end of the study.

After completing the experimental trials, participants were presented with a voluntary demographic questionnaire, which also asked them if they were native speakers of English, and, if not, whether they had any difficulty understanding the words describing the emotions.

Participants were then presented with the feedback page showing their own score (i.e., the number of correct answers) in comparison to the average score of 26 as reported for the adult population reported in the original study [2].

Participants

The experiment was completed by 131,785 first-time participants. To match the populations used in the original study (British students and general British adult population), we selected participants who reported currently living in the United Kingdom and being native speakers of English, and excluded participants who reported having an impairment affecting their ability to use a computer. To match the age distribution of the student population, we selected participants aged 19–22 and to match the general adult population we selected participants aged 36 through 75. Table 5 summarizes the demographics of the participants in both the original study and in our analysis.

Measures and Analysis

The main dependent variable was the score, computed as the number of images for which a correct emotion was given. For each sub-population included in the study, we first used an F-test to test for significant differences in variance in scores. Because no such differences were found, we compared mean scores using a t-test for independent samples with equal variances. For both analyses, we applied Bonferroni correction to account for multiple hypotheses being tested.

		Original study			Lab in the Wild			F-test		t-test	
		N	Score	Stdev	N	Score	Stdev	F	p	t	p
General British population	male	55	26.0	4.2	817	27.2	3.8	0.83	ns	2.31	ns
	female	67	26.4	3.2	1156	27.6	3.8	1.42	ns	2.48	ns
British student-aged participants	male	53	27.3	3.7	1064	27.3	3.9	1.14	ns	0.08	ns
	female	50	28.6	3.2	455	27.6	3.8	1.39	ns	-1.77	ns

Table 6. Results from the "Reading the mind in the eyes" study. Scores were computed as the number of trials for which the correct answer was given. There were no significant differences in the mean scores or variances in scores between in-lab participants and the matching LabintheWild participants.

Results

Table 6 contrasts results from the original study with those obtained on LabintheWild. While adult British participants had slightly higher scores on LabintheWild than in lab and while British student-aged female participants had slightly higher scores in lab than on LabintheWild, none of these differences were statistically significant. There were also no statistically significant differences in the variances across the two experimental settings for any of the four populations. In summary, the results suggest that LabintheWild participants put in as much effort as those in the laboratory study.

We conducted an additional analysis of variance for all participants who reported being native speakers of English, who were at least 11 years old, who provided gender information, and who were currently living in the same country in which they grew up. We excluded participants from countries for which fewer than 100 such participants were available. In the end, 59,934 participants from 10 countries were included in this analysis. After controlling for age and gender, we observed a significant effect of country ($F_{9,59917} = 8.9, p < 0.0001$). Post hoc Tukey HSD test revealed that this effect can be explained by scores from India and Malaysia being significantly lower than those from Australia, Singapore, New Zealand, United States, United Kingdom and Canada. Given these results, we find it likely that the test is sensitive to participants' prior cultural exposure (the way emotions are communicated in media and in person; exposure to facial features of a particular ethnic group) and that it may not be robust for comparing participants from substantially different cultures.

In summary, our results showed no significant differences in mean scores or score variance between the participants from the original in-lab experiment and the equivalent sample of LabintheWild participants. However, an analysis of a broader sample of native English speakers indicated a significant effect of country, suggesting that the design of this test may be culture specific.

Participants' Role in Ensuring Data Quality

The results of the three experiment replications show that the data obtained in online studies with uncompensated study participants can be as reliable as data collected in lab. However, the uncontrolled environment online means that participants might have technical difficulties, they might get distracted, or they might deliberately try to "cheat" to get a better result. We found that an essential part of LabintheWild studies is asking participants to leave open-ended comments at the end of an experiment to report on possible distractions, technical difficulties, or cheating. About 5% of participants

(depending on the experiment) leave feedback in these comment boxes. We learned from the comments that participants are taking the tests in various situations and locations, and that this often leads to distractions or interruptions by other humans or even pets:

> *My daughter came in and I missed one picture.* [P42515, aesthetics test]
>
> *cat sat briefly on keyboard.* [P36211, aesthetics test]

Participants' comments also made us aware of potential distractions resulting from the lack of control over other software running on their computers and disrupting the participant while taking a test:

> *I had a software update popup during one of the slides, which prevented me from viewing the image, and so I may have given an average rating, when I might have rated it otherwise.* [P4958, aesthetics test]
>
> *my friend spoke to me in a voice chat client halfway through.* [P933, memory test].

