Pay Attention, Please: Formal Language Improves Attention in Volunteer and Paid Online Experiments

Tal August
University of Washington
Seattle, Washington
taugust@cs.washington.edu

Katharina Reinecke
University of Washington
Seattle, Washington
reinecke@cs.washington.edu

ABSTRACT

Participant engagement in online studies is key to collecting reliable data, yet achieving it remains an often discussed challenge in the research community. One factor that might impact engagement is the formality of language used to communicate with participants throughout the study. Prior work has found that language formality can convey social cues and power hierarchies, affecting people's responses and actions. We explore how formality influences engagement, measured by attention, dropout, time spent on the study and participant performance, in an online study with 369 participants on Mechanical Turk (paid) and LabintheWild (volunteer). Formal language improves participant attention compared to using casual language in both paid and volunteer conditions, but does not affect dropout, time spent, or participant performance. We suggest using more formal language in studies containing complex tasks where fully reading instructions is especially important. We also highlight trade-offs that different recruitment incentives provide in online experimentation.

CCS CONCEPTS

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

KEYWORDS

online experiments; language formality; participant engagement; data quality; attention

CHI 2019, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300478

ACM Reference Format:
Tal August and Katharina Reinecke. 2019. Pay Attention, Please: Formal Language Improves Attention in Volunteer and Paid Online Experiments. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3290605.3300478

1 INTRODUCTION

While online experimentation offers many benefits for collecting data, researchers still struggle with participant engagement. Ensuring participants read instructions fully, complete the experiment, and provide high quality responses are key in drawing reliable results from the data [30, 35]. One important design consideration for studies is the wording of questions and instructions, which can impact participant responses [1, 44]. Previous work on the language of questions and instructions mostly focused on type of instructions or question framing (i.e., emotive or biased survey questions) [44]. However, more subtle changes in language *style* can also affect user behavior [6, 11, 22, 48].

One of the most commonly studied stylistic dimensions of language is formality, which can convey politeness, authority, amount of shared context, and social distances [17, 21]. Language styles associated with formality, such as politeness and rudeness, can significantly affect user behaviour online [5, 9]. For example, Wikipedia editors using polite language are more likely to attain positions of authority than those who do not [9], and language that is more restrained and thankful, or refers to authority, improves a Kickstarter campaign's chances of being funded [10, 32]. It is therefore possible that language formality also influences participant behavior in online studies. Does formality affect participant engagement in online studies, whether participants read instructional text, complete studies, and exert themselves?

We explored this question by varying the formality of an online experiment's written instructions and evaluating how this affected participant effort and engagement in the study, defined by attention, dropout rate, time spent on the study, and participant performance. To test whether the effect holds for both financially compensated and volunteer participants, we conducted the study on two sites that are commonly used

by researchers for recruiting online participants: Amazon's Mechanical Turk, an online labor market where participants receive monetary compensation for completing studies, and on LabintheWild, a volunteer online experimentation site.

We found that more formal language improved one aspect of participant engagement: participants were around 10% more likely to pass an attention check when the study used formal language compared to casual language, regardless of platform (i.e., Mechanical Turk or LabintheWild). Those who passed the attention check also exerted themselves more by spending significantly longer on the study and receiving higher scores on average.

While a higher level of formality increased attention, we additionally found that Mechanical Turk participants—with approval ratings of 99% or higher and presumably being accustomed to attention checks [16]—were more likely to pass our attention check than volunteers recruited through LabintheWild. However, Mechanical Turk participants' compliance in reading the instructions did not mean that they exerted themselves more. In actuality, we found that Mechanical Turk participants tended to score lower on the study (perform worse) than LabintheWild participants, suggesting that volunteers exert themselves more than paid crowd-workers in tasks that require significant cognitive effort.

Our results are the first to show that language formality in online studies can impact participant engagement by improving attention, but that it does not impact other variables of engagement, such as time spent on a study or participant performance. When designing studies, researchers should use more formal language if a close reading of instructions is important, as is often the case for complex tasks.

2 RELATED WORK

Related to the current work is (1) measurements of formality in language, (2) formality's effect on users, (3) participant effort in online studies and citizen science, and (4) the effect of study language on participant behaviour.

Formality measurements

Formality has been dubbed one of the "most important dimension of variation between styles" [17] and subsumes a number of linguistic dimensions, including politeness, amount of shared context and social distance [21]. Heylighen and Dewaele [17] measured formality using a metric called the *F*-score, which used frequency counts of context-dependent words, such as pronouns or verbs, and context-independent words, such as nouns or prepositions, to classify document formality. They rated the formality of a number of genres, finding that the *F*-score predictably scored essays and speeches as more formal than phone conversations [17].

Many other methods of measuring formality also look at document-level formality using word counts. For example, Chengyu Fang and Cao [8] used adjective density in a document to measure formality, while Brooke and Hirst [4] defined a lexicon of words whose frequencies either increase or decrease formality, similar to Heylighen and Dewaele [17]'s context-dependent and independent words.

While many formality measures are purely automatic, recent work has begun defining formality in terms of human ratings. Lahiri [27] collected an English corpus of sentences drawn from blogs, news, and forums, with annotations by human raters for formality on a 1-7 scale. They found that human raters could reliably rate sentence formality, with moderate to strong agreement across raters [27]. Pavlick and Tetreault [37] extended this work by building a sentence-level formality classifier trained on the corpus to predict subjective formality ratings. Their classifier used a number of features, including the *F*-score and formality lexicons, and improved over these previous automatic metrics in predicting subjective ratings of formality [37].

Work on automatically translating sentences from informal to formal has also found that formal rewrites of informal sentences often expand contractions (*"don't"* to *"do not"*), change punctuation (*!!!!* to *!*), and paraphrase to more restrained language (*"awesome"* to *"very nice"*) [37, 40]. The current paper draws on these measurements of formality, analyzing whether differences in formality lead to noticeable differences in participant behaviour for an online study.

Effects of formality

Formality measures have allowed researchers to identify how formality use differs across social situations and its effects on individuals. In one study on the Enron email corpus, researchers showed that personal communication was less formal than business related emails, and that when talking to a superior or large group, people used more formal language [38]. This suggests that formal language can be used to enforce power hierarchies or social distance.

