The Potential for Scientific Outreach and Learning in Mechanical Turk Experiments

Eunice Jun, Morelle Arian, and Katharina Reinecke
DUB Group, Paul G. Allen School of Computer Science and Engineering, University of Washington
{emjun, morellea, reinecke}@cs.washington.edu

ABSTRACT

The global reach of online experiments and their wide adoption in fields ranging from political science to computer science poses an underexplored opportunity for learning at scale: the possibility of participants learning about the research to which they contribute data. We conducted three experiments on Amazon's Mechanical Turk to evaluate whether participants of paid online experiments are interested in learning about research, what information they find most interesting, and whether providing them with such information actually leads to learning gains. Our findings show that 40% of our participants on Mechanical Turk actively sought out post-experiment learning opportunities despite having already received their financial compensation. Participants expressed high interest in a range of research topics, including previous research and experimental design. Finally, we find that participants comprehend and accurately recall facts from post-experiment learning opportunities. Our findings suggest that Mechanical Turk can be a valuable platform for learning at scale and scientific outreach.

ACM Classification Keywords

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

Author Keywords

online experimentation; learning; scientific outreach

INTRODUCTION

Online studies have become an increasingly common, validated method of data collection in various fields, including the social sciences, medicine, and computer science [10, 13, 35, 3, 15, 8]. For example, in two top social psychology journals, the percentage of articles with online experiments, many of them conducted on the online labor market Mechanical Turk (MTurk), increased five-fold between the years 2012 and 2015 [41]. One of the reasons for this increase is that MTurk provides convenient access to a large number of participants; a survey of MTurk workers in 2015 found 30,002 unique MTurk workers [30]. MTurk's growing popularity in research is especially profound considering that MTurk was not designed for online experimentation.

Despite its popular use, MTurk (and other paid online experiment platforms) has been largely overlooked in the research community as a context for learning, perhaps because MTurk workers are paid to complete an experiment. This is in contrast to volunteer-based online experiments and citizen science platforms, which commonly offer learning opportunities as an incentive to compensate participants for their time. Prior work has found that volunteers often greatly appreciate additional information about the background of studies they take part in or the broader research objectives [32]. Despite previous research that shows the diversity of motivations MTurkers have for partcipating in tasks, including to have fun and learn new skills [17, 20, 6], it is unclear whether financially-compensated participants, such as those on MTurk, will be interested in information about research given that they often use the platform to earn money.

In this paper, we evaluate whether online experiments on Mechanical Turk can provide an opportunity for scientific outreach and learning. Our three research questions are: (1) Are MTurk workers participating in online experiments interested in learning about research?; (2) What kind of information about research interests them most?; and (3) Do participants engage with information about research in a way that they comprehend, and can they recall the information later? Drawing on prior work and a landscape analysis of existing scientific outreach efforts provided in volunteer-based online experiments, we focused on four topics for scientific outreach and learning: research impact, previous research findings, experimental design, and other research motivations. 40% of our Mechanical Turk participants actively sought out post-experiment learning opportunities that were completely optional and offered after providing financial compensation. We found that participants are interested in all four topics. Finally, our third experiment showed that all participants later accurately recalled some of the information provided, suggesting that post-experiment learning opportunities can successfully serve as a platform for scientific outreach.

Our work makes two main contributions:

1. We provide empirical evidence for the importance and promise of designing post-experimental scientific outreach.
2. Our research opens up a new space for learning at scale by showing how researchers can provide valuable scien-

L@S 2018, June, 2018, London, UK.

ISBN ISBN to appear here. . . $15.00
DOI: DOItoappearhere

tific learning opportunities to a population that is distinct from those accessing MOOCs, citizen science projects, or volunteer-based online experiments.

We conclude with a call to action for researchers to consider the learning needs and interests participants have and to engage in scientific outreach on MTurk and other paid platforms.

BACKGROUND AND RELATED WORK

Scientific outreach has been central to the relationship between science and society [26]. The kind of learning achieved by scientific outreach can (1) increase public awareness of how research is conducted and the uncertainty or debate in scientific findings, (2) open up a dialogue for the public to share their concerns and ideas, and (3) facilitate a mutual understanding between scientists and the general citizenry. Scientific outreach is distinct from classroom learning because rather than the acquisition of new skills or collaboration in the process of conducting research, it is focused on comprehending scientific findings and facilitating two-way communication between scientists and participants.

Our focus on scientific outreach in this paper is distinct from the forms of learning in citizen science projects in that we target people who do not necessarily plan to engage with research. Further, our focus on Mechanical Turk, an online labor market, reaches a different population from the scientific outreach efforts in citizen science projects or in volunteer-based online experiments.

Learning Opportunities in Crowd Work

Research has suggested that there is an "information asymmetry" between requesters (i.e., those who post a task or study) and workers on Mechanical Turk; workers often do not know what their work contributes towards [37]. This can be especially problematic if researchers collect data from MTurk participants for scientific goals that may not align with participants' ethical principles or ideals. As a response to such concerns, researchers have developed Turkopticon [18], a website workers can use to rate and review requesters and share information with each other, and Daemo [12], a crowd-designed online labor market which involved crowd workers through the design, development, and launch. While these efforts greatly help the MTurk community, neither of these approaches focuses on rethinking the labor market used by researchers as a platform for scientific outreach and learning. Our work fills this gap and reconsiders MTurk as a new place for scientific learning. This is a challenge because participants are usually treated as anonymous, paid workers. The transaction between participants and researchers is considered complete with the last trial in a study. This challenge is important to face because the sustainability of crowd work and the possibility to leverage crowds for online studies relies on integrating and reconsidering learning in these contexts [22].

Learning in Citizen Science Projects

Citizen science projects, such as eBird [38] and GalaxyZoo [25], have provided volunteer participants with learning experiences by involving them directly in the research process. Ebird involves bird enthusiasts who share information about bird sightings in order to promote conservation and biodiversity. GalaxyZoo involves the public in labeling and categorizing astronomical images. Through these volunteer activities, publicly available datasets, and forums, both of these citizen science projects provide learning opportunities to volunteers and the interested public. Another citizen science project, Gut Instinct, also increases people's understanding of a specific domain area; participants contribute data about the human microbiome while also learning about their own and posing hypotheses and theories to be tested [33].

