Types of Motivation Affect Study Selection, Attention, and Dropouts in Online Experiments

EUNICE JUN, Computer Science & Engineering, DUB Group, University of Washington
GARY HSIEH, Human Centered Design & Engineering, DUB Group, University of Washington
KATHARINA REINECKE, Computer Science & Engineering, DUB Group, University of Washington

Understanding whether and how motivation affects participation in online experiments is critical because who contributes and how they contribute can affect the validity of findings. Analyzing data from 7,674 participants across three different studies on the volunteer-based online experiment platform LabintheWild, we identified five motivation types for participating: boredom, comparison, fun, science, and self-learning. We found that these motivation types affect study selection, attention, and dropouts. Participants who were highly motivated by boredom paid less attention and were more likely to dropout than those who were motivated by the possibility of contributing to science. We additionally show that motivation can impact study results and suggest how researchers can take participants' motivation into account when designing and analyzing data from volunteer-based online experiments.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**;

Additional Key Words and Phrases: Large-scale online experiments; motivation; online experimentation

ACM Reference Format:
Eunice Jun, Gary Hsieh, and Katharina Reinecke. 2017. Types of Motivation Affect Study Selection, Attention, and Dropouts in Online Experiments. *Proc. ACM Hum.-Comput. Interact.* 1, 1, Article 56 (November 2017), 15 pages. https://doi.org/10.1145/3134691

1 INTRODUCTION

With an increasing number of online experiments contributing data and results to the literature, much research has investigated threats to their internal and external validity —systematic errors that can influence an experiment's reliability and generalizability [2, 3, 10, 24]. Examples of such threats are sample biases stemming from self-selection or participant drop-outs and low-quality data as a result of participants paying insufficient attention to instructions or the task [2, 10]. These threats are particularly difficult to mitigate in an uncontrolled and often anonymous online setting.

One potential factor that can affect the validity of the experimental results is the motivation of the participants. When comparing volunteer participants to student participants who are often "coerced," researchers have found that the volunteers "seemed more interested and cooperative" and put more effort into the task [22]. Those less motivated, especially when they are in a less interesting study or condition, may also be more likely to drop out and/or use strategies to complete the study quickly [29, 35]. However, what is less well understood is whether and

Authors' addresses: Eunice JunComputer Science & Engineering, DUB Group, University of Washington; Gary HsiehHuman Centered Design & Engineering, DUB Group, University of Washington; Katharina ReineckeComputer Science & Engineering, DUB Group, University of Washington.

2573-0142/2017/11-ART56
https://doi.org/10.1145/3134691

how certain *types* of motivation (and not just *levels* of motivation) may predict problematic behaviors. Do people with different motivations self-select into certain tasks? Can types of motivations predict how much attention they spend on experiments or how likely they are to drop out? And does motivation ultimately affect study results? Answering these questions (summarized in Figure 1) may enable experimenters to foresee unwanted behaviors, make more informed decisions for data cleaning and analysis, detect potential sample biases, and identify threats to the validity of their study.

To answer these questions, we analyzed the behavior of 7,674 participants in three online experiments on a volunteer-based online experiment platform, LabintheWild. Similar to ProjectImplicit [26] and TestMyBrain [6], LabintheWild offers a number of scientific experiments that enable volunteer participants to receive personalized feedback in exchange for study participation. We chose to focus our study on a volunteer-based platform over paid online labor markets (e.g., Mechanical Turk) because reasons for participating on LabintheWild are likely to be more diverse and less likely to be dominated by financial reasons.

Our results make the following contributions:

(1) We identified five motivations for participating in online experiments on LabintheWild: boredom, the desire to compare themselves to others, fun, the possibility of contributing to science, and the desire to learn about themselves. The motivations to participate out of boredom and for comparison extend the motivations to have fun, to learn, and to contribute to science found among volunteers on other volunteer-based online communities and citizen science projects (e.g., Wikipedia [20] and GalaxyZoo [27]).
(2) We found that participants' type of motivation varies significantly across studies. For example, one study attracted people who are more motivated to compare themselves to others, whereas another attracted people who are more motivated to learn about themselves. This finding extends recent research on participation bias due to incentives offered [11] and suggests that other aspects of the experiments can also differentially attract who participates. This type of self-selection bias can undermine the generalizability of online experiments as subsets of a population may self-select out of an experiment.
(3) We contribute empirical evidence that motivation types have differential effects on attention, dropouts, and experiment results. These findings indicate that many of the common data cleaning practices, such as removing cases where participants do not pay close attention to instructions or exhibit other forms of satisficing behavior, may undermine the external validity of experimental results. The results also show that people with different types of motivations perform differently, which can again pose a threat to an experiment's validity.

Based on our findings, we make recommendations for experimenters and discuss the need for design guidelines that ensure the validity of experimental results.

2 RELATED WORK AND RESEARCH QUESTIONS

To develop our research questions, we describe prior work on (1) people's motivation to participate in online experiments and other communities and (2) how that translates into behavioral differences.

2.1 Motivation to Participate in Online Experiments

Many studies have explored why people participate on Amazon's Mechanical Turk, an online labor market that researchers in various disciplines use to recruit and compensate participants. The general finding is that while the financial incentives (as primary or additional income) play a major factor in why people participate [12, 14, 31], people also participate for many non-monetary reasons, such as to kill time, for fun, for the challenge, to sharpen skills, and to learn English [12]. Others have more succinctly categorized these motivations as: to kill time, to make extra money, for fun, and because it gave me a sense of purpose [1].