Another example of the differential use of formal language is from a study on Wikipedia editors. Researchers found that editors that would eventually be elected as administrators tended to use more polite language (a trait often associated with formality) than editors who would not be elected [9], suggesting that polite language was used to gain social capital among editors. The researchers also found that once Wikipedia editors were elevated to administrators, a position with higher authority, their average politeness went down [9]. Burke and Kraut [5] found that polite language led to greater response rates in online math forums, while rudeness (impoliteness) elicited more responses in political forums. Additionally, Kickstarter campaigns that use more

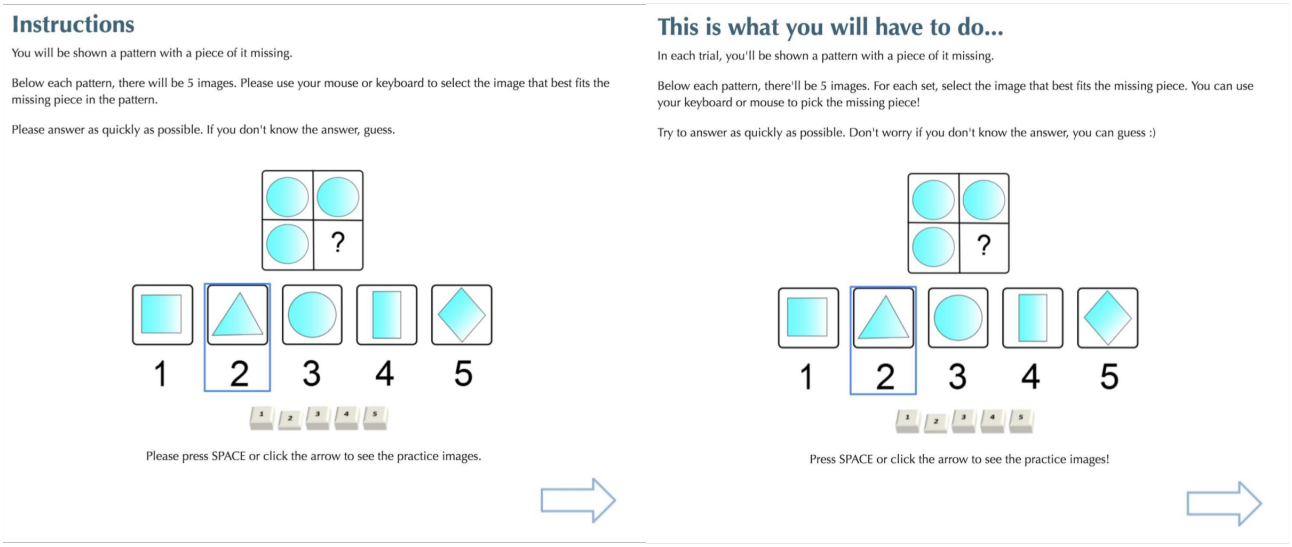

Figure 1: The instruction page in the formal condition (left) and informal condition (right). Formal and informal language was used throughout the study, not just in these instructions.

restrained, thankful language (also associated with formality) are more likely to be funded [10, 32].

While prior work explored the effects of formal and polite language in contexts such as online communities [5] or Kickstarter campaigns [10], our study adds to this literature by showing the effects of formal and informal language in a new context: online experiments. In this paper we explore whether the effects of language formality also occur in online studies: does formal language improve participant effort in online studies? Or is it the opposite?

Effort in online experimentation and citizen science

Online experimentation has grown in popularity from its ability to collect large data samples quickly and cheaply compared to in-lab studies [30]. Many researchers collect study participants on Amazon Mechanical Turk, a major general purpose crowdsourcing site, by offering small financial compensation. As an alternative to financially compensating participants, volunteer-based studies use participants' intrinsic motivations—such as wanting to support science [24, 42]—and have become popular with sites such as TestMyBrain [49], GamesWithWords [14] or LabintheWild [42].

In both volunteer and paid online studies, researchers must take precautions to reduce drop-out rates and guarantee high quality data due to the uncontrolled setting of the internet [24, 30, 43]. Many researchers have developed procedures that can improve the quality of participant responses in online studies and crowd-sourcing, such as attention checks [35], screening [30], detecting unmotivated participants [20], and voting on best submissions [33].

Similar work has investigated effort in citizen science platforms. Raddick et al. [39] identified 12 motivations, including a desire to support science or join a community, for members of GalaxyZoo, a citizen science platform where members classify astronomy images. Lee et al. [28] measured how successful recruiting emails highlighting these motivations were at attracting new members to GalaxyZoo. They found that appealing to supporting science recruited members who contributed more over time than other recruitment emails, suggesting that participants' motivations influence how much effort they invest in a citizen science project.

Some work has shown that volunteer workers provide higher data quality than paid workers in crowd-sourcing tasks. For example, Gil et al. [12] found that quality control measures were required for paid crowd-workers to attain data quality similar to volunteer crowd-workers in a text annotation task. Similarly, Borromeo et al. [2] found that volunteer crowd-workers provided accurate data in both simple and complex data extraction tasks, while paid crowd-worker's accuracy suffered in more complex tasks. Volunteers have also been shown to provide more reliable data than Mechanical Turk participants in a subjective rating task for online experiments, yet this difference disappeared for Mechanical Turk participants with approval ratings above 98% for over 1000 tasks [51].

In contrast, other studies have shown that paid crowd-workers provide data quality similar to volunteers given the right incentives, and at a faster rate [29]. We explore these differences further by quantifying four measures of study engagement–attention, drop out, time spent on the study,

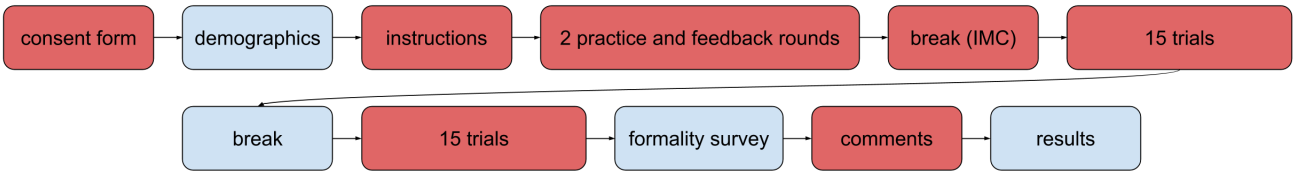

Figure 2: The flow of the experiment. All red steps are instrumented with formal or informal text.

and participant performance–and compare the effects that both experimentation platform and formality have on these measures in an online study.