Learning Opportunities in Volunteer-based Online Studies

Participants in laboratory studies commonly receive more extensive information about how they are contributing to a research project through informal conversations with researchers upon completion of a task. For example, before a participant leaves, best practice guidelines suggest that researchers should ask whether participants have any more questions, in line with the APA debriefing guideline to provide "a prompt opportunity for participants to obtain appropriate information about the nature, results, and conclusions of the research" [1]. Such opportunities often lead researchers to provide additional information about their study, the broader research goal, or how they decided to work on a specific topic. These informal interactions help to bridge the gap between participants and researchers and increase general global awareness about the scientific process and the numerous niche areas of inquiry.

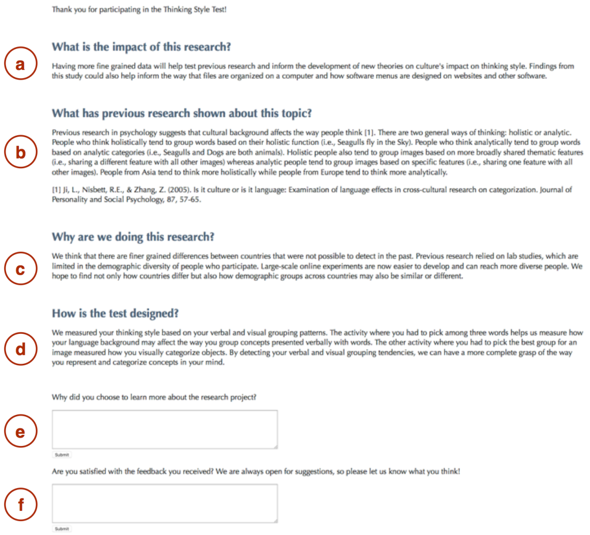
In the online setting, volunteer-based online experiment platforms, such as MySocialBrain, LabintheWild, and GamesWithWords do not pay participants for study completion. Instead, they compensate participants with information about their performance or a personal characteristic (e.g., thinking style or writing style) in return for voluntarily completing an online study. To do so, they include a scientific outreach page at the end of each study, which often enables participants to see their personalized results and share the results with others [35]. Recent work has found that participants on LabintheWild, for example, often participate to learn about themselves or to help science [19] and that they appreciate opportunities for learning more about the background of studies or the broader research goals [32]. Oliveira et al. identified three broad areas that volunteer participants in large-scale online experiments were interested in learning about: themselves (e.g., in comparison to others), the research project, and experimental design [32].

Research has also shown that participants can benefit from learning in online studies. For example, online studies can provide casual observational learning throughout participation, increasing participants' understanding about nutrition [9]. Other research has found ways to teach the public about the scientific process, and involve them in developing scientific ideas [33, 39].

We extend this related work by investigating if post-experimental learning outcomes, as experienced in the lab and provided in the volunteer-based setting, are feasible – and fruitful – in the paid online setting.

IDENTIFYING LEARNING OPPORTUNITIES

To establish a set of possible learning opportunities, we conducted an analysis of post-experimental information pages



Thank you for participating in the Thinking Style Test!

What is the impact of this research?

Having more fine grained data will help test previous research and inform the development of new theories on culture's impact on thinking style. Findings from this study could also help inform the way that files are organized on a computer and how software menus are designed on websites and other software.

What has previous research shown about this topic?

Previous research in psychology suggests that cultural background affects the way people think [1]. There are two general ways of thinking: holistic or analytic. People who think holistically tend to group words based on their holistic function (i.e., Seagulls fly in the Sky). People who think analytically tend to group words based on analytic categories (i.e., Seagulls and Dogs are both animals). Holistic people also tend to group images based on more broadly shared thematic features (i.e., sharing a different feature with all other images) whereas analytic people tend to group images based on specific features (i.e., sharing one feature with all other images). People from Asia tend to think more holistically while people from Europe tend to think more analytically.

[1] Ji, L., Nisbett, R.E., & Zhang, Z. (2005). Is it culture or is it language: Examination of language effects in cross-cultural research on categorization. Journal of Personality and Social Psychology, 87, 57-65.

Why are we doing this research?

We think that there are finer grained differences between countries that were not possible to detect in the past. Previous research relied on lab studies, which are limited in the demographic diversity of people who participate. Large-scale online experiments are now easier to develop and can reach more diverse people. We hope to find not only how countries differ but also how demographic groups across countries may also be similar or different.

How is the test designed?

We measured your thinking style based on your verbal and visual grouping patterns. The activity where you had to pick among three words helps us measure how your language background may affect the way you group concepts presented verbally with words. The other activity where you had to pick the best group for an image measured how you visually categorize objects. By detecting your verbal and visual grouping tendencies, we can have a more complete grasp of the way you represent and categorize concepts in your mind.

Why did you choose to learn more about the research project?

Are you satisfied with the feedback you received? We are always open for suggestions, so please let us know what you think!

Figure 1: Scientific outreach page from the choice test. Each study included sections covering (a) the research goal and impact, (b) a summary of relevant previous research, (c) explanation of the motivations researchers have for conducting the online study, and (d) description of the experimental design. Part (e) was only included in Experiment 1 to ask about participants' motivations for choosing to learn more. Part (f) was a place for participants to give open-ended general feedback.

offered on volunteer-based online experiment platforms. Our goal was to find examples of learning opportunities that could be easily generated by researchers without requiring too much of their time. To collect as many different examples as possible, we used a purposive sampling method, landscape sampling, following the procedure described in [5] and [40].

We started our search for online studies that matched our criteria (i.e., behavioral online experiments and surveys that provide some kind of learning opportunity) on the Social Psychology Network [31] and SciStarter [36], two publicly available lists of online studies and citizen science projects. Once we found an initial set of online studies with post-experimental learning opportunities, we contacted the researchers and asked for nominations of other, similar projects, following a snowball recruiting technique. We included any post-experimental outreach pages that (a) were listed on a website previously not included in our dataset and (b) included content that we had not seen on other pages before.