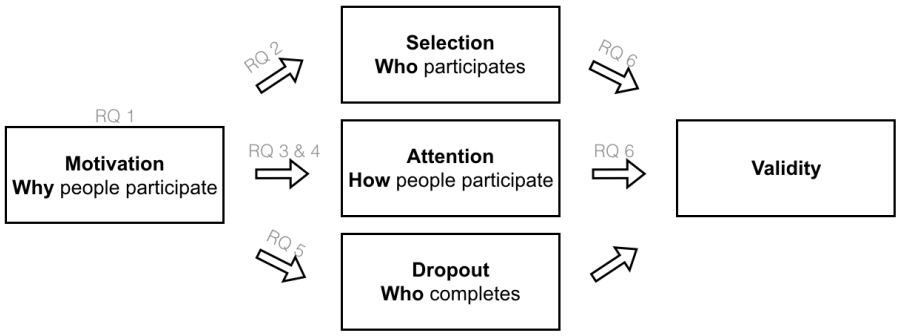

Fig. 1. Theoretical framing of how motivation affects validity. The arrows are labeled with the specific research question that addresses the links between the concepts.

Similar types of motivations have also been identified in community-driven and citizen science projects. Studies of GalaxyZoo and Wikipedia contributors have found motivations such as "for fun" and "to learn" [20, 27, 34]. Firstly, financial motivations often do not play a role. Secondly, there are additional motivations related to contributing and interest in science as these citizen science communities are more specifically focused on science advancement in comparison to Mechanical Turk which is a general labor market.

In contrast to financially-compensated online experiments and citizen science projects, we know little about why people participate on volunteer-based experiment platforms, such as Project Implicit [26], LabintheWild [16], and TestMyBrain [32]. One exception is a study by Reinecke and Gajos [28] who analyzed how people tweet about their own results, sometimes to encourage others to participate. They found four main categories: comparison to others, curiosity about themselves, desire to know if their own characteristics could be predicted, and improvement of particular skills. These ways for sharing are likely linked to their motivation for participating. To extend this literature, our first goal is to directly ask participants for their motivations for taking part in experiments on the volunteer-based experiment platform LabintheWild: **RQ 1: What are people's motivations for participating in experiments on LabintheWild?**

2.2 Motivation and Participation Behaviors

Recent research has shown that motivation affects how participants evaluate incentives offered and how this means that they self-select into different tasks on Mechanical Turk [11]. Motivation is likely to affect participants' choices for participating beyond the incentives offered for completing a task. Our second research question therefore addresses whether a similar self-selection also occurs in a volunteer-based context, such as on LabintheWild. **RQ 2: Does motivation affect whether people self-select into experiments?**

Research on the validity of online experiments has also found that careless or inattentive responses can raise methodological concerns by increasing noise and decreasing validity [15, 23]. This is especially a challenge in the online context where participants may be multitasking or interrupted by their environments [25]. However, using conventional data cleaning techniques, much prior work has shown that Mechanical Turk experiments and those conducted on volunteer-based online experiment platforms can accurately replicate in-lab studies (e.g., [7, 10, 28]).

Oppenheimer et al. found that participants' motivation levels affected whether they passed Instructional Manipulation Checks (IMCs), which were inserted in place of regular study questions to observe whether participants

were closely reading the instructions [23]. We might assume that it is not just the level of motivation that impacts whether participants pay close attention to instructions, but also the *type* of motivation that led participants to take part in a study. **RQ 3: Does motivation affect whether people closely read instructions?**

Similarly, motivation might impact how much participants are willing to exert themselves in an experiment. One strategy to cope with the high cognitive demand of answering survey questions, for example, is to satisfice by "straightlining," i.e., by providing answers to a series of questions in the same place of a rating scale (e.g., all 1s or all 2s) [13]. Straightlining and other satisficing strategies compromise the quality of data from online experiments. A common data cleaning practice is to exclude data that look suspicious of satisficing behavior, which could also lead to exclusion of legitimate data, another possible threat to validity.

While there are other strategies to decrease the cognitive effort when answering rating scale questions, prior work suggests that those who are less motivated are more likely to use straightlining [35]. Here we seek to explore if certain motivation predicts the use of straightlining. **RQ 4: Does motivation affect if people satisfice?**

Another methodological concern impacting the validity of online experiments is participant mortality — whether participants drop out of an experiment [4]. Unlike participation bias, which is noncompliance at the start of the study, dropout is about the decision not to complete a study after starting it. Neither is random, so both can lead to considerable sample biases [4]. Prior work showed that participants in an online experiment who were assigned to a less motivating, boring, or very difficult experimental condition were more likely to drop out [29, 30]. This prior work primarily focused on incentives, examining how an increase in the incentives can make tasks more intrinsically motivating. We hope to additionally find out how motivation affects whether volunteers complete an experiment even without manipulating any incentives. **RQ 5: Does motivation affect the likelihood of a participant dropping out of an online experiment?**

If motivation indeed affects who participates and how people participate, we have to assume that it could also affect the results of the study, and thus, study validity. If it does, this would suggest that experimenters need to ensure to equally attract participants with different motivations. **RQ6: Does motivation ultimately affec t study validity?**

3 EXPERIMENTAL METHODS

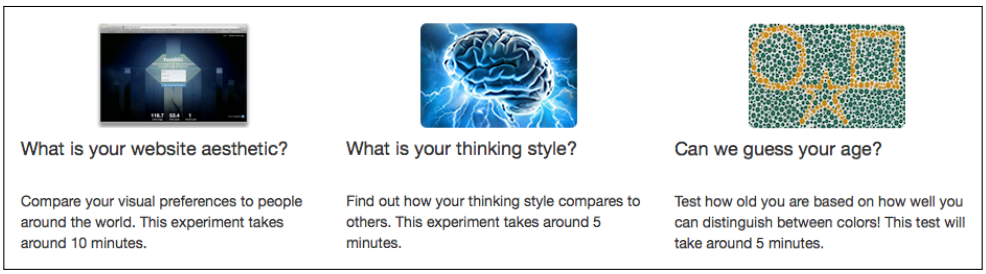

Fig. 2. Experiment teasers used to advertise the three experiments on the LabintheWild platform.