Effect of study language

The types of questions and instructions in a study can influence the amount of effort required from a participant, as well as the quality of data they provide. In a study on job security, researchers found that almost three times as many participants (21%) answered that a steady job was most important when it was given as an option in a closed question form, compared to when asked in a free response form (8%) [44]. Sniderman and Theriault [46] showed that almost double the number of survey respondents (85%) were in favor of allowing hate group rallies when the question began with a free speech reference, compared to it being introduced with a warning about violence (45%).

Longer surveys or irrelevant questions can also impact participant engagement by causing more participants to drop out [18, 31]. Survey language can mitigate certain types of response errors, such as social desirability bias [34], by wording questions in a neutral or nonthreatening way [25, 34].

In the context of crowd-sourcing tasks, researchers showed that providing basic training and examples in instructions can significantly improve participant responses [33]. While useful in recognizing language's impact on participant response, previous studies have focused on different types of instructions (e.g., providing examples or not [33]) or question framing (e.g., neutral or biased survey questions [44]). We explore how language *style*, specifically formality, impacts user effort in online studies.

3 ONLINE EXPERIMENT

To explore how language formality affects participant engagement in online studies, we designed a between-subjects study testing participants' problem solving skills with two alternate versions of text (formal and informal language). The 10-minute study was launched online on Mechanical Turk (MTurk), a paid crowdsourcing platform, and LabintheWild (LITW), a volunteer-based online experimentation site. In both cases, it was advertised with the slogan "What is your problem solving score?" and participants saw their score on a personalized results page after completing the experiment. We chose this study because it is a cognitively demanding task with verifiable answers and has a similar amount of instructional text as other online studies.

Materials

To test participants' problem solving skills, we developed stimuli similar to Raven's Progressive Matrices [41], a popular test used to assess a component of a person's Intelligence Quotient (IQ). Each trial included a picture of a specific geographic pattern with one piece missing, as well as a choice of five answers, one of which was the correct answer (see Figure 1 for an example). We developed two easy practice trials and 30 trials of increasing difficulty.

To vary the formality of the study's language, we developed two versions of the study instructions: formal and informal. The wording for each was based on common rewrites from informal to formal language found in [37] and [40]. The instrumented language included the study header on the informed consent page, the study's initial instructions, feedback on practice questions, break instructions, prompts for each round and general feedback and comment instructions. Figure 1 provides examples of differences in one section of the study, the initial instructions. Formal text is often longer than informal text [17, 37]; however, we sought to maintain comparability between our conditions by keeping roughly equal text length between our formal and informal conditions. Overall the formal condition had actually 21 words less than our informal condition (271 versus 292). On the Instructional Manipulation Check (IMC) page (Figure 3) the number of words is exactly the same between conditions.

Procedure

Participants began by agreeing to an informed consent and filling out a demographic questionnaire that included questions on their age, gender, education, and what countries they had lived in, as these variables can impact participant motivation for taking studies and formal language use [17, 24]. Participants then read an instruction page and completed two practice rounds with feedback on whether they were correct. Participants saw a break screen with an attention check at the end, as shown on Figure 3, after the two practice

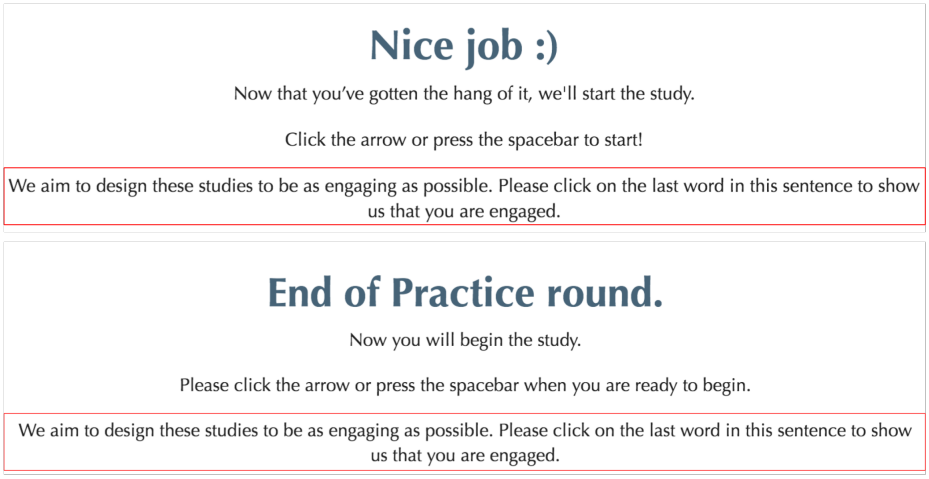

Figure 3: The attention check (highlighted for this paper) in our online experiment in the informal condition (top) and formal condition (bottom). To pass the check, participants needed to press the final word in the sentence: *'engaged.'*

rounds. Participants then completed 15 trials without feedback, followed by a break and 15 more trials. After these 30 trials, participants completed the formality survey, described in Section 3. Participants then submitted any comments or feedback on the study and saw their results. Figure 2 details the study procedure and the text that was instrumented.

Measures

Research in online experimentation and crowdsourcing has identified a number of measurable behaviors related to participant effort that are important for data reliability and statistical power. Among these are participants fully reading instructions [13, 35], participants dropping out [43], the time it takes participants to complete a task [7] and participant error rate [7]. Below we explain how we measured each of these behaviours in our study to quantify participant effort.