Our final sample consisted of 35 post-experimental pages from 12 online experiment platforms or websites. We found the following categories that could lead to science learning:

- Research goals and potential impact: What are the short-term and/or long-term research goals? How will this data benefit society or advance a discipline?
- Previous research: What has been found or studied in this research area before?
- Study and experimental design: Why was the online experiment designed the way that it was with specific instructions, timers, progression of trials, etc.?
- Other research motivations: If there is a specific hypothesis, what is it? If not, what is the need for collecting data using this study?

We later use these categories to design post-experimental outreach pages to be used in our experiments, as described below. An overview of the three experiments can be found in Figure 2.

Another common category in our analysis that has interested participants in previous work is personalized results. We deliberately excluded this category from our experiments because we acknowledge that personalized results, while interesting to participants, is not always feasible for researchers to include in scientific outreach while collecting data. Our findings, therefore, can be applied to a broader range of online studies where personalized feedback is not readily available.

EXPERIMENT 1: ARE PARTICIPANTS ON MECHANICAL TURK INTERESTED IN LEARNING ABOUT RESEARCH?

Our first research aim was to assess how desirable post-experiment scientific outreach and learning opportunities would be for participants. To address this aim, we asked, "What proportion of participants on Mechanical Turk would *without compensation* actively seek out optional learning opportunities after completing an online study?"

Materials

We developed post-experimental outreach pages for three studies. These studies cover a range of topics and experimental tasks to improve the generalizability of our results:

i a *subjective survey* that asked participants to re-create and answer questions about a particular social situation,

ii a *choice test* where participants had to group or categorize stimuli, and

iii an *objective recall test* where participants had to remember geometric configurations and reproduce them.

The first author wrote and designed simple text-only informative pages for each of the three studies, as shown in Figure 1. The pages were intentionally designed with text only because (1) we wanted to eliminate the possible confounding factor of presentation style (i.e., visualization vs. text vs. both) because prior work has found differences in information recall with these modalities [21] and (2) our landscape analysis showed that post-experimental pages offered to participants also primarily relied on text.

Each page included an average of 361 words (sd=58.5). To find out how much time would be needed to read these pages, we asked three people to read the page word by word. The average time it took to read was 58 seconds (sd=26.8s).

Experimental Design and Procedure

Our experiment used a between-subjects design with the three studies described above.

After giving informed consent, participants completed one of the studies. Afterwards, a page offered the token they needed to receive payment. Underneath the token, participants saw an optional button to learn more about the research project

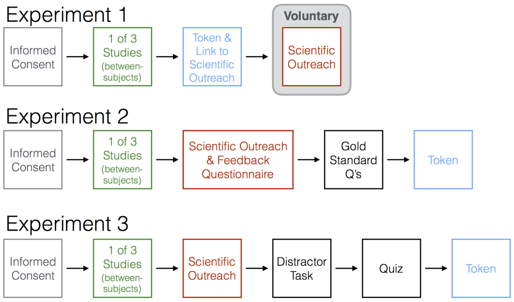
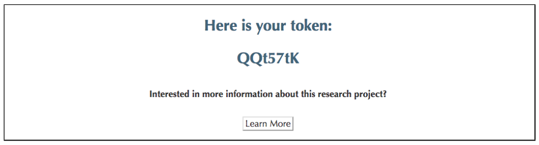

Experiment 1
Informed Consent
1 of 3 Studies (between-subjects)
Token & Link to Scientific Outreach
Voluntary
Scientific Outreach
Experiment 2
Informed Consent
1 of 3 Studies (between-subjects)
Scientific Outreach & Feedback Questionnaire
Gold Standard Q's
Token
Experiment 3
Informed Consent
1 of 3 Studies (between-subjects)
Scientific Outreach
Distractor Task
Quiz
Token

Figure 2: Overview of experimental progression. The token is a code needed to receive monetary compensation.



Here is your token:

QQt57tK

Interested in more information about this research project?

Figure 3: Experiment 1 token page, where participants receive their Mechanical Turk token for payment and optionally can click "Learn More" to continue to the scientific outreach page.

(Figure 3). We provided the outreach page as a separate page (rather than adding the information to the page with the token) to be able to accurately record the time that people spend reading through the information. In addition, recording how many participants *actively* click on the button to access the outreach page provided us with a conservative measure of the percentage of participants interested in this information.

The size and font of the token was designed to be more visually salient in order to signal clearly to the MTurk participants that they had completed the task and their payment did not depend on their doing anything more. There was no other incentive to voluntarily learn more about the research project. Moreover, we referred to the learning opportunities as "more information" in order to transparently communicate that participants would receive additional information only and no bonuses. The Turkers who clicked on the link to learn more saw the outreach pages described in the Materials section and Figure 1.

Participants were compensated $0.67 for taking the choice test and subjective survey, each of which took approximately 5 minutes to complete, and $1.35 for taking the objective recall test which took approximately 10 minutes.

Measures

We calculated the *percentage of participants* who chose to click on the "Learn more" button on the landing page (Figure 3). We also tracked the *amount of time* participants spent on the conclusion page. To gain a better understanding of participants' motivations and reactions to seeking the optional learning opportunities, we posed two *open-ended questions* on the outreach pages: participants could optionally explain why they clicked to learn more ("Why did you choose to learn more about the research project?"), and they could provide general open-ended feedback on the research information pages ("Are you satisfied with the feedback you received? We are always open for suggestions, so please let us know what you think!").

Table 1: Mean percentages and standard deviations of participants who clicked to learn more across studies and the overall 95% confidence interval in Experiment 1. The 95% confidence interval was calculated using a Student's t-distribution.

Task	Mean	SD	95% CI
Objective Recall Test	52.63	11.45	
Subjective Survey	33.33	12.17	
Choice Test	27.27	13.43	
Aggregate	40	7.30	[25.2, 54.8]

Participants

We collected data from 61 participants on MTurk. Due to repeated participation and missing data, we report on data from 45 unique participants (19 objective recall test, 5 subjective survey, and 11 choice test). For repeat participants, we only included their earliest submission in our dataset. We did not specify any prerequisites (e.g., minimum approval rates, skill sets, etc.) to increase the generalizability of our results to the larger population of people who take online experiments on MTurk. We collected data on different days and at different times of the day in order to obtain as diverse a sample of participants as possible.