To answer our research questions, we conducted three experiments on LabintheWild, a large volunteer-based experiment platform that usually hosts between 9 and 12 experiments and attracts an average of 1,000 participants from numerous countries each day. Participants are volunteers who, instead of receiving financial compensation, receive personalized feedback at the end of each online experiment. Participants hear about the platform through online social media (e.g., Twitter and Facebook), news articles, and/or word of mouth. To recruit and maintain a diverse sample of volunteers and reduce the barrier to participation, LabintheWild does not require participants to create accounts.

We launched three experiments that were designed to represent a diverse set of common psychology and HCI studies (see Figure 2 for the corresponding experiment teasers):

(1) *Visual preferences:* A subjective experiment collecting Likert scale ratings on website images, advertised as "What is your website aesthetic? Compare your visual preferences to people around the world."
(2) *Thinking style:* A choice experiment, which asked participants to select related pairs from word triplets, advertised as "What is your thinking style? Find out how your thinking style compares to others."
(3) *Color vision:* An experiment collecting participants' responses to a set of color stimuli, advertised as "Can we guess your age? Test how old you are based on how well you can distinguish between colors!"

Each advertisement slogan also noted the length that an experiment could be expected to take. The visual preferences experiment was estimated to take 10 minutes while the other two were estimated to take 5 minutes.

All three experiments were online and on the LabintheWild homepage at the same time.

3.1 Assessing LabintheWild Motivations

To explore potential reasons for participating on LabintheWild, we included a preliminary survey at the beginning of each of our three studies (after the informed consent). In line with research on motivation on Mturk [25], we asked participants an open-ended question: "What is your motivation for participating in an experiment on LabintheWild?". The question was intentionally placed before starting the actual study and referenced the overall platform in order to elicit volunteer motivations to participate on LabintheWild as a whole as opposed to a specific study. After collecting initial responses from 178 participants, two coders thematically analyzed the open-ended responses [8] and identified five predominant types of motivation: (1) An interest in supporting science, (2) an interest in learning about themselves, (3), an interest in comparing themselves to others, (4) because of boredom and (5) for fun. Note that the first three motivation types have been identified in prior work on volunteer-based online experiments; these motivations were extracted from participants' tweets about their results [28]. Our additions of out of boredom and for fun did not come up in their analysis of tweets, most likely because this prior work focused on participants' public discussions. However, fun and boredom as motivational reasons have been identified on Mechanical Turk and in citizen science projects where participants were directly surveyed or interviewed [1, 12, 21].

After identifying these five motivation types, we modified all experiments to include the single question "To what extent are you participating in an experiment on LabintheWild for the following reasons?" followed by five Likert scale items whose order was randomized across participants: (1) "I want to help science," (2) "I want to learn about myself," (3) "I want to compare myself to others," (4) "I am bored," and (5) "For fun."

3.2 Measures

We collected the following data:

- *Initial manipulation check (IMC):* To test whether participants were closely reading the instructions at the beginning of the study, we placed an attention check on the IRB consent form (see Figure 3). Participants were asked to click on the last word of one of the sentences. This proxy for attention was measured as a binary variable (1=clicked on the last word, 0=did not click on the last word).
- *Self-reported motivation:* After participants gave their consent to participate on the IRB consent form, they were asked to give their ratings on five 5-point Likert scale questions in randomized order in response to the question "To what extent are you participating in an experiment on LabintheWild for the following reasons?", as described above.
- *Dropout:* Dropout was coded as a binary variable based on data we logged for study completion. Participants were considered to have completed the study if they proceeded beyond the informed consent page (to ensure they accepted being in the study) and reached the final page of the experiment showing their personalized feedback.

- *Straightlining:* To assess whether participants were satisficing, we used the 4-item Intrinsic Motivation Inventory (IMI) from [19] because it is a well-tested inventory, is comprised of questions that are easy to answer, and would not raise suspicions that participants' attention was being tested. We included the inventory at the end of the study before participants proceeded to the personalized feedback page to test whether they were paying sustained attention even after a 5-10 minute task. The IMI asks the question "For each of the following statements, please indicate how true it is for you:," followed by four 7-point Likert scale items labeled with "This task was fun to do.", "I thought this was a boring task." (reverse question), "I would describe this task as very interesting.", and "I thought this task was quite enjoyable." Straightlining was coded as a binary variable (1=same ratings across the four Likert items, 0=different ratings on at least one of the four items).

Additionally, we also collected demographics information, including age and gender, and trials data that participants gave directly as part of each online experiment. We collected age and gender data because previous work has shown that age and gender affect civic engagement [27, 33] and online contributions [9, 17].The instructional manipulation check was added later and this data is only available for 1,999 participants.