(1) **Attention:** We measured participant attention as a binary variable based on if participants passed our instructional manipulation check (IMC) (i.e., attention check), in which they had to click on the final word on the break screen after the practice rounds (see Figure 3). We chose this location because prior work indicates that Mechanical Turk participants have become accustomed to attention checks on the instruction page or on specific questions [16]. Previous studies have also placed attention checks in various study locations and differed in the action they required by participants (e.g., submitting an answer or clicking a word) [16, 24, 35]. For our attention check, participants passed (1) if they clicked on the final word, otherwise they failed (0). The attention check language was the same across both formality conditions.
(2) **Dropout:** We measured dropout as a binary variable. If a participant dropped out of the study before submitting their comments, they were marked as a drop out (1), otherwise they completed the study (0).
(3) **Time:** We measured the amount of time it took participants to complete the study, from the informed consent page to submitting comments before seeing their results.
(4) **Participant performance:** We counted a participant's score on the study itself (how many of the questions they got correct) as a measure of participant performance. This was a score out of 30. While problem solving ability naturally varies between individuals [41], measuring score could capture whether participants gave up on more difficult questions, providing a useful proxy for effort.

To validate our intervention, we also asked participants to rate the formality of a randomly drawn sentence from the study instructions (*"Please read the following sentence and determine its formality."*) on a 1-7 Likert scale from 1 (very informal) to 7 (very formal). We presented examples of formal and informal sentences for participants to anchor their responses, following the procedure of [37] and [27]. The ratings were averaged across each formality condition to gather a general subjective rating of formality from participants.

Participants

Mechanical Turk participants were paid $2 to complete the study. We followed common practices for Mechanical Turk studies and required all participants to have an approval rating of 99% or higher in order to take the study. We accepted all participants from LabintheWild.

Table 1: Participant numbers and demographics from each platform (LabintheWild/Mechanical Turk) and in each condition (formal/informal) after data cleaning. % English refers to the percent of participants who came from English-speaking countries.

	LabintheWild		Mechanical Turk	
	Formal	Informal	Formal	Informal
N	130	142	41	56
Age (sd)	30.93 (12.89)	30.09 (12.77)	34.46 (13.73)	31.93 (9.93)
% Female	40.00%	45.07%	58.54%	48.21%
% English	48.46%	52.11%	87.81%	87.50%
Education in years (sd)	15.77 (3.37)	15.98 (4.14)	15.85 (3.14)	14.93 (3.10)

A total of 492 participants completed our study (383 participants from LabintheWild and 109 from Mechanical Turk). We removed 59 participants who indicated in the questionnaire that they had taken the study before, specified that they had technical difficulties, simply clicked through the study, or spent longer than 60 minutes. We took a completion time of over 60 minutes as an indication that participants attended to other tasks or left their computer for long periods of time (87% of participants took less than 20 minutes). Because the study did not require participants to provide all demographic information, 64 participants did not provide their age, gender, or education level, whom we also excluded from analysis. In total we had 369 participants (167 female), 272 from LabintheWild and 97 from Mechanical Turk. Based on a power analysis for identifying small effects (power level = .8, probability level = .05), we required a minimum sample size of 307 participants. Overall participants came from 65 countries, with 60% coming from English speaking countries. Table 1 details the participants in the formal and informal conditions and from LabintheWild and Mechanical Turk.

Analysis

We first validated our intervention by comparing the average formality rating participants gave to instructions in the two formality conditions. We used an independent samples t-test to determine whether the formality ratings were significantly different between the two formality conditions. We augmented our t-test with Cohen's d effect size [47]. This subjective rating of formality clarified how participants perceived the formality versions differently.

We explored how language formality and platform influenced engagement by constructing a regression model for each of our engagement measures. Since other variables, such as age or gender, can play a role in participant engagement [24], we included in our regression models a participant's age, education, gender, and a variable indicating whether the participant came from an English speaking country (*English*). We added *English* because participants whose native language is not English might react differently to formality [26, 50], or struggle with study procedures that could impact our measures of effort (e.g., they might take longer to read instructions or complete the study) [16]. Past work has linked attention checks with other measures of participant engagement [36] so we included attention in our model for time spent and participant performance to analyze how our measures of engagement related to one another. We did not include these additional variables for the regression model on participant dropout because less than 1% of participants who had submitted their demographic information or passed the attention check dropped out of the study, making any model that used these variables skew towards no drop outs.

We initially constructed a secondary model that included the interaction effect of formality and platform on measures of engagement to see if there was a significant differential effect of formality on Mechanical Turk versus LabintheWild participants. However, this was not significant and did not improve any models' fit based on the Akaike's Information Criteria (AIC) [3], so we left it out of our final analyses.

For our linear regression results, we report on the adjusted R^2, which measures how much of the observed variance is explained by the model (e.g., $R^2 = .1$ means 10% of the variance in the data is explained by the model). We also present the coefficients of each variable, which is the expected increase or decrease in the dependent variable that a one unit increase in the independent variable (or change to other condition in the case of categorical variables) causes. When reporting on logistic regression results, we report the McFadden's Pseudo R^2, a substitute for the the R^2 statistic for logistic regression [19], and the beta coefficients for each independent variable in the model. The beta coefficient represents the odds ratio of the independent variable: the relative increase in likelihood of the dependent variable given a one point increase in the independent variable. When the independent variable is categorical (e.g., formality), the odds ratio is the increase in likelihood compared to the other condition.