The diversity of LabintheWild participants also means that participants can be more systematically distracted or unable to perform the test at their best due to temporary or long-term conditions:

> *I suddenly lost hearing in my right ear during the first set of images. As a result I was a bit disoriented and confused for a little while and this may have affected my ability to accurately appraise.* [P35721, aesthetics test]
>
> *epilepsy* [P27753, aesthetics test]
>
> *I was really, REALLY drunk when I did this... and I don't in any way believe it is reflective of my true memory capacity.* [P1545, memory test]

Some comments additionally provide more context explaining why their data might not be good for inclusion:

> *Not sure if I am a good subject. I do UX professionally, and was distracted by what I perceived as usable or unusable design.* [P35395, aesthetics test]
>
> *I'm in a bit of a mood and probably rated a few things one or two lower than i normally would, but all of the ones i put as ones were pretty solid.* [P12134, aesthetics test]
>
> *Difficult to press no with my right hand. I'm right-handed and it just felt the wrong way around.* [P1212, memory test]
>
> *Please note that I have my screen set to Black & White, which is how I normally view the internet.* [P36395, aesthetics test]

Participants also reveal strategies that they apply to improve their performance. In response to the experiment testing the working memory processing speed, participants revealed how they supported their memory:

> *I felt it didn't test my memory so much as my ability to arrange the characters quickly into a memorable string (I put them in alphabetical then numerical order, e.g. AUVW67). Then I could just chant the string to myself constantly. Oh dear, was that cheating?* [P2257, memory test]
>
> *I don't know if this is cheating, but I would repeat it out loud to myself.* [P1515, memory test].

Apparently, these participants were unsure whether a particular strategy is "allowed", suggesting that instructions need to precisely address expectations on participants' behavior while taking part in an online experiment.

Participants also commented that they knowingly cheated or provided untruthful information. For example, in our aesthetics experiment with short stimulus exposure times, two participants reported employing strategies to work around the short timing:

> *You can click and hold the image, and put it into a new tab to view the image for a longer amount of time.* [P801, aesthetics test]
>
> *USED A PRINTSCREEN :)) FOR COUPLE OF PIC...* [P28434, aesthetics test]

Such strategies — even if not revealed by the participants — can be efficiently caught by recording the time it takes participants to provide answers and removing outliers before analysis. However, there is no straightforward way to deal with dishonest responses to demographic questions, except to rely on participants' comments. In fact, we found that even when demographic questions are voluntary, participants sometimes admit to having provided wrong information:

> *I shaved a few years from my actual age - still same decade.* [P40989, aesthetics test]
>
> *I said I was 99.* [P8508, memory test]

While participants' corrections are essential for us, we currently do not know how many participants provide wrong information or cheat without revealing it later.

Discussion

Replicating three experiments from the literature on LabintheWild, the previous sections confirm that the data collected in an unsupervised online environment without compensating participants can replicate in-lab study results. Our findings suggest that this setting enables reliable data quality for both subjective and performance-based experiments.

We also showed how providing participants with the option to leave comments leads around 5% of them to report untruthful behavior, distractions, or unusual settings and situations that might have compromised the quality of the data. Many of these comments show that participants are distracted by other humans, pets, a TV, or computer software that diverts their attention. Participants also report on health conditions that impact their performance. These comments can inform necessary changes of instructions, the experiment design, or the feedback pages, and therefore provide essential support for iteratively improving experiments.