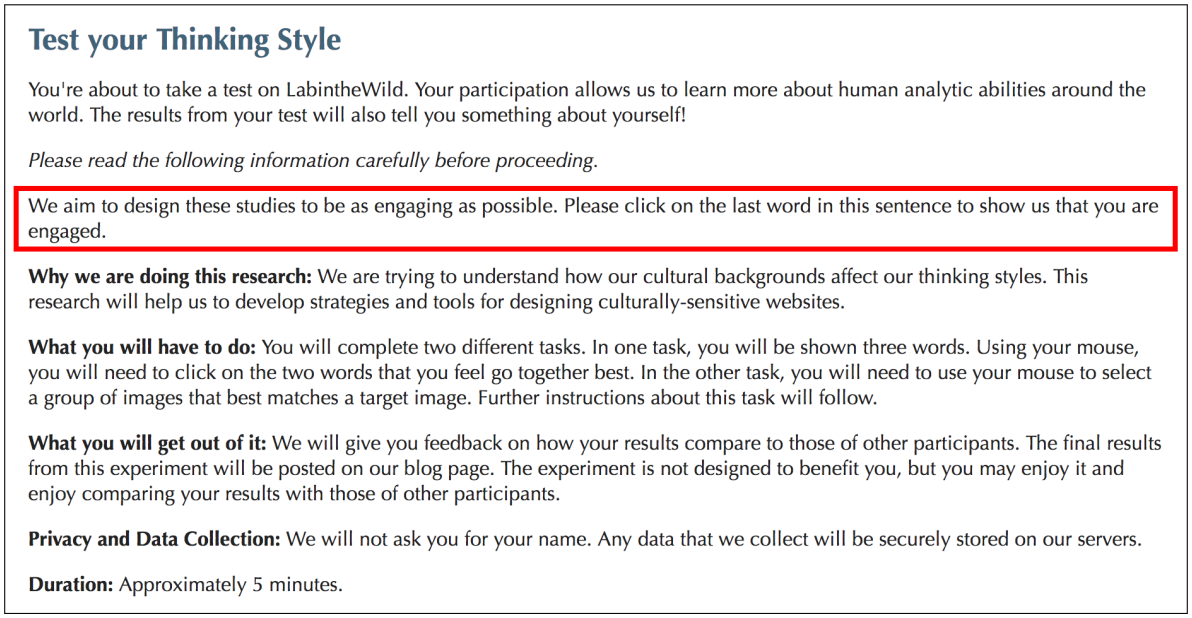

Fig. 3. As an initial attention check, the IRB consent forms of all three experiments (here shown for the Thinking Style experiment) were modified to include an additional sentence asking people to prove that they were reading closely.

3.3 Participants

We collected data for approximately three months in 2016/2017, during which, across the three experiments, 7,868 participants took part. The participants came from 117 different countries. Participants also represented a diversity of education levels, ranging from pre-high school to college/professional school to postdoctoral studies. The education level of participants was roughly the same across studies with approximately a third of participants

reporting to have at least some college education. Table 1 gives more details of participant makeup across all three studies.

We excluded 194 participants who reported to have taken the study before. The final data set consisted of 7,674 participants (53% female, 2% declined to answer or considered themselves as "other"). The mean age of participants was 24 years (sd = 14.5 years).

Table 1. Overview of self-reported age and gender in the three experiments (excluding participants who retook an experiment).

	Visual Preferences	Thinking Style	Color Vision
N	1,165	3,347	3,162
Mean age (SD)	25 (11)	27 (12)	20 (18)
% female	50	54	52
% male	46	44	47

3.4 Analysis

Our first research question was analyzed using thematic analysis as described earlier.

To answer our second research question (self-selection), we conducted three logistic regression models (one model for each study). For each model, the dependent variable is coded as true if the participant participated in the study, false if they participated in the other two studies. The independent variables were the five motivation types.

To answer the research questions 3-5, we conducted a series of logistic regression models with participants' ratings on the five motivation types and the study name as independent variables. The dependent variable was either the binary measure for IMC (RQ3), straightlining (RQ4), or dropouts (RQ5), depending on the focus of the analysis. We also tested models with age and gender as control variables because prior work has identified that they might affect motivations to contribute online [9, 17].

The color vision study did not ask participants for their age until the very end (resulting in missing age data for many participants), so we exclude this study from any analyses that include age. For those analyses that include gender, we excluded 153 participants who declined to specify their gender or chose "other" because their numbers were too low to calculate statistically stable models.

When presenting the effect of the independent variables in the logistic regressions, we will often present the beta coefficients. These beta coefficients are the odds ratio and represent the constant effect of the predictor variable on the likelihood of the outcome to occur. For example, when interpreting boredom's effect on the likelihood for failing the IMC, the odds ratio of 1.44 means that for each one point increase in the boredom rating (collected on a 5-point scale), participants are 1.44 times more likely to fail.

4 RESULTS

4.1 RQ 1: What are people's motivations for participating in experiments on LabintheWild?

The results of our thematic analysis showed that there are five motivation types for participating on LabintheWild: out of boredom, to compare oneself to others, for fun, to contribute to science, and to learn about oneself (hereafter self-learning). These motivation types are only weakly correlated (Pearson's r: $-.14$ to .34), suggesting that they represent distinct reasons for participating but that volunteers usually have multiple motivations for participating in experiments on LabintheWild.

We also found that age was positively correlated with self-learning and science (Pearson's r .06 and r .09, $p < .001$) and negatively correlated with boredom and fun (r $-.25$ and r $-.07$, $p < .001$). The correlation between age and the motivation to compare was not significant.

Recognizing that participants have multiple motivation types, we did not superficially want to categorize participants as belonging to any one motivation group. Instead, we included all five motivation ratings in the models in the following analyses, thereby controlling for the influence of other motivations.

Table 2. The mean scores and standard deviations (in parentheses) for the five different motivation types from participants in each study. The mean scores are based on self-reports that ranged from 1, "not at all", through 6, "very much."