Due to differences in pay, the three studies were hosted as separate HITs on MTurk. Ad hoc comparisons revealed no significant differences between participants across the three studies. Among the participants, 19 self-identified as female and 26 as male. The mean age was 36.6 (sd=12.0). The majority of them, 34 participants, reported being from the United States, 9 from India, 1 from Singapore, and 1 from Venezuela. Our sample was slightly older than the median age of 30 years (mean = 32) that has previously been reported for workers on MTurk [29] but was representative of the country distribution found in [34].

Results

Of 45 participants across the three studies, 18 participants (40%) clicked on the button to learn more. The median time these participants voluntarily spent on the research information pages was 41.1 seconds (mean = 49.1s, sd = 35.5s). This is slightly less time than the 58 seconds that we found it takes on average to read the pages, suggesting that participants most likely read a substantial part, but not the entire page.

There was no statistically significant difference across studies in the percentage of participants who sought additional learning opportunities about research (F(2,42) = 1.12, p = .34). Table 1 shows the study and aggregate means, standard deviations, and 95% confidence interval.

We additionally used Bayesian statistics to estimate how our results generalize from our sample to a larger MTurk population. Our Bayesian analysis gives a more conservative interpretation of our findings. We followed the procedure outlined in [27] and used the brms package [7] for R. We created a mixed effects linear Bayesian model with an expert-determined prior. We asked an expert in crowdsourcing and online experiments who had significant experience with MTurk and who was not involved in this project for a prior probability distribution (short *prior*) that expressed their belief about the percentage of

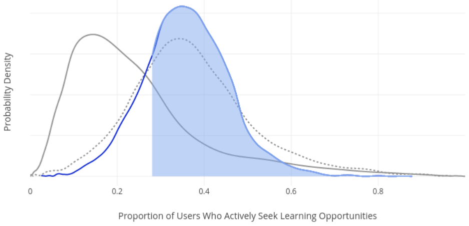


Figure 4: The uninformed (dashed grey curve) and posterior (blue curve) Bayesian models of the percentage of participants seeking scientific outreach learning opportunities, given an expert prior (solid grey curve). The shaded blue region shows the ~75% credible interval. Approximately 75% of the time, at least 30% of participants on Mechanical Turk will voluntarily seek out additional information about the research.

participants who might click to learn more about the research. The prior was defined using a Student's t-distribution with 3 degrees of freedom (defined over log-odds scale because we are modeling a Bernoulli random variable: click or not click). Given that t-distributions are normal distributions with heavier tails, this distribution simulated the small sample size and unknown variance in our data. The prior assumed that approximately 20% of participants on MTurk would be interested in optional scientific learning and that it would be *extremely* surprising to see more than 50% of participants interested.

Visualizations of the uninformed, expert (prior), and final (posterior) Bayesian models are shown in Figure 4. The final model shows a ~75% credible interval (the Bayesian analog for confidence interval), suggesting that 75% of the time, researchers can expect that at least 30% of participants will actively seek learning opportunities. This result suggests a promising potential for learning at scale given the large number of MTurk workers who participate in research experiments.

Why did participants want to learn more?

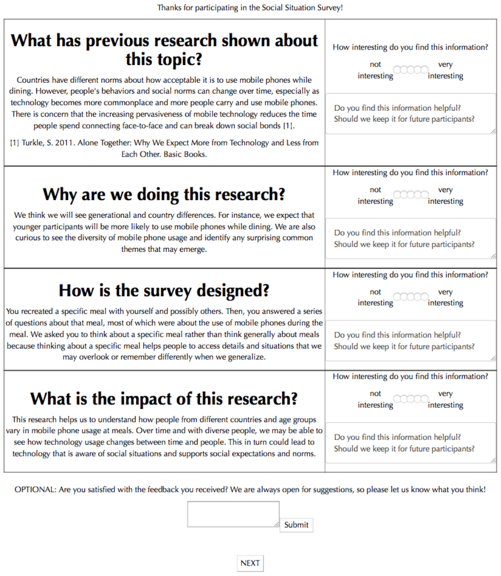
Ten of the 18 participants who proceeded to the learning pages voluntarily left comments explaining why they chose to click through and see the outreach pages. Participants uniformly described their motivation for accessing the page with curiosity; they were curious to hear more about the research they had just participated in. For example, P 2, who had participated in the objective recall test, described their curiosity as "I was curious what you would be looking at with this seemingly inconsequential task." P 36 more directly spoke to their desire to learn: "I would like to learn more on this survey. So I chose this page." Altogether, the comments emphasized participants' desires to find out what kind of research they had just contributed to and what the research might be used for.

EXPERIMENT 2: WHAT KINDS OF INFORMATION ABOUT RESEARCH INTERESTS PARTICIPANTS THE MOST?

The aim of our second experiment was to assess and compare the relative "interestingness" of the various categories included in post-experiment outreach pages.

Materials

We used the same three studies and outreach pages as in the first experiment.



Thanks for participating in the Social Situation Survey!

What has previous research shown about this topic?

Countries have different norms about how acceptable it is to use mobile phones while dining. However, people's behaviors and social norms can change over time, especially as technology becomes more commonplace and more people carry and use mobile phones. There is concern that the increasing pervasiveness of mobile technology reduces the time people spend connecting face-to-face and can break down social bonds [1].

[1] Turkle, S. 2011. Alone Together: Why We Expect More from Technology and Less from Each Other. Basic Books.

How interesting do you find this information?
not interesting ○○○○○ very interesting
Do you find this information helpful? Should we keep it for future participants?

Why are we doing this research?

We think we will see generational and country differences. For instance, we expect that younger participants will be more likely to use mobile phones while dining. We are also curious to see the diversity of mobile phone usage and identify any surprising common themes that may emerge.

How interesting do you find this information?
not interesting ○○○○○ very interesting
Do you find this information helpful? Should we keep it for future participants?

How is the survey designed?

You recreated a specific meal with yourself and possibly others. Then, you answered a series of questions about that meal, most of which were about the use of mobile phones during the meal. We asked you to think about a specific meal rather than think generally about meals because thinking about a specific meal helps people to access details and situations that we may overlook or remember differently when we generalize.