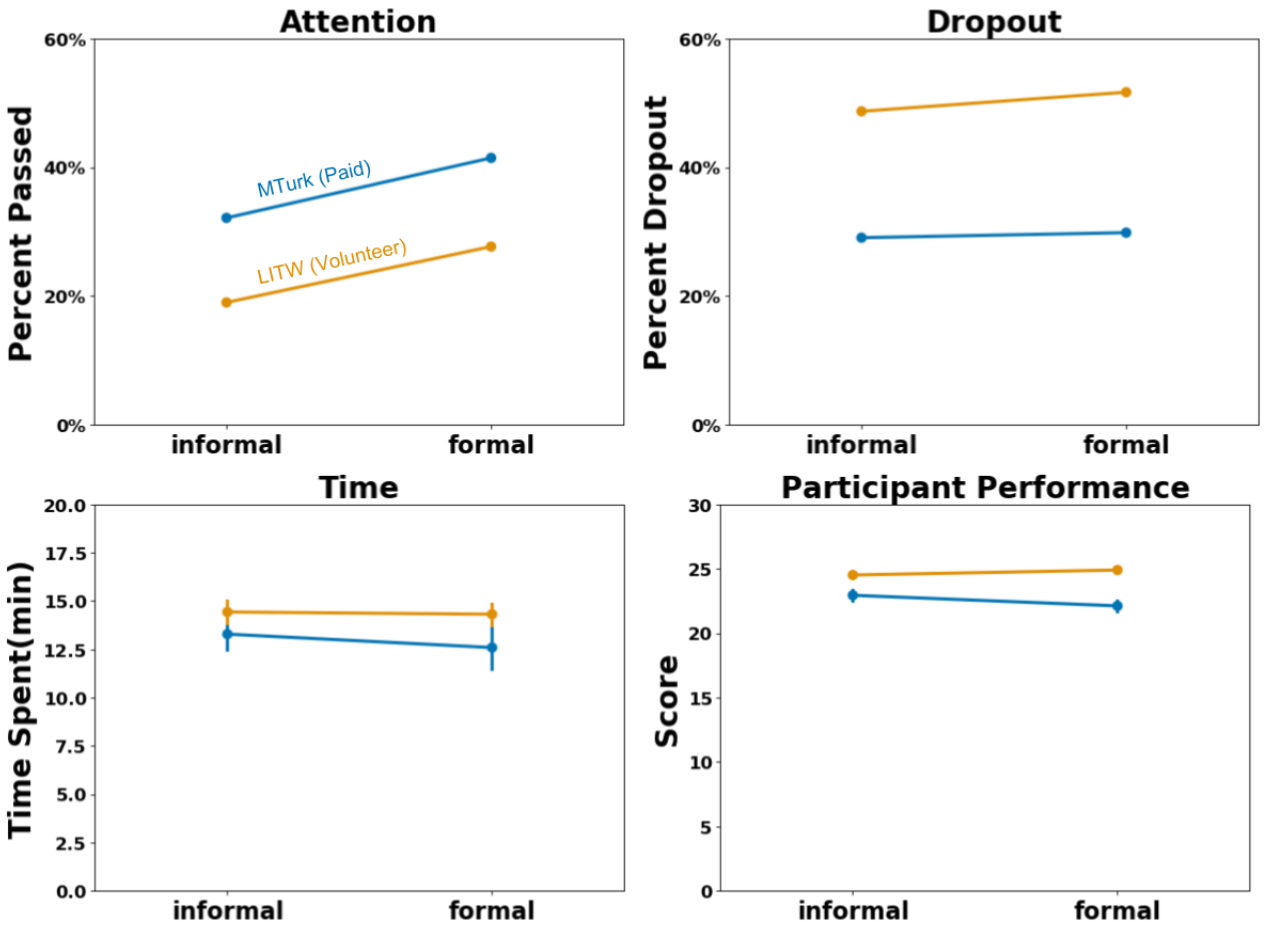

Figure 4: Study engagement measures across conditions (informal/formal) and participant samples (MTurk/LabintheWild). For attention and dropout, percentages are without controlling for participant demographics. Time and study performance report marginal means with error bars using standard error. For performance, score is out of 30 and higher is better.

All analyses were made in Python using the *statsmodels* and *SciPy* libraries [23, 45]. Our data set and analysis scripts are publicly available[1].

4 RESULTS

Validating formality ratings. Participants rated sentences from the formal condition (m=5.09, sd=1.39) significantly higher than those in the informal condition (m=3.96, sd=1.74 on a 1-7 scale, $t_{359} = 6.65$, $p < .0001$, $d = .70$). This shows that the two conditions were perceived as two different levels of formality. Because highly informal language is linked to lower informativeness [27], lack of proper grammar [37], and even swear words [4], we aimed to not write instructions on the extremes of the formal/informal dimension, feeling that this was unrealistic. However, there is still a range of formality that is acceptable in an online study that we sought to compare in our two conditions. For example, social media posts for academic surveys can range from using emojis and causal language (e.g., *"Please take my study :D"*) to more formerly worded requests for participation with included citations (e.g., *"This study builds on past work from August et al."*). We were interested in how these subtler differences in formality can lead to differences in participant behavior.

We measured participant engagement across a number of dimensions that past work has identified as important in online experimentation [24, 30, 35]: attention, dropout, time spent on the study, and participant score. Below we outline our regression results for each of these measures, evaluating how instruction formality (formal vs. informal) and platform (LabintheWild vs. MTurk) impacted engagement while controlling for participant demographics.

[1]https://github.com/talaugust/CHI2019_Formality

Table 2: Coefficients for each independent variable in the regression models with participant engagement measures (i.e., attention, dropout, time spent on the study, and performance) as the dependent variables. Adjusted R^2 (pseudo R^2 in the case of logistic regression) is listed along with the dependent variable. * = $p < .05$, ** = $p < .01$. The coefficients for attention and dropout are presented as odds ratios.

	Attention (Psd. R^2 = .06)**	**Dropout (Psd. R^2 = .02)****	**Time (min) (R^2 = .08)****	**Performance (R^2 = .12)****
Formality (formal)	**1.68***	1.11	-.63	-0.02
Platform (MTurk)	**1.83***	**0.42****	-1.35	**-1.86****
English (True)	1.58	N/A	**-1.99***	-0.53
Age	1.00	N/A	**.138****	-0.01
Gender (male)	.75	N/A	-.93	**.99****
Education	**1.08***	N/A	-.03	0.08
Attention (Passed)	N/A	N/A	**1.68***	**0.78***

Attention. Overall, 26.56% of participants passed the attention check, suggesting that most participants did not read the instructions on the break page fully. Participants in the formal condition passed the attention check 31.00% of the time, while participants in the informal condition passed the check 22.72% of the time. Mechanical Turk participants passed the check 36.08% of the time, while LabintheWild participants passed it 23.16% of the time (see Figure 4). When controlling for participant demographics (age, gender, education, and *English*), participants in the formal group were 1.68 times more likely to pass the attention check than those in the informal group. Mechanical Turk participants were also 1.83 times more likely to pass the attention check compared to volunteers (see the first column in Table 2).