	Visual Preferences	Thinking Style	Color Vision
Boredom	3.1(1.4)	3.1(1.4)	3.1(1.4)
Comparison	3.4(1.3)	3.4(1.3)	3.1(1.4)
Fun	4.1(1.1)	4.2(.9)	4.1(1.0)
Science	3.8(1.2)	3.7(1.2)	3.6(1.3)
Self-learn	4.1(1.1)	4.5(.8)	4.0(1.1)

Table 3. Summary of the results of our research question RQ2 on whether people self-select into experiments. All coefficients are presented as odds ratios. An odds ratio greater than 1 means that the outcome variable is more likely to occur, whereas an odds ratio less than 1 means that the outcome variable is less likely to occur. *p < .05, **p < .01, ***p<.001, ****p<.0001

	Visual Preferences vs Others	Thinking Style vs Others	Color Vision vs Others
Boredom	0.99	1.03	0.98
Compare	1.09**	1.04	0.92****
Fun	0.91**	0.88****	1.19****
Science	1.13****	0.93**	1.01
Self-learn	0.82****	1.64****	0.71****

4.2 RQ 2: Does motivation affect whether people self-select into experiments?

Table 2 shows the means and standard deviations of motivation ratings in each of the three studies. Table 3 shows the results of three logistic regression models (one model for each study) that analyze whether people self-select into experiments depending on their motivation.

We found significant differences in motivation types across experiments. In particular, self-learning stands out as the strongest motivator in predicting self-selection into the thinking style experiment, which was advertised as "What is your thinking style?". For the visual preferences experiment ("What is your website aesthetic?"), we found an interest in science to be the strongest motivator, albeit closely followed by a motivation to compare to others. The strongest motivator for participating in the color vision experiment, advertised as "Can we guess your color age?", was fun.

Table 4. Summary of our results from our analyses to answer research questions 3-6. Research questions 3-5 were explored using two models, one without age and gender and one including age and gender (available for a subset of studies). The coefficients for RQs 3-5 are presented as odds ratios (the greater the odds ratio value, the greater the likelihood is of the measured phenomenon occurring). Research question 6 was explored using only one model without age and gender. The coefficients for RQ 6 are presented as standardized beta coefficients. *p < .05, **p < .01, ***p<.001

	RQ 3: Failed IMC	RQ 3: Failed IMC	RQ 4: Straightlining	RQ 4: Straightlining	RQ 5: Dropout	RQ 5: Dropout	RQ 6: Visual Preferences	RQ 6: Thinking Style	RQ 6: Color Vision
Boredom	1.44***	1.40***	1.2***	1.2***	1.06*	1.03	0.01	-0.02	-0.01
Compare	0.92	0.92	1.14**	1.15**	0.98	0.96	0.09*	0.01	0.06**
Fun	0.92	0.99	0.84**	0.83**	0.92*	1	0.07	0.01	0.07***
Science	0.85*	0.94	0.9*	0.88**	0.85***	0.85*	0.05	-0.03	0.08***
Self-learn	1.14	1.19	0.76**	0.76***	0.88***	1	-0.03	-0.0005	0.03
Visual study	0.72*	0.76*	1.03		1.99***	5.81***			
Thinking study	1.14		0.99		0.41***				
Age		0.97**		1		1			
Gender (female)		0.51**		1.11		0.99			

These results suggest that different experiments on LabintheWild attract people with different motivations for participating on the platform. In other words, people self-select into an experiment depending on their motivation. As we will discuss later, the slogan used to advertise the experiment on the platform could explain these differences between experiments.

Practically, the differences of study participation based on motivation across studies necessitate that study is an important variable to control for when examining the effects of motivation on attention and dropout, which we do for the remaining analyses.

4.3 RQ 3: Does motivation affect whether people closely read instructions (IMC)?

Only 7.6% of participants (between 6.4% and 10.9% depending on the experiment) passed the instructional manipulation check on the consent form, suggesting that only a minority of participants closely read the instructions on that page. However, the type of motivation influences who is more likely to pass the check. Table 4 shows that for each point increase in boredom as a motivation, participants were 1.44 times more likely to fail the attention check ($p < .001$). Including age and gender in the model also shows that for each one-year increase in age, participants are 97% as likely to fail the IMC. In other words, participants are more likely to pass the IMC with an increase in age. Males are also 1.8 times more likely to fail the attention check compared to female participants with the same motivations and age.

4.4 RQ 4: Does motivation affect if people satisfice?

A relatively low 10% of participants (between 9.76% and 10.49% depending on the experiment) satisficed in the final survey section by giving the same responses to all questions (and thereby ignoring reverse questions). Table 4 shows an overview of the model results. Participants were more likely to straightline if they were motivated by boredom and comparison (log odds are 1.2 and 1.14, respectively, with $p < .01$). In contrast, participants motivated more by fun, science, and self-learning were less likely to satisfice (log odds are 0.84, 0.9, and 0.76, respectively, $p < .05$). The effects persisted even when controlling for age and gender.

4.5 RQ 5: Does motivation affect the likelihood of a participant dropping out of an online experiment?

About 20% of the participants who started the experiments did not complete them. Dropout rates varied quite significantly across experiments. The thinking style experiment had the lowest dropout rate at 10%, color vision had 24%, and the visual preferences experiment had the highest at 38%. Potential explanations for this high divergence are differences in length and tasks between the studies: For one, the visual preferences experiment took twice as long (10 minutes) as the other two experiments. The experiments also differed in their tasks, which alternated between two different versions in the thinking style experiment (with the lowest dropout rate), but remained the same for the other two.

In addition to such differences between studies, we found that motivation type significantly predicted dropouts. In particular, people motivated by boredom were 1.06 times ($p < .05$) more likely to drop out than others (see Table 4). People motivated by science were least likely to drop out (odds ratio = 0.85, $p < .001$). When age and gender were included in the model, only motivation for science remained a significant negative predictor of dropout. For each point increase in motivation for science, participants were 1.17 times more likely to complete an online experiment.