How interesting do you find this information?
not interesting ○○○○○ very interesting
Do you find this information helpful? Should we keep it for future participants?

What is the impact of this research?

This research helps us to understand how people from different countries and age groups vary in mobile phone usage at meals. Over time and with diverse people, we may be able to see how technology usage changes between time and people. This in turn could lead to technology that is aware of social situations and supports social expectations and norms.

How interesting do you find this information?
not interesting ○○○○○ very interesting
Do you find this information helpful? Should we keep it for future participants?

OPTIONAL: Are you satisfied with the feedback you received? We are always open for suggestions, so please let us know what you think!

Figure 5: Scientific outreach page with feedback questions as included in Experiment 2.

Experimental Design and Procedure

Upon giving informed consent, participants completed one of the three aforementioned studies and then saw the outreach pages right away. Each major content section (addressing the different categories identified in the landscape analysis) was outlined in a box with the following questions: 1) "How interesting do you find this information?" with a five-point response scale with end points marked with "not interesting" and "very interesting" and 2) "Do you find this information helpful? Should we keep it for future participants?" with a text box for responding. We required participants to provide feedback on each of the pieces of information. In addition, we provided a space for participants to voluntarily comment on anything else about the conclusion page. Figure 5 shows the layout for the subjective survey.

To control for the possible effect of the order of content on interesting ratings (e.g., learning opportunities at the top of the page could receive higher ratings, perhaps due to novelty), we randomized the order of the content shown on the outreach pages between participants.

Finally, we included two gold standard questions after the outreach pages in order to exclude those who may have circumvented reading the conclusion pages. The gold standard questions measured simple recall. We also determined that participants who gave the same rating and text feedback verbatim on the entire outreach page were satisficing and excluded them from the analysis.

Measures

Our primary dependent variable was the interesting ratings and, secondarily, the open-ended feedback. We scored the interesting ratings from 1 through 5 with 1 corresponding to the endpoint marked as "not interesting" and 5 corresponding to the endpoint marked as "very interesting." Higher scores meant that the learning opportunities were more interesting.

Participants

An a priori power analysis ($\alpha = 0.05$, power = 0.80, F effect size = .25) for an ANOVA with 3 between-subject groups (3 studies) and 4 within subject groups (4 categories of outreach information) showed the need for at least thirty participants. We collected data from ninety-five participants and had to exclude twenty-seven due to repeat participation, participation in Experiment 1, or missing data. We further narrowed the sample to thirty-two participants (12 objective recall test, 8 subjective survey, and 12 choice test) who answered at least one of two gold standard questions correctly and who did not exhibit satisficing behavior. We chose to include participants who correctly answered 1 out of 2 gold standard questions correctly because participants seemed to exhibit an understanding of the material in their responses but still answered the gold standard questions incorrectly.

Eleven participants self-identified as female and twenty-one as male. The mean age was 35.6 years (sd=14.3). 30 self-reported being from the United States, 1 from Canada, and 1 did not respond.

As in Experiment 1, we did not specify any prerequisites and collected data on different days and times of day to diversify our sample. Participants were compensated $1.60 for tasks involving the choice test or subject survey that were estimated to take 12 minutes and $2 for tasks involving the objective recall test that was estimated to take 15 minutes.

Analysis and Results

Our analysis showed that participants were roughly equally interested in all four topics, with median ratings of the different categories ranging between 4-5 on a 5-point scale.

Collapsing across studies, there was no statistically significant difference among categories of outreach information (F(3,124) = 0.79, p = .50). Analyzing the different studies separately, there were no statistically significant differences between the interesting ratings for the four categories in the subjective survey (F(3,28) = 0.25, p = .86) or the choice test (F(3,44) = 0.41, p = 75). In the the objective recall test, there was a statistically significant difference across categories (F(3,44) = 3.00, $p < .05$). The median rating for the information about previous research (median = 5) and the researchers' motivations (median = 5) were higher than the median ratings for experimental design (median = 4.5) and research impact (median = 4).

Confirming their ratings, most participants left comments to keep the various pieces of information. Nonetheless, there was still a spread of reactions. For instance, some expressed a clear like of the information by leaving comments such as "This is cool to tell us how the survey was designed. That is something that I like." (P 141). Others, however, provided less enthusiastic feedback but still asked to keep the content: "This part is less interesting, but still important" (P 61). P 62 captured the variability but overall positive response of the participants to the various types of content: "Leave this and all information available. Even if it's not helpful it's really neat."

Participants also came up with other content they would like to see in the outreach pages in the future. Many participants, such as P 61, desired personalized content after the study, stating "I would like to know how I did compared to the tendencies." P 166's response pointed to possible future avenues of research, "Cultural context is fascinating; it would also be interesting to see how personality and background (e.g., I'm an artist) plays into reactions."

Some participants demonstrated future engagement as a result of the outreach. P 58 responded, "This is new information to me and I'm going to look it up after the sudy, thank you for the source." The scientific outreach information also led to increased self-awareness and self-reflection: "yes I wondered myself about some of my choices" (P 59).

In general, participants responded positively to seeing how their participation was contributing to the scientific community. P 147 said, "Yes, I think it's interesting and I like to see the results of my efforts."

EXPERIMENT 3: DO PARTICIPANTS LEARN FROM POST-EXPERIMENT SCIENTIFIC OUTREACH INFORMATION?

With Experiment 2 confirming that the four learning opportunities offered were indeed interesting to participants on MTurk, we next asked "Do participants engage with information about research in a way that they comprehend and can recall later?".

Materials

We used the same three studies as in the previous two experiments. To evaluate comprehension and recall of the post-experiment outreach pages, we developed 12 quiz questions (one question for each of the four learning opportunities for three studies). For example, we asked questions such as "What is the potential impact of this study?" (research impact), "How do people from Europe and Asia differ in their memory of geometric configurations?" (previous research), "Why did we measure both your visual and verbal grouping tendencies?" (experimental design), and "We expect to see two broad differences in mobile phone usage. One difference is between countries. What is the other main difference we expect to see?" (other motivations).