Dropout. A total of 928 participants landed on the consent page, with 494 completing the study (434 dropping out) for a drop out rate of 46.77%. The linear regression model reported in the second column of Table 2 shows that Mechanical Turk participants were less than half (.42 times) as likely to drop out compared to LabintheWild participants (see Figure 4).

Time. Participants on average completed the study in 14.01 minutes (sd=6.61). Neither formality nor platform was significantly associated with time spent in the study based on the regression model (3rd column in Table 2 and Figure 4). Attention, however, did correlate with time spent on the study: participants who passed the attention check spent approximately 1.68 minutes longer on the study compared to participants that did not pass the attention check.

Participant performance. Our final measure of participant engagement was the participant's performance on the study itself. Participants scored on average 24.15 (sd=3.12) out of 30. When controlling for our demographic variables, platform correlated significantly with performance. Turkers scored approximately 1.86 points out of 30 lower than volunteers (4th column in Table 2 and Figure 4). Attention also significantly correlated with performance, with participants who had passed the attention check scoring approximately 0.78 points higher than those who did not pass.

5 DISCUSSION

This paper set out to answer how language formality influences participant engagement and effort in online studies. Are participants engaged and exert themselves more when reading formal instructions compared to instructions in casual language? While past work has identified the effects formality or politeness has on user behavior in contexts like online communities [5] or fundraising [10], our results identify the effect formality has on user behaviour in a new context: online experimentation.

We found that more formal instructions improves participant engagement in online studies: participants who read formal instructions were close to 10% more likely to pass our attention check compared to participants who read casually written instructions. More participants in our study passed the attention check compared to previous work (26.56% versus 7.6% in [24], where the attention check was presented on the informed consent page) using the same type of check (i.e., having participants click on a word in the instructions).

Participants who passed our attention check also spent over a minute and a half longer on the study and scored close to a full point higher out of 30, suggesting that our attention check, similar to past work [36], captured part of a participant's engagement and effort in the study. However, language formality did not impact these other measures of participant engagement. Using more formal language causes participants to read instructions more carefully, increasing participant effort in one dimension, but this effect did not carry over to other measures of engagement.

Formal language tends to be clearer and more precise than casual language [17]. It is possible for this reason that participants read the instructions and identified the attention check more easily in the formal condition compared to the informal condition. Another explanation for our finding is that formal language is associated with higher-stakes environments (e.g., talking to many people or to a superior [9]). This could lead participants to pay closer attention to instructions when the study used more formal language.

Mechanical Turk can be seen as such a higher-stakes environment, given that participants strive for high approval ratings and receiving compensation. In support of this, we found that Mechanical Turk participants were more likely to pass the attention check and complete the study (i.e., not drop out) than LabintheWild participants. Past work in online experimentation and crowdsourcing suggests that Mechanical Turk participants have learned to identify and pass attention checks due to their ubiquity in paid online experiments and crowdsourcing tasks [16]. Studies using Mechanical Turk also will sometimes use attention checks as a screening mechanism, only paying participants who pass the check. This raises the stakes for Mechanical Turk participants to become adept at identifying and passing these checks, while LabintheWild volunteers have no such compulsion. Mechanical Turk participants had probably come across many more attention checks while taking studies (especially given that they were required to have a 99% approval rating to take our study, suggesting that they were experienced workers on Mechanical Turk) so were better at picking out our attention check than volunteer participants on LabintheWild.

While Mechanical Turk participants passed our attention check more often and dropped out less, they preformed *worse* (scored lower) on our study overall. These results could be seen as contradictory because in general our participants who passed the attention check scored *higher* on the study. Research has shown that paid participants give up on cognitively demanding tasks quicker than volunteers [2, 29]. It is possible that since Mechanical Turk participants might have been motivated to finish the study as quickly as possible (there is little benefit to getting a higher score), they gave up on more demanding problems in the study faster, investing less effort in the study and resulting in lower scores overall.

Design Implications

Participants engaging and investing effort in an online study—reading instructions, completing the study, and trying their best—is important for gathering reliable results. We found in our study that participants in the formal condition passed our attention check more often than participants in the informal condition, showing that formal language has an impact on participant engagement in our online study. While reading instructions sentence by sentence is not necessarily important for all experiments, there are many tasks that contain instructional sentences that participants must read in order to provide reliable data (e.g., a sentence explaining that stimuli are timed and therefore require focusing on the screen). For these tasks, we recommend researchers to use more formal language to encourage participants' sustained focus and mitigate any errors resulting from them skimming instructions. It is also worth noting that formal language did not harm any measure of engagement, so researchers do not have any risks associated with using formal language.

Researchers interested in redesigning a study with formal language are faced with the issue of *how* to redesign their writing – a notoriously difficult process. The first step is to quantify the formality of the language a study currently uses. Researchers can use simple automatic measures of text formality, such as the F-score [17], more advanced text classifiers like the one presented in [37], or follow our procedure and gather subjective ratings of formality from participants (similar to the ratings that Pavlick and Tetreault [37] trained their classifier on). Once researchers have quantified how formal their study language is, they can rewrite instructions to increase the formality by expanding contractions (*"won't"* to *"will not"*), adding more polite language (using *"please"*) or paraphrase to more restrained language (*"this is the best ever!!"* to *"great work"*) [9, 37, 40].

Researchers can also use our results to weigh the trade-offs of using platforms that offer different incentives. We found that Mechanical Turk (paid) participants exerted less effort in the study compared to LabintheWild (volunteer) participants, shown by lower performance on the study. While Mechanical Turk participants read instructional text more closely, researchers might be better served by recruiting volunteers to guarantee high quality data for cognitively demanding tasks where performance and exertion is independent of having read each sentence in the instructions.

6 LIMITATIONS & FUTURE WORK

Labinthewild and Mechanical Turk are two different platforms, with different populations and norms of behaviour. This makes it difficult to draw any final conclusions about the influence of incentive on participant engagement. It could be that the differences in engagement we observed are based on other factors that differ between these two platforms. It would be interesting to deploy studies on a hybrid platform, one providing monetary and non-monetary incentives, in order to explicitly analyze the effect of incentive on participant engagement and effort.