The absence of main effects of age and gender on participants' likelihood to dropout does not discredit the impact age and gender have on dropout. Motivation and age are correlated, so the observation that only the motivation to contribute to science remains a significant predictor of dropout after controlling for age and gender suggests that age is also an important predictor of dropouts. The younger a participant, the higher the likelihood of dropping out (and the more likely they are to be motivated by boredom).

4.6 RQ 6: Does motivation ultimately affect study validity?

To assess the effect of motivation on study validity, we conducted three linear regressions with the five motivation types as independent variables predicting the performance variable, which was unique to each study, as the dependent variable. Table 4 has a column per study showing the impact of motivation on study outcomes.

For the visual preferences study, participants rated the subjective appeal of various websites twice on a 9-point Likert scale. We used the consistency between the first and second ratings as the performance variable (ranging from 0=same rating both times to 8=one extreme rating to another) since the consistency of appeal ratings was the research question in [18] where the study originally appeared. The coefficients in Table 4 show the effect of motivation on rating difference in terms of the standard deviation of rating differences. Motivation to compare had a significant but weak effect (β 0.09) on the average absolute difference between the two ratings participants gave each website stimulus. The more participants were motivated by comparison, the less consistent they were in their ratings. The other motivation types had no effect on rating consistency.

In the thinking style study, participants' performance measure was their overall score, which indicated how holistic or analytic they think. Motivation did not have any effect on the thinking styles participants exhibited.

The color vision study asked participants to indicate the orientation of 84 stimuli that changed in color and contrast from the background. A perfect score was 84 correct answers. We found that an increase in motivation for comparison, fun, and science lead to a statistically significant, but again, weak, increase in accuracy (β 0.06, 0.07, 0.08, respectively). The motivations to participate out of boredom and self-learning did not have statistically significant effects on accuracy.

Since our previous findings showed that motivation affects how much attention participants pay, we also investigated potential changes in study outcomes as a result of common data cleaning practices that would remove participants who satisfice and/or fail instructional manipulation checks [23]. Out of the three experiments, only the results of the color vision study changed between groups of participants who passed or failed the attention check and between groups of participants who did or did not straightline. Using the same dependent variable as above and

t-tests for samples of unequal variance, we found that those who passed the IMC gave more accurate responses (M=57.4 vs. M=55.2, $t97.7\ -2.80, p < .01$, Cohen's $d\ .28$), and that those who did not straightline also gave more accurate responses (M=56.0 to M=53.8, $t181.1\ -2.41, p < .05$, Cohen's $d\ .22$) than participants who did not pass the IMC or straightlined. The small to medium effect sizes suggest that satisficing behaviors are indeed important to consider; whether people pass these attention checks depends on their motivations and can ultimately affect experimental conclusions if experimenters follow standard data cleaning procedures.

5 SUMMARY AND DISCUSSION

Our findings show that motivation affects who participates in online experiments and how they participate. These, in turn, have strong implications for the validity of online experiments (see also Figure 1).

[RQ 1] Our results showed that volunteers have five types of motivations to participate in experiments on LabintheWild: out of boredom, for comparison to others, for fun, for self-learning, and to benefit science. Our finding that participants are motivated by fun and to learn were also found in previous work on volunteers' motivations for participating in Wikipedia and GalaxyZoo [20, 27, 34], suggesting that these types of motivation are general to communities dependent on volunteers. These similar types of motivations have also been identified in prior work on studying participation motivation on Mechanical Turk [1, 12, 14, 31], with the key exception being the "for science" motivation that we identified in our work. This is likely due to the fact that LabintheWild is specifically designed for online experiments, whereas Mechanical Turk is broadly framed as an online marketplace for work. Because of similar motivation types between LabintheWild and Mechanical Turk, it is likely that our results will generalize to other experiment platforms.

[RQ 2] Participants' motivations also influenced which studies they self-select into. The result is especially interesting because it suggests that different aspects of the study design, including the recruitment strategy and marketing of the experiments, may systematically encourage or discourage participation. For example, one likely reason that the visual preferences study attracted significantly more participants who were motivated by comparison may be due to its slogan, which explicitly invites people to "compare...to people around the world." Similarly, the thinking style experiment may have attracted significantly more self-learning participants because it allows participants to "find out" about their thinking style. While our experiments were not intended to produce design guidelines for the marketing of experiments, our results suggest that such guidelines are needed in order to understand the impact of such design decisions on the experiment's validity.

[RQ 3 + 4] Motivation affected how attentive participants were. The starkest contrast was between participants motivated by boredom and science. Participants motivated by boredom were less likely to pass an early attention check and more likely to satisfice survey responses than those motivated by science. The results of passing the attention check and satisficing are related. Participants motivated by science may pay significantly more attention and exert more effort rather than satisfice. On the other hand, participants motivated by boredom may pay significantly less attention (and therefore are more likely to fail the initial attention check) and exert less effort and satisfice. These results on attention and satisficing are important because they suggest that motivation can be used as a tool in data cleaning and analysis as we will discuss in the next section.

[RQ 5] Motivation also impacted who dropped out of online experiments. Regardless of experimental task and duration, participants motivated by boredom were more likely to drop out than participants motivated by science. Our results affirm previous work that participant mortality is systematic, albeit sometimes covert [4]. The effect of motivation on dropout, as observed in the odds ratio of 1.99 for dropouts in the visual preferences study (see Table 4) was comparable to the effect size that [5] found of offering financial incentives to people who completed a demographic questionnaire online.

The finding that people who are motivated by science are more likely to complete a study might suggest that experimenters should primarily appeal to a population interested in contributing to science. However, such a sample bias could affect the overall results, and thus, the experiment's validity, as we found in answering RQ 6.