To confirm that these quiz questions can be answered after reading the outreach pages, we tested them in-person with people who were unaware of our experiment's intentions. Six volunteers (two per study) read the learning pages (unaware that they would be quizzed) and then answered the quiz questions after two minutes. The volunteers were able to answer the questions accurately. They also shared verbal feedback about their thoughts on the difficulty of the questions and whether they were too pointed or vague. Based on this evaluation, one quiz question was determined confusing and was changed before data collection.

Experimental Design and Procedure

The experiment was designed such that the online studies were the between-subjects variable (3 studies) and the type of outreach information was the within-subjects variable (4 categories of information).

Participants gave informed consent, completed one of the three studies, and were then presented with the outreach page. They were not required to stay on the outreach page for a predetermined amount of time but instead were free to move on as quickly or slowly as they desired. They were then asked to complete a paper folding task [11] for at most 3 minutes, which served as a distractor task. Afterwards, they were presented with the four text-entry quiz questions that measured their recall and comprehension of the research information. Participants were not forewarned that they would be quizzed on the material.

To avoid possible ordering effects, we randomized the order of the categories of information on the outreach page and the order of the quiz questions across participants. In addition, only one quiz question was shown at a time in order to prevent participants from changing their answers between questions.

Measures

We wanted to know if one of the four categories of information found on the outreach pages (research impact, previous research, experimental design, and other research motivations) was more conducive to comprehension and accurate recall, a metric that is the basis of all higher order learning [4]. Our primary learning measure was therefore *comprehension and accuracy of recall* between categories of learning opportunities and within participants. The accuracy of participants' answers was determined by two researchers who rated the responses and resolved any conflicts. Without recall of the learning opportunities, we would not expect any more complex forms of learning about research to take place (such as creating follow-up hypotheses or experiments). We additionally captured *time* spent on the conclusion pages to enable comparison with the results of Experiment 1.

Participants

To determine the minimum number of participants needed to detect a small-medium F effect size of .25, we conducted an a priori power analysis (α = .05, power = .80) for an ANOVA with 3 between-subject groups (3 studies) and 4 within-subject groups (4 categories of outreach information). The minimum participants required was thirty. We collected data from forty-five and had to omit 6 participants' data due to missing data and repeat or prior participation in Experiments 1 or 2. We report on data from thirty-nine unique participants.

Thirteen participants self-identified as female and twenty-six as male. The mean age was 28.9 (sd=5.84). Thirty-three participants were from the United States and 6 from India.

As in Experiments 1 and 2, we did not specify any prerequisites and collected data on different days and at different times of day in order to reach a diverse sample of MTurk participants. Participants were compensated $1.60 for tasks involving the choice test or subject survey that were estimated to take 12 minutes and $2 for tasks involving the objective recall test that was estimated to take 15 minutes.

Analysis and Results

The median recall accuracy score across all studies was 1 question out of four (mean = 0.82). The median time voluntarily spent on the outreach pages was 9.4s (sd=9.4s, min=3.4s, max=45.6s).

To analyze whether specific categories were comprehended more than others, we ran a general linear model across all three studies, treating the four categories as a main effect, participant number as a random effect (because of the within-subjects analysis), and quiz score as the dependent variable. We found that participants were significantly more likely to answer correctly questions about other research motivations than the information about the experimental design and research impact (Table 2).

Table 2: Differences in comprehension and recall accuracy among the four learning opportunities presented on the scientific outreach pages. Experimental design was used as the default comparison category. *p <.05

	Estimate	Standard error	t value	Pr(>|t|)
Intercept	0.12821	0.06322	2.028	0.0443*
Other research motivations	0.20513	0.08941	2.294	0.0231*
Previous research	0.15385	0.08941	1.721	0.0873
Research Impact	-0.05128	0.08941	-0.574	0.5671

We found that the mean comprehension was different across the studies (F(2,36) = 4.79, p < .05). Following up with t-tests, we found that comprehension of the subjective survey (m = 1.36) was statistically significantly higher than the choice test (m = 0.33, t(21.06) = -2.84, p < .01). There was also a statistically significant difference in comprehension between the subjective survey and the objective recall test (m = 0.69, t(20.45) = 1.88, p < .05). This suggests that participants comprehended and remembered most information about the research behind the subjective survey, which asked for comparatively more personal social information.

DISCUSSION

The results of our three experiments showed that a significant proportion of study participants on Mechanical Turk are interested in learning about research. Providing median ratings between 4-5 on a 5-point scale, they appreciated information about research goals and the potential impact of a study, prior work, the study and experimental design, and other research motivations. Moreover, our participants accurately recalled at least one of four facts when spending a mere 10 seconds on the outreach page. Altogether, our findings reveal an opportunity for scientific outreach benefiting workers on online labor markets. In the following, we will discuss the implications of our findings in detail.

MTurk participants are interested in learning about research

When we embarked on this project, we were unable to pose a one-directional hypothesis on whether or not MTurk participants would be interested in learning about research. For one, much research has discussed that people predominantly use MTurk to supplement their income [28, 24, 6], and that they often attempt to minimize the time spent on a task in order to maximize financial compensation over time [28, 14]. On the other hand, research has also shown that some people on MTurk participate for fun, or to learn new skills [6, 28]. Our own results show that 40% of participants voluntarily sought additional information about the research. Our Bayesian analysis additionally suggests that for 75% of online experiments offered on MTurk, at least 30% of participants will actively seek learning opportunities. Because participants needed to click on a link to see the research information (rather than seeing it right away), these numbers are likely a conservative estimate of the number of MTurk participants who may be interested in these kinds of materials. In particular, we believe that it is likely that more participants would have read and spent time on the outreach page if it had been on the same page as the token at the end of the study (a design we avoided in our experiment to enable accurate timing).

We also found that participants spent a median time of 41 seconds on the outreach pages we provided – time in which they forgo the opportunity to complete financially compensated HITs. The finding provides empirical evidence that a proportion of MTurk workers are not solely interested in monetary incentives, as suggested in prior work (e.g., [6, 17, 20]).