Our measures of participant effort present a few limitations as well. We tried to remove participants who we suspected left their computers or attended to other tasks, yet due to these confounds time spent on a study is an imperfect

measure of effort [7]. We placed our attention check on the break page to avoid more common locations that participants might have expected [16], yet this also meant that the location we selected had less information for completing the study successfully, reducing its practical significance.

Work has also shown that attention checks influence participant behavior later in a study [15], making it difficult to disentangle how our attention check related to other measures of engagement like time spent on the study. Were participants who passed the attention check more engaged before the check, or did noticing the check push them to pay more attention later on? An important future direction of this work is defining more rigorous measures of participant effort in the context of online experimentation, possibly drawing on work from measuring effort in crowdsourcing tasks [7] or using behavioral measures like eye-tracking to quantify participant engagement.

Sentences for the formal and informal study conditions were written based on the rewrites found in [37] and [40]. It would be interesting to explore how specific syntactic differences affect formality ratings. For example, do emojis shift perceived formality more substantially than extra exclamation points, and does this shift lead to a larger impact on participant engagement? An exciting future direction of this research is identifying these specific linguistic features and their effect on participant engagement online.

7 CONCLUSION

This is the first study that looked at the effects of language formality on participant engagement and effort in an online study. Participants reading formal instructions are almost twice as likely to pass an attention check compared to those reading more casually written instructions. We therefore recommend that researchers write more formal instructions for complex tasks that require participants' close attention. Participants also behave differently depending on the platform they are recruited from. Mechanical Turk (paid) participants dropped out less but scored lower overall compared to LabintheWild (volunteer) participants, making volunteer participants ideal for collecting high quality data in cognitively demanding tasks.

8 ACKNOWLEDGMENTS

This work was partially funded by NSF award 1651487. We greatly thank the LabintheWild participants who make this research possible. We also thank the reviewers for their time and input.

9 DATASET

We make available our datasets and Python code for analysis at https://github.com/talaugust/CHI2019_Formality.

REFERENCES

[1] Michael Barber, David Gordon, Ryan Hill, and Joseph Price. 2017. Status Quo Bias in Ballot Wording. *Journal of Experimental Political Science* 4, 2 (2017), 151–160.

[2] Ria Mae Borromeo, Thomas Laurent, and Motomichi Toyama. 2016. The influence of crowd type and task complexity on crowdsourced work quality. In *Proceedings of the 20th International Database Engineering & Applications Symposium*. ACM, 70–76.

[3] Hamparsum Bozdogan. 1987. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika* 52, 3 (1987), 345–370.

[4] Julian Brooke and Graeme Hirst. 2014. Supervised Ranking of Co-occurrence Profiles for Acquisition of Continuous Lexical Attributes. *Coling* (2014), 2172–2183. https://pdfs.semanticscholar.org/3694/d5b26cd9bbeb72f6a85552396adffff5e4fa.pdfhttp://www.cs.utoronto.ca/pub/gh/Brooke+Hirst-COLING-2014.pdf

[5] Moira Burke and Robert Kraut. 2008. Mind your Ps and Qs: the impact of politeness and rudeness in online communities. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. ACM, 281–284.

[6] Justine Cassell and Timothy Bickmore. 2003. Negotiated Collusion: Modeling Social Language and its Relationship Effects in Intelligent Agents. *User Modeling and User-Adapted Interaction* 13, 1/2 (2003), 89–132. https://doi.org/10.1023/A:1024026532471

[7] Justin Cheng, Jaime Teevan, and Michael S Bernstein. 2015. Measuring crowdsourcing effort with error-time curves. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1365–1374.

[8] Alex Chengyu Fang and Jing Cao. 2009. Adjective Density as a Text Formality Characteristic for Automatic Text Classification: A Study Based on the British National Corpus. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 1*. http://www.aclweb.org/anthology/Y09-1015http://aclweb.org/anthology/Y09-1015

[9] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. *arXiv preprint arXiv:1306.6078* (2013).

[10] Nihit Desai, Raghav Gupta, and Karen Truong. 2015. Plead or pitch? The role of language in kickstarter project success.

[11] Samantha Finkelstein, Evelyn Yarzebinski, Callie Vaughn, Amy Ogan, and Justine Cassell. 2013. The effects of culturally congruent educational technologies on student achievement. In *International Conference on Artificial Intelligence in Education*. Springer, 493–502.

[12] Yolanda Gil, Felix Michel, Varun Ratnakar, Matheus Hauder, Christopher Duffy, Hilary Dugan, and Paul Hanson. 2015. A task-centered framework for computationally-grounded science collaborations. In *e-Science (e-Science), 2015 IEEE 11th International Conference on*. IEEE, 352–361.

[13] Joseph K. Goodman, Cynthia E. Cryder, and Amar Cheema. 2012. Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples. *Journal of Behavioral Decision Making* 26, 3 (jul 2012), 213–224. https://doi.org/10.1002/bdm.1753

[14] Joshua Hartshorne. 2018. Games With Words. https://www.gameswithwords.org/

[15] David J Hauser and Norbert Schwarz. 2015. It's a trap! Instructional manipulation checks prompt systematic thinking on "tricky" tasks. *Sage Open* 5, 2 (2015), 2158244015584617.

[16] David J Hauser and Norbert Schwarz. 2016. Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior research methods* 48, 1 (2016), 400–407.

[17] Francis Heylighen and Jean-Marc Dewaele. 1999. Formality of language: definition, measurement and behavioral determinants. *Interner Bericht, Center "Leo Apostel", Vrije Universiteit Brüssel* (1999).

[18] Michael Hoerger. 2010. Participant dropout as a function of survey length in Internet-mediated university studies: Implications for study design and voluntary participation in psychological research. *Cyberpsychology, Behavior, and Social Networking* 13, 6 (2010), 697–700.

[19] Bo Hu, Jun Shao, and Mari Palta. 2006. Pseudo-R 2 in logistic regression model. *Statistica Sinica* (2006), 847–860.

[20] Jason L Huang, Paul G Curran, Jessica Keeney, Elizabeth M Poposki, and Richard P DeShon. 2012. Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology* 27, 1 (2012), 99–114.