[RQ 6] Motivation predicted outcomes for each of the three studies differently. When examining the relationship between motivations and study outcome measures, we found significant, but generally weak relationships. For example, participants motivated by comparison were less consistent in their visual preference ratings. This is perhaps due to their desire to compete, leading them to respond in a way that maximizes some unknown objective (i.e., they did not respond truthfully). And thus, when asked the second time, they could not re-produce their initial ratings, resulting in the high inconsistencies. We also found that in the color vision study with objective tasks, participants motivated by comparison, fun, and science were more accurate in their color perception than participants motivated by boredom or self-learning. Participants motivated by boredom likely did not exert enough mental effort to answer as accurately as others while participants motivated by self-learning may have intentionally answered incorrectly at times to see how good their worst performance could be.

Our additional analyses of satisficing behaviors also showed that satisficing behaviors, which are correlated with participants' motivations, can lead to differences in study outcomes. Those who exhibited satisficing behaviors performed worse in the color vision study (i.e., they answered fewer questions correctly) than those who paid more attention. We only saw a difference in performance in the color vision experiment, perhaps largely because the study requires a higher level of attention than the other two experiments.

We conclude that the ultimate effect motivation has on experimental results depends on the task and the nature of the independent variables (whether motivation moderates the effect of the independent variables). For instance, if the experimental task is completely subjective and the trials do not build on each other, attention and motivation may not matter as much as in experiments where future trials are dependent on previous trials and each trial requires a high level of mental effort.

Altogether, this work contributes empirical evidence of how motivation affects study validity and adds to the body of research on how levels of motivation affect experiment behavior (e.g., [29, 35]). We show that considering motivation *types* – beyond motivation *levels* – can be insightful. Participant's study-selection, attention and mental effort, and likelihood to dropout depend on *what* motivates them, not just *how* motivated they are overall. Motivation types can affect study outcomes and validity.

6 PRACTICAL IMPLICATIONS

Our results have several practical implications for both researchers using online experiments and designers of online experiment platforms.

Based on our findings that motivation affects how volunteers participate in online experiments through their attention and question answering behavior, we recommend researchers use motivation data to support their data analysis. Researchers could collect participants' motivation data along with other demographic data in order to differentiate between a participant's lack of attention and poorly written questions or tasks. For instance, if participants with a wide range of motivation types satisfice a particular survey or question, researchers may be dissuaded from presumptuously excluding this data and persuaded to examine the reliability of the survey or question item.

The use of motivation as a tool for data analysis also opens the possibility of potentially using motivational surveys, possibly at various points of a study, as alternatives to attention checks in online experiments, which by inclusion imply that researchers think the participants are inattentive and may compromise the rapport between researcher and participant [23]. In order to reduce the complication of data cleaning and analysis for online experiments and increase confidence in data quality and trust, we recommend that researchers collect motivation data before online experiments to help them make more sense of suspicious, problematic data.

Based on our finding that motivation affects who participates, we suggest strategies to improve sample representativeness. We recommend researchers and platform designers to diversify their marketing efforts (e.g., by rotating through different slogans). For example, our results indicate that specific keywords, such as "compare" or "find out" may predominantly attract people interested in comparing themselves or learning about themselves. While more research is needed to find out how exactly the phrasing of slogans impacts participation of specific groups, we suggest that platform designers should provide a way to auto-rotate through a variety of slogans. In addition, if a study predominantly attracts participants who are motivated by boredom, researchers may have a difficult time retaining participants (high dropouts) and obtaining enough data. A more comprehensive tactic to mitigate self-selection and improve dropout could be to take participant motivations and demographics into consideration and tailor both slogans and the messages throughout an online experiment to a specific participant.

7 LIMITATIONS AND FUTURE WORK

As with most studies, future work is needed to explore whether the results are generalizable. Because the motivation types uncovered are similar to those that other researchers have found on related platforms used for science, (e.g., Mechanical Turk and GalaxyZoo), and the participation behaviors studied (i.e., selection, attention, and dropout) are common behaviors in both experiments and non-experimental descriptive survey research, we believe that our results can generalize. However, additional work is needed.

In this work, we chose to examine the effect of motivation on three existing online studies. This limited our experimental control in a number of dimensions related to the studies, such as slogan used and nature of the tasks. In our analyses, we controlled for these factors together in the study variable, but future work is needed to further tease apart these aspects of a study that could be interacting with motivation in affecting experimental behavior. For instance, we can imagine future work that employs Natural Language Processing metrics and techniques to capture the semantics of slogans to more precisely understand what aspects of slogans referring to online experiments and their duration appeal differentially to participants with certain motivation types.

Because LabintheWild does not track participants across the studies, we could not account for changes in proximate motivation and behavior due to study participation. For instance, it is reasonable to think that a volunteer who has completed one or more experiments before taking one of the experiments in this study is more likely to skim the IRB page or surveys and thereby less likely to pass the IMC and more likely to straightline.

By considering motivation in the design and analysis of online experiments, researchers can more accurately assess the internal and external validity of their findings. Although the first step is to recognize the impact of motivation on the validity of an increasingly popular research methodology, we believe that the next step would be to try interventions that change participant motivations to improve validity and then derive design guidelines to support more robust and valid online experiments. We are excited to explore the directions outlined above.

8 CONCLUSION

In this paper, we provide empirical evidence that motivation affects which experiment participants select, whether they pay attention at the beginning and the end of an experiment, and whether they will drop out in volunteer-based online experiments.