Perhaps surprisingly, we found that people spent more time on the scientific outreach pages in Experiment 1, which did not pay them for their time on the pages, than in Experiment 3, where they were paid for their time. This suggests a selection bias and differences in motivations. The 40% of participants who voluntarily accessed the outreach pages after having received financial compensation are most likely more motivated to learn additional information than those who are in the midst of completing a HIT. In other words, had we allowed MTurkers to voluntarily seek out the outreach pages in Experiment 3, we would have likely seen greater learning gains than we observed here. It also suggests that outreach pages will be most impactful when participants voluntarily access them even though participants still accurately recall some of the information when the outreach page is part of a paid HIT.

Our second experiment showed that our participants were highly interested in all information about research: research impact, previous research, experimental design, and researchers' motivations for conducting the studies. Their open-ended comments also showed their interest, excitement, and appreciation. These results extend the findings presented in [32], which showed that volunteer participants often state their interest in hearing more about the research background and experimental design decisions. Our work demonstrates that this interest is not unique to a population who self-selects to participate in volunteer-based online experiments or citizen science projects; instead, this interest is also prevalent among participants in online labor markets.

Participants accurately recall research information

Our third study showed that participants remembered information about researchers' motivations for conducting the online studies more than information about the research impact. Additionally, participants demonstrated in their open-ended answers comprehension of the research. Participants in Experiment 2 spent an average of just under 10 seconds (median) on the research information pages, suggesting that they skim the content or read only a small portion at best. Given that participants who voluntarily sought the additional information spent four times longer with the material, it is highly likely that those who actively seek the information will actually learn more than what we observed among participants who were paid to go through the pages as part of a longer HIT.

Scientific outreach is an imperative

In contrast to laboratory studies, online studies are usually unsupervised and often lack opportunities to learn, share, and foster greater social understanding between researchers and participants [23]. Researchers have raised ethical concerns that online participants might not receive adequate information and support after participating in a study [23, 2] and have identified an information asymmetry when participants cannot control the use of the data they contribute [18, 2]. The reframing of MTurk as a promising opportunity for large-scale scientific outreach in a wide variety of domains is one way to work towards reducing the current knowledge and power imbalance between researchers and participants in online labor markets.

The shift from the lab to online platforms such as MTurk does not diminish the importance of researcher-participant communication. Given our findings that complement previous research [32], it is an imperative and even an ethical obligation for researchers conducting online experiments to include scientific outreach pages. Scientific outreach pages are different and more impactful than informed consent pages that may claim to serve a similar purpose. Scientific outreach pages leverage participants' recent experiences in completing online experiments to construct new scientific learning; participants are more likely to seek out additional research information after a study, which can function similarly to a foot-in-the-door [16], than to read information on the consent forms beforehand. It is time for researchers using online experiments to take advantage of the large numbers of participants they can reach to share knowledge, shift attitudes towards science, and meet participants' demonstrated interests and needs.

To support the widespread creation and adoption of effective scientific outreach pages, more work is needed to develop design guidelines as well as tools to minimize the amount of time it may take researchers to create scientific outreach pages. To further embody two-way scientific outreach and communication between researchers and participants, we also imagine the importance of collaboration to create online experiments and post-experiment learning opportunities that take into consideration the needs of both groups.

LIMITATIONS AND FUTURE WORK

Our experiments focused on behavioral and cognitive studies on Mechanical Turk. The samples of participants who are attracted to these studies may be different from those who choose to participate in other kinds of MTurk studies or tasks, such as writing stories to evaluate a Natural Language Processing model. Additional work should evaluate the interest and best practices for these sorts of studies. Furthermore, there might be a specific subset of participants on MTurk who can spend time without earning money to learn about research. Future work will have to investigate the motivations, personalities, and interest of those who choose to learn more compared to others who do not.

We evaluated learning based on short-term comprehension and recall. In the future, we plan to evaluate how deeper forms of learning can occur through scientific outreach on MTurk, and whether it can change attitudes towards science.

We are also interested in evaluating different ways of presenting outreach information and whether this would affect interest and recall for people from diverse backgrounds. We speculate that content that adapts to diverse participants will lead to the greatest learning.

CONCLUSION

Online studies provide an exciting place for learning about research for participants, doing scientific outreach for researchers, and, ultimately, increasing public awareness and engagement with science.

Our landscape sample analysis and the series of three experiments paint a convincing story of the possibility for scientific outreach and learning at scale in online experiments on Mechanical Turk. We identified four main opportunities for learning and found that 40% of participants in online experiments are likely to actively seek out learning. We also found that participants are very interested in research goals and impact, previous research, experimental design, and researchers' motivations. Finally, we found that participants learn from outreach pages. Based on these findings, the domain of financially compensated online experiments is an exciting new research space for the learning at scale community.

ACKNOWLEDGEMENTS

We thank all the Mechanical Turk participants for their data and feedback that made this work possible. We thank Jesse Dodge, Matt Kay, Adam Elder, and Jonathan Bragg for insightful conversations about Bayesian statistics. We also thank Yeaseul Kim for sharing code for the distractor task, the volunteers who helped us improve our quizzes, and Arvind Krishnamurthy for valuable early feedback. This work was partially funded by NSF award #1651487.