GENERAL DISCUSSION

One of the first questions we ask ourselves before designing an experiment for LabintheWild is "What can participants learn from this?". By offering participants personalized feedback on their performance, we have been able to conduct a variety of studies that, at first, might seem to have little intrinsic appeal. For example, our study of cross-cultural differences in aesthetic appeal involved rating 60 websites, which can be a tedious and exhausting task. Yet this study attracted more than 42,000 participants who provided as reliable ratings as participants in lab. Performing a series of Fitt's law tasks (clicking on a number of dots on the screen) for five minutes, as in our study of age-related differences in motor performance, is usually perceived as monotonous and tiring, yet the experiment was completed by more than 550,000 participants. In some cases, the measurement we report back to the participants is close to the one we are interested in. For example, in the study of cross-cultural differences in aesthetic appeal, participants were able to compare their own preferences to those of other people in their country. If the measurement related to our research interest might not be exciting for participants, we find another facet of the results that might be more suitable. In the case of the study on age-related differences in motor performance, we did not compare participants' motor performance to others, but instead "guessed" participants' age based on their mouse movements and pointing behavior.

The feedback provided by the experiments also appears to be the feature that most influences people's decision to share the studies with others. Participants share their experience in their online social networks, in blog posts, or newspaper articles, with the result that LabintheWild receives traffic from more than 5,000 referral sources.

Although LabintheWild is currently only available in English, previous experiments have already attracted participant samples that are more diverse than those of conventional laboratory studies and even those on Mechanical Turk. In particular, LabintheWild participants have a wider age range and come from a larger number of countries across six continents. This has enabled us to conduct comparisons of factors influencing subjective perception of aesthetic appeal between more than 40 countries and various demographic groups in the past [32], suggesting that LabintheWild is a feasible tool for conducting cross-cultural studies.

However, LabintheWild's participant sample is still far from being perfectly representative of the broader population. Because people hear about LabintheWild through social media channels and web pages, the sample is almost certainly biased towards people who frequently use the Internet, and are most likely more WEIRD (in particular, more educated, more industrialized, and richer) than the actual world population. The comparison-based incentives might additionally attract particular kinds of participants. As the demographics data across nine experiments (Table 3) show, even different experiments on LabintheWild attract slightly different populations. Understanding why varying demographic groups are drawn to certain experiments will be part of our immediate next steps. Much of our future work will therefore focus on attracting even more diverse samples: Apart from translating the content into a variety of languages, we are also in the process of making experiments available for use on touchpads and smartphones in the hope of reaching different kinds of populations.

All LabintheWild experiments are currently 5–15 minutes long on the assumption that participants would be unlikely to choose to participate in a longer study. Researching the influence of the length of studies on participant recruitment, engagement, and data quality will be one of our immediate next steps. Study duration is one aspect where Mechanical Turk compares favorably to LabintheWild: While recruiting participants on Mechanical Turk for studies that last 30 minutes or more is a matter of increasing the financial compensation, we currently do not know whether the social comparison feedback on LabintheWild will be sufficient for recruiting and engaging similarly large numbers of participants.

In addition, many LabintheWild visitors arrive with the intention of participating in a specific study. It typically takes a few weeks for a new study to receive as much traffic as the established ones. Even though most LabintheWild studies eventually attract thousands of participants, recruiting participants via online labor markets is still a more effective recruitment mechanism when immediate participation is required.

It is important to note that the goal of LabintheWild is not to replace laboratory studies or experiments on Mechanical Turk. LabintheWild is intended to complement existing methods by enabling large scale replications of previous results, the extension of results with more diverse subject populations, and the generation of new results with studies that do not require specialized hardware or direct supervision of participants. In-lab studies will remain the best option for conducting experiments that require controlled settings, close observation of participant behavior, or specific devices. Mechanical Turk is arguably the best option for fast recruitment of participants and it is effective for recruiting participants for studies longer than 15 minutes. It might also be a better option for experiments that cannot be packaged into intrinsically motivating stories.

CONCLUSION

This research has addressed the issue that most research findings in HCI, psychology, behavioral economics, and related fields are based on studies with WEIRD (i.e., Western, Educated, Industrialized, Rich and Democratic) participants, and thus, might not generalize to other populations [1, 17]. Our goal was to find a way to conduct experiments with larger sample sizes including participants from non-WEIRD populations, and without compromising the data quality.

Reporting on almost two years of experience with our online experiment platform LabintheWild, we showed that online experiments that provide interesting personalized feedback (but no money) in exchange for study participation can fulfil this goal: LabintheWild attracts more than 1,000 participants on an average day, the sample population is significantly more diverse in terms of geographic dispersion and demographic

composition than what is common in laboratory studies, and LabintheWild experiments accurately replicate the results of in-lab studies.