We contribute identification of five motivation types that compel volunteers to participate in online experiments on LabintheWild, evidence for how motivation types could affect sample representativeness, links between motivation, attention, and dropout that can compromise data quality and generalizability of results, a discussion of how common data cleaning practices without considering motivation can be flawed, and an early idea for ways to intentionally incorporate motivation measures in online experiments as proxies for attention. We argue for the inclusion of motivation in demographics questionnaires and the need for more research on how *types* of motivation affect

experimental behaviors. Based on our findings and existing theory, we conclude that participant motivation must be considered in the design and analysis of data from online experiments.

9 DATA SETS

The data sets used for the analysis of research questions 2-5 including participants' demographics can be accessed at http://www.labinthewild.org/data.

10 ACKNOWLEDGEMENTS

We greatly thank the LabintheWild participants who make this research possible. We also thank the reviewers for their time and input.

REFERENCES

[1] Judd Antin and Aaron Shaw. 2012. Social desirability bias and self-reports of motivation: a study of amazon mechanical turk in the US and India. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2925–2934.
[2] Adam J Berinsky, Gregory A Huber, and Gabriel S Lenz. 2012. Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk. *Political Analysis* 20, 3 (2012), 351–368.
[3] Michael H Birnbaum. 2004. Human research and data collection via the Internet. *Annu. Rev. Psychol.* 55 (2004), 803–832.
[4] Chris Fife-Schaw. 2006. Quasi-experimental designs. *Research methods in psychology* (2006), 88–103.
[5] Andrea Frick, MT Bächtiger, and Ulf-Dietrich Reips. 2001. Dimensions of Internet science. (2001).
[6] 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.
[7] 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.
[8] Barney G Glaser and Anselm L Strauss. 2009. *The discovery of grounded theory: Strategies for qualitative research*. Transaction publishers.
[9] Benjamin Mako Hill and Aaron Shaw. 2013. The Wikipedia gender gap revisited: characterizing survey response bias with propensity score estimation. *PloS one* 8, 6 (2013), e65782.
[10] John J Horton, David G Rand, and Richard J Zeckhauser. 2011. The online laboratory: Conducting experiments in a real labor market. *Experimental economics* 14, 3 (2011), 399–425.
[11] Gary Hsieh and Rafał Kocielnik. 2016. You get who you pay for: The impact of incentives on participation bias. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 823–835.
[12] Panos Ipeirotis. 2008. Mechanical turk: The demographics. *A Computer Scientist in a Business School* 19 (2008).
[13] Olena Kaminska, Allan L McCutcheon, and Jaak Billiet. 2010. Satisficing among reluctant respondents in a cross-national context. *Public Opinion Quarterly* 74, 5 (2010), 956–984.
[14] 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.
[15] Aniket Kittur, Ed H Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 453–456.
[16] LabintheWild. 2017. (2017). http://www.labinthewild.org, last accessed April 22, 2017.
[17] Cliff Lampe, Rick Wash, Alcides Velasquez, and Elif Ozkaya. 2010. Motivations to participate in online communities. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 1927–1936.
[18] Gitte Lindgaard, Gary Fernandes, Cathy Dudek, and Judith Brown. 2006. Attention web designers: You have 50 milliseconds to make a good first impression! *Behaviour & information technology* 25, 2 (2006), 115–126.
[19] Edward McAuley, Terry Duncan, and Vance V Tammen. 1989. Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A confirmatory factor analysis. *Research quarterly for exercise and sport* 60, 1 (1989), 48–58.
[20] Oded Nov. 2007. What motivates wikipedians? *Commun. ACM* 50, 11 (2007), 60–64.
[21] Oded Nov, Ofer Arazy, and David Anderson. 2011. Technology-Mediated Citizen Science Participation: A Motivational Model.. In *ICWSM*.
[22] William Oakes. 1972. External validity and the use of real people as subjects. *American Psychologist* 27, 10 (1972), 959.

[23] Daniel M Oppenheimer, Tom Meyvis, and Nicolas Davidenko. 2009. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology* 45, 4 (2009), 867–872.
[24] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5, 5 (2010).
[25] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. Running experiments on amazon mechanical turk. (2010).
[26] ProjectImplicit. 2017. (2017). http://www.projectimplicit.net, last accessed April 22, 2017.
[27] M Jordan Raddick, Georgia Bracey, Pamela L Gay, Chris J Lintott, Phil Murray, Kevin Schawinski, Alexander S Szalay, and Jan Vandenberg. 2009. Galaxy zoo: Exploring the motivations of citizen science volunteers. *arXiv preprint arXiv:0909.2925* (2009).
[28] Katharina Reinecke and Krzysztof Z Gajos. 2015. LabintheWild: Conducting large-scale online experiments with uncompensated samples. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 1364–1378.
[29] Jochen Musch Ulf-Dietrich Reips. 2000. A brief history of Web experimenting. *Psychological Experiments on the Internet* (2000), 61.
[30] Ulf-Dietrich Reips. 2009. Internet experiments: Methods, guidelines, metadata. In *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 724008–724008.
[31] Joel Ross, Lilly Irani, M Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who are the crowdworkers?: shifting demographics in mechanical turk. In *CHI'10 extended abstracts on Human factors in computing systems*. ACM, 2863–2872.
[32] TestMyBrain. 2017. (2017). http://www.testmybrain.org, last accessed April 22, 2017.
[33] Lori M Weber, Alysha Loumakis, and James Bergman. 2003. Who participates and why? An analysis of citizens on the Internet and the mass public. *Social Science Computer Review* 21, 1 (2003), 26–42.
[34] Heng-Li Yang and Cheng-Yu Lai. 2010. Motivations of Wikipedia content contributors. *Computers in human behavior* 26, 6 (2010), 1377–1383.
[35] Chan Zhang and Frederick Conrad. 2014. Speeding in web surveys: The tendency to answer very fast and its association with straightlining. In *Survey Research Methods*, Vol. 8. 127–135.