REFERENCES

1. American Psychological Association. 2017. Ethical Principles of Psychologists and Code of Conduct. (2017).
2. Benjamin B Bederson and Alexander J Quinn. 2011. Web workers unite! addressing challenges of online laborers. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*. ACM, 97–106.
3. 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), 351–368.
4. Benjamin S Bloom, Max D Engelhart, Edward J Furst, Walker H Hill, David R Krathwohl, and others. 1956. Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain. New York: David McKay Company. *Inc.(7th Edition 1972)* (1956).
5. Nathan Bos, Ann Zimmerman, Judith Olson, Jude Yew, Jason Yerkie, Erik Dahl, and Gary Olson. 2007. From shared databases to communities of practice: A taxonomy of collaboratories. *Journal of Computer-Mediated Communication* 12, 2 (2007), 652–672.
6. 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 *Proc. ACM Conference on Human Factors in Computing Systems*. ACM, 2246–2257.
7. Paul-Christian Buerkner and others. 2016. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* 80, 1 (2016), 1–28.
8. 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.
9. Marissa Burgermaster, Krzysztof Z Gajos, Patricia Davidson, and Lena Mamykina. 2017. The role of explanations in casual observational learning about nutrition. In *Proc. ACM Conference on Human Factors in Computing Systems*. ACM, 4097–4145.
10. Matthew JC Crump, John V McDonnell, and Todd M Gureckis. 2013. Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS one* 8, 3 (2013), e57410.
11. Ruth B Ekstrom, John W French, Harry H Harman, and Diran Dermen. 1976. Manual for kit of factor-referenced cognitive tests. *Princeton, NJ: Educational testing service* (1976).
12. Snehal Neil Gaikwad, Durim Morina, Rohit Nistala, Megha Agarwal, Alison Cossette, Radhika Bhanu, Saiph Savage, Vishwajeet Narwal, Karan Rajpal, Jeff Regino, and others. 2015. Daemo: A self-governed crowdsourcing marketplace. In *Adjunct Proc. of the 28th Annual ACM Symposium on User Interface Software & Technology*. ACM, 101–102.
13. 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.
14. Kotaro Hara, Abi Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey Bigham. 2018. A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. *Proc. ACM Conference on Human Factors in Computing Systems* (2018).
15. J J Horton, D G Rand, and R J Zeckhauser. 2011. The online laboratory: Conducting experiments in a real labor market. *Experimental Economics* (2011).
16. Shih-Wen Huang, Jonathan Bragg, Isaac Cowhey, Oren Etzioni, and Daniel S Weld. 2016. Toward Automatic Bootstrapping of Online Communities Using Decision-theoretic Optimization. In *Proc. of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 582–594.
17. Panagiotis G Ipeirotis. 2010. Demographics of mechanical turk. (2010).
18. Lilly C Irani and M Silberman. 2013. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proc. ACM Conference on Human Factors in Computing Systems*. ACM, 611–620.

19. Eunice Jun, Gary Hsieh, and Katharina Reinecke. 2017. Types of motivation affect study selection, attention, and dropouts in online experiments. In *CSCW*.

20. Nicolas Kaufmann, Thimo Schulze, and Daniel Veit. 2011. More than fun and money. Worker Motivation in Crowdsourcing-A Study on Mechanical Turk.. In *AMCIS*, Vol. 11. 1–11.

21. Yea-Seul Kim, Katharina Reinecke, and Jessica Hullman. 2017. Explaining the Gap: Visualizing One's Predictions Improves Recall and Comprehension of Data. In *Proc. ACM Conference on Human Factors in Computing Systems*. ACM, 1375–1386.

22. Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work. In *Proc. of the 2013 conference on Computer supported cooperative work*. ACM, 1301–1318.

23. Robert Kraut, Judith Olson, Mahzarin Banaji, Amy Bruckman, Jeffrey Cohen, and Mick Couper. 2004. Psychological research online: Report of Board of Scientific Affairs' Advisory Group on the conduct of research on the internet. *American psychologist* 59, 2 (2004), 105.

24. Siou Chew Kuek, Cecilia Paradi-Guilford, Toks Fayomi, Saori Imaizumi, Panos Ipeirotis, Patricia Pina, and Manpreet Singh. 2015. The global opportunity in online outsourcing. (2015).

25. Kate Land, Anže Slosar, Chris Lintott, Dan Andreescu, Steven Bamford, Phil Murray, Robert Nichol, M Jordan Raddick, Kevin Schawinski, Alex Szalay, and others. 2008. Galaxy Zoo: The large-scale spin statistics of spiral galaxies in the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society* 388, 4 (2008), 1686–1692.

26. N Lane. 1997. An open letter to scientists and engineers:"Let's get the word out together about why science matters." National Science Foundation, June [online]. (1997).

27. Jisoo Lee, Erin Walker, Winslow Burleson, Matthew Kay, Matthew Buman, and Eric B Hekler. 2017. Self-experimentation for behavior change: Design and formative evaluation of two approaches. In *Proc. ACM Conference on Human Factors in Computing Systems*. ACM, 6837–6849.

28. David Martin, Benjamin V Hanrahan, Jacki O'Neill, and Neha Gupta. 2014. Being a turker. In *Proc. ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 224–235.

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

30. Kristy Milland. 2015. (2015). `https://docs.google.com/spreadsheets/d/1T3yP_Jo4qELrwsE2NAPNs07L1AWmpAEr9vnhreGJ-K0/edit#gid=1993074859`, last accessed January 18, 2018.

31. Social Psychology Network. 2017. (2017). `http://www.socialpsychology.org`, last accessed August 20, 2017.

32. Nigini Oliveira, Eunice Jun, and Katharina Reinecke. 2017. Citizen science opportunities in volunteer-based online experiments. In *Proc. ACM Conference on Human Factors in Computing Systems*. ACM, 6800–6812.

33. Vineet Pandey, Amnon Amir, Justine Debelius, Embriette R Hyde, Tomasz Kosciolek, Rob Knight, and Scott Klemmer. 2017. Gut Instinct: Creating Scientific Theories with Online Learners. In *Proc. ACM Conference on Human Factors in Computing Systems*. ACM, 6825–6836.

34. Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5, 5 (2010).

35. Katharina Reinecke and Krzysztof Z Gajos. 2015. LabintheWild: Conducting large-scale online experiments with uncompensated samples. In *Proc. ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 1364–1378.

36. SciStarter. 2017. (2017). `http://www.scistarter.com`, last accessed August 20, 2017.

37. M Silberman, Joel Ross, Lilly Irani, and Bill Tomlinson. 2010. Sellers' problems in human computation markets. In *Proc. of the ACM SIGKDD Workshop on Human Computation*. ACM, 18–21.

38. Brian L Sullivan, Christopher L Wood, Marshall J Iliff, Rick E Bonney, Daniel Fink, and Steve Kelling. 2009. eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation* 142, 10 (2009), 2282–2292.

39. 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. (2017).

40. Andrea Wiggins and Kevin Crowston. 2011. From conservation to crowdsourcing: A typology of citizen science. In *Hawaii International Conference on System Sciences*. IEEE, 1–10.

41. Haotian Zhou and Ayelet Fishbach. 2016. The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *American Psychological Association* (2016).