We additionally analyzed statistics derived from Google Analytics, tweets that mention LabintheWild, and participants' qualitative feedback to explain who the participants are, how they hear about LabintheWild, why they participate, and how they contribute to ensuring the reliability of data.

Our findings emphasize the power of social comparison: In their tweets and comments, participants repeatedly refer to the personalized feedback that enables them to compare themselves to others. This is the main factor in motivating participants to recruit others, generating a self-perpetuating recruitment mechanism that spreads around the world. After less than two years, LabintheWild participants now come from 200 countries on six continents and have a variety of demographic backgrounds.

Our experience with LabintheWild demonstrates that many people are willing to contribute. Over the course of almost two years, we not only saw large numbers of participants volunteer the time, but we also received an overwhelming number of comments and emails containing feedback on technical issues, experiment design, suggestions for improvements, people offering help with the recruitment, sending us translations of our studies, or even offering the donation of bitcoins. We believe that one of our main future tasks should be to give back, perhaps by creating more ways for participants and researchers to interact, and by providing meaningful learning opportunities. Exploring participants' role in recruiting others, designing and debugging experiments, and analyzing and interpreting the data will be exciting avenues in the future.

DATA SETS

The data sets used for the analyses of the three replication experiments can be accessed at www.labinthewild.org/data.

ACKNOWLEDGMENTS

We thank the members of the Test My Brain project: Ken Nakayama, Laura Germine, and Sam Anthony for sharing expertise and code. Ofra Amir, Sarita Yardi Schoenebeck, and Trevor Croxson provided valuable feedback on the earlier drafts of this manuscript. Yuechen Zhao, Dianna Hu, Jonathan Taratuta-Titus, Rishav Mukherji, and William Xiao contributed to the implementation of LabintheWild. We also thank the Intelligent Interactive Systems Group at Harvard, and in particular Kenneth Arnold, for brainstorming the name for the platform, as well as experiment incentives and slogans. This work was funded in part by a research fellowship from the Alfred P. Sloan foundation, by a grant from the Harvard Mind Brain and Behavior Initiative, and by the Swiss National Science Foundation.

REFERENCES

1. Arnett, J. The neglected 95%: Why American psychology needs to become less American. *American Psychologist 63(7)* (2008), 602–14.
2. Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., and Plumb, I. The "Reading the Mind in the Eyes" Test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of child psychology and psychiatry, and allied disciplines 42*, 2 (Feb. 2001), 241–251.
3. Berinsky, A. J., Huber, G. A., and Lenz, G. S. Using Mechanical Turk as a subject recruitment tool for experimental research. *Political Analysis 20* (2012), 351–68.
4. Buhrmester, M., Kwang, T., and Gosling, S. D. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science 6*, 1 (2011), 3–5.
5. Cavanagh, J. P. Relation between the immediate memory span and the memory search rate. *Psychological Review 79*, 6 (1972), 525–530.
6. Cavanagh, J. P., and Chase, W. G. The equivalence of target and nontarget processing in visual search. *Perception & Psychophysics 9*, 6 (Nov. 1971), 493–495.
7. Chase, W. G., and Calfee, R. C. Modality and similarity effects in short-term recognition memory. *J Exp Psychol 81*, 3 (1969), 510–514.
8. Corbin, L., and Marquer, J. Effect of a simple experimental control: The recall constraint in Sternberg's memory scanning task. *European Journal of Cognitive Psychology 20*, 5 (Sept. 2008), 913–935.
9. Cruse, D., and Clifton, Jr, C. Recoding strategies and the retrieval of information from memory. *Cognitive Psychology 4*, 2 (1973), 157–193.
10. Downs, J. S., Holbrook, M. B., Sheng, S., and Cranor, L. F. Are your participants gaming the system?: Screening Mechanical Turk workers. In *Proc. CHI'10*, ACM (2010), 2399–2402.
11. Eriksson, K., and Simpson, B. Emotional reactions to losing explain gender differences in entering a risky lottery. *Judgment and Decision Making 5*, 3 (2010), 159–163.
12. Festinger, L. A theory of social comparison processes. *Human relations 7(2)* (1954), 117–140.
13. Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., and Wilmer, J. B. 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.
14. Glaser, B. G., and Strauss, A. L. *The discovery of grounded theory: Strategies for qualitative research*. Transaction Publishers, 2009.
15. Goodman, J. K., Cryder, C. E., and Cheema, A. Data collection in a flat world: The strengths and weaknesses of mechanical turk samples. *Journal of Behavioral Decision Making 26*, 3 (2013), 213–224.
16. Heer, J., and Bostock, M. Crowdsourcing graphical perception: Using Mechanical Turk to assess visualization design. In *Proc. CHI'10* (2010), 203–212.