[21] Judith T. Irvine. 1979. Formality and informality in communicative events. *American Anthropologist* 18, 2 (dec 1979), 773–790. https://doi.org/10.1525/aa.1979.81.4.02a00020 arXiv:arXiv:1011.1669v3

[22] Aaron Jaech, Victoria Zayats, Hao Fang, Mari Ostendorf, and Hannaneh Hajishirzi. 2015. Talking to the crowd: What do people react to in online discussions? *arXiv preprint arXiv:1507.02205* (2015).

[23] Eric Jones, Travis Oliphant, Pearu Peterson, et al. 2001–. SciPy: Open source scientific tools for Python. http://www.scipy.org/ Online.

[24] Eunice Jun, Gary Hsieh, and Katharina Reinecke. 2017. Types of Motivation Affect Study Selection , Attention , and Dropouts in Online Experiments. *Proceedings of the ACM on Human-Computer Interaction Archive* 1, 1 (2017), 1–15. https://doi.org/10.1145/3134691

[25] Ivar Krumpal. 2013. Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity* 47, 4 (2013), 2025–2047.

[26] Riadh Ladhari, Frank Pons, Grégory Bressolles, and Michel Zins. 2011. Culture and personal values: How they influence perceived service quality. *Journal of Business Research* 64, 9 (sep 2011), 951–957. https://doi.org/10.1016/j.jbusres.2010.11.017

[27] Shibamouli Lahiri. 2015. SQUINKY! A Corpus of Sentence-level Formality, Informativeness, and Implicature. *arXiv* (2015). arXiv:1506.02306 https://arxiv.org/pdf/1506.02306.pdfhttp://arxiv.org/abs/1506.02306

[28] Tae Kyoung Lee, Kevin Crowston, Mahboobeh Harandi, Carsten Osterlund, and Grant Miller. 2018. Appealing to different motivations in a message to recruit citizen scientists: results of a field experiment. *JCOM: Journal of Science Communication* 17, 1 (2018), A1–A1.

[29] Andrew Mao, Ece Kamar, Yiling Chen, Eric Horvitz, Megan E Schwamb, Chris J Lintott, and Arfon M Smith. 2013. Volunteering versus work for pay: Incentives and tradeoffs in crowdsourcing. In *First AAAI conference on human computation and crowdsourcing*.

[30] Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods* 44, 1 (mar 2012), 1–23. https://doi.org/10.3758/s13428-011-0124-6 arXiv:http://ssrn.com/abstract=1691163

[31] Jim McCambridge, Eleftheria Kalaitzaki, Ian R White, Zarnie Khadjesari, Elizabeth Murray, Stuart Linke, Simon G Thompson, Christine Godfrey, and Paul Wallace. 2011. Impact of length or relevance of questionnaires on attrition in online trials: randomized controlled trial. *Journal of medical Internet research* 13, 4 (2011).

[32] Tanushree Mitra and Eric Gilbert. 2014. The language that gets people to give: Phrases that predict success on kickstarter. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 49–61.

[33] Tanushree Mitra, Clayton J Hutto, and Eric Gilbert. 2015. Comparing person-and process-centric strategies for obtaining quality data on amazon mechanical turk. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1345–1354.

[34] Anton J Nederhof. 1985. Methods of coping with social desirability bias: A review. *European journal of social psychology* 15, 3 (1985), 263–280.

[35] 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.

[36] Leonard J Paas, Sara Dolnicar, and Logi Karlsson. 2018. Instructional Manipulation Checks: A longitudinal analysis with implications for MTurk. *International Journal of Research in Marketing* (2018).

[37] Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association of Computational Linguistics* 4, 1 (2016), 61–74.

[38] Kelly Peterson, Matt Hohensee, and Fei Xia. 2011. Email formality in the workplace: A case study on the enron corpus. *Proceedings of the Workshop on Languages in Social Media* June (2011), 86–95. https://dl.acm.org/citation.cfm?id=2021120http://dl.acm.org/citation.cfm?id=2021120

[39] M Jordan Raddick, Georgia Bracey, Pamela L Gay, Chris J Lintott, Carie Cardamone, Phil Murray, Kevin Schawinski, Alexander S Szalay, and Jan Vandenberg. 2013. Galaxy Zoo: Motivations of citizen scientists. *arXiv preprint arXiv:1303.6886* (2013).

[40] Sudha Rao and Joel Tetreault. 2018. Dear Sir or Madam, May I introduce the YAFC Corpus: Corpus, Benchmarks and Metrics for Formality Style Transfer. *arXiv preprint arXiv:1803.06535* (2018).

[41] John Raven. 2008. The Raven progressive matrices tests: their theoretical basis and measurement model. In *Uses and abuses of Intelligence. Studies advancing Spearman and Raven's quest for non-arbitrary metrics*. 17–68. https://www.researchgate.net/publication/255605513

[42] 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.

[43] Ulf-Dietrich Reips. 2002. Standards for Internet-based experimenting. *Experimental psychology* 49, 4 (2002), 243.

[44] Howard Schuman and Stanley Presser. 1977. Question wording as an independent variable in survey analysis. *Sociological Methods & Research* 6, 2 (1977), 151–170.

[45] J.S. Seabold and J. Perktold. 2010. Statsmodels: Econometric and Statistical Modeling with Python. In *Proceedings of the 9th Python in Science Conference*.

[46] P.M. Sniderman and S.M. Theriault. 2004. The Structure of Political Argument and the Logic of Issue Framing. (jan 2004), 133–164.

[47] Gail M Sullivan and Richard Feinn. 2012. Using Effect Size-or Why the P Value Is Not Enough. *Journal of graduate medical education* 4, 3 (sep 2012), 279–82. https://doi.org/10.4300/JGME-D-12-00156.1

[48] Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic-and author-controlled natural experiments on Twitter. *arXiv preprint arXiv:1405.1438* (2014).

[49] TestMyBrain. 2017. TestMyBrain. http://dx.plos.org/10.1371/journal.pone.0165100

[50] Judith Yaaqoubi and Katharina Reinecke. 2018. The Use and Usefulness of Cultural Dimensions in Product Development. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, CS18.

[51] Teng Ye, Katharina Reinecke, and Lionel P Robert Jr. 2017. Personalized Feedback Versus Money: The Effect on Reliability of Subjective Data in Online Experimental Platforms. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 343–346.