17. Henrich, J., Heine, S. J., and Norenzayan, A. The weirdest people in the world. *Behavioral and Brain Sciences 33* (2010), 61–83.
18. Horton, J. J., Rand, D. G., and Zeckhauser, R. J. The online laboratory: Conducting experiments in a real labor market. *Experimental Economics* (2011).
19. Ipeirotis, P. Demographics of Mechanical Turk. NYU Working Paper No. CEDER-10-01, March 2010.
20. Ipeirotis, P. G. Analyzing the Amazon Mechanical Turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students 17*, 2 (2010), 16–21.
21. Kapelner, A., and Chandler, D. Preventing satisficing in online surveys. In *Proceedings of CrowdConf* (2010).
22. Kittur, A., Chi, E. H., and Suh, B. Crowdsourcing user studies with Mechanical Turk. In *Proc. CHI'08*, ACM (2008), 453–456.
23. Komarov, S., Reinecke, K., and Gajos, K. Z. Crowdsourcing performance evaluations of user interfaces. In *Proc. CHI'13* (2013), 207–216.
24. Kristofferson, M. W. Effects of practice on character-classification performance. *Canadian Journal of Psychology/Revue canadienne de psychologie 26*, 1 (1972), 54–60.
25. Lindgaard, G., Fernandes, G., Dudek, C., and Brown, J. Attention web designers: You have 50 milliseconds to make a good first impression! *Behaviour & Information Technology 25*, 2 (2006), 115–126.
26. Mason, W., and Suri, S. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods 44*, 1 (2012), 1–23.
27. Nickerson, R. S. Response times with a memory-dependent decision task. *J Exp Psychol 72*, 5 (Nov. 1966), 761–769.
28. Oppenheimer, D., Meyvis, T., and Davidenko, N. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology 45*, 4 (2009), 867–872.
29. Orne, M. T. Demand characteristics and the concept of quasi-controls. In *Artifacts in Behavioral Research*, R. Rosenthal and R. L. Rosnow, Eds. Oxford University Press, July 2009, ch. 5.
30. Paolacci, G., Chandler, J., and Ipeirotis, P. G. Running experiments on Amazon Mechanical Turk. *Judgment and Decision making 5*, 5 (2010), 411–419.
31. Rand, D. G. The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of theoretical biology* (2012).
32. Reinecke, K., and Gajos, K. Z. Quantifying visual preferences around the world. In *Proc. CHI'14*, ACM (2014), 11–20.
33. Reinecke, K., Nguyen, M. K., Bernstein, A., Näf, M., and Gajos, K. Z. Doodle around the world: Online scheduling behavior reflects cultural differences in time perception and group decision-making. In *Proc. CSCW'13*, ACM (2013), 45–54.
34. Shaw, A. D., Horton, J. J., and Chen, D. L. Designing incentives for inexpert human raters. In *Proc. CSCW'11*, ACM (2011), 275–284.
35. Sternberg, S. High-speed scanning in human memory. *Science 153*, 3736 (Aug. 1966), 652–654.
36. Sternberg, S. Two operations in character recognition: Some evidence from reaction-time measurements. *Perception & Psychophysics 2*, 2 (1967), 45–53.
37. Sternberg, S. Memory scanning: New findings and current controversies. *The Quarterly journal of experimental psychology 27*, 1 (1975), 1–32.
38. Suri, S., and Watts, D. Cooperation and contagion in networked public goods experiments. *Arxiv preprint arXiv10081276* (2010).
39. Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., and Malone, T. W. Evidence for a collective intelligence factor in the performance of human groups. *science 330*, 6004 (2010), 686–688.