Personalized Feedback Versus Money: The Effect on Reliability of Subjective Data in Online Experimental Platforms

Teng Ye[1], Katharina Reinecke[2], Lionel P. Robert Jr.[1]

[1] School of Information
University of Michigan
Ann Arbor, MI 48109 USA
{tengye, lprobert}@umich.edu

[2] Computer Science & Engineering, DUB Group
University of Washington
Seattle, WA 98195 USA
reinecke@cs.washington.edu

Abstract

We compared the data reliability on a subjective task from two platforms: Amazon's Mechanical Turk (MTurk) and LabintheWild. MTurk incentivizes participants with financial compensation while LabintheWild provides participants with personalized feedback. LabintheWild was found to produce higher data reliability than MTurk. Our findings suggest that online experiment platforms providing feedback in exchange for study participation can produce more reliable data in subjective preference tasks than those offering financial compensation.

CSCW '17 Companion, February 25 - March 01, 2017, Portland, OR, USA
ACM 978-1-4503-4688-7/17/02.
http://dx.doi.org/10.1145/3022198.3026339

Author Keywords

Crowdsourcing; Online Experimentation; Mechanical Turk; Compensation; Motivation; Incentives; Data Quality.

ACM Classification Keywords

H.5.2. Information Interfaces and Presentation: User Interfaces—Interaction styles, Evaluation/Methodology; H.5.3. Information Interfaces and Presentation: Group and Organization Interfaces---Web-based interaction

Introduction

Crowdsourcing platforms, such as Amazon's Mechanical Turk (MTurk), support rapid and mass experimental data collection by allowing researchers to easily recruit and pay their online participants. Since online experiment is conducted unsupervised and in uncontrolled environment, researchers have been worried that participants can pay inadequate attention to instructions, provide untruthful responses, putting insufficient effort, or are distracted (e.g., [2]). Pre- and post-experimental actions have been explored to improve MTurk data reliability so that it's comparable to that of conventional lab studies (e.g., [7]).

In contrast to MTurk, other online experiment platforms, such as TestMyBrain [1] or LabintheWild [5], rely on the participation of volunteers and are designed to be

intrinsically motivating. Instead of receiving financial compensation, these participants are provided with personalized feedback (such as comparing them with other participants) at the end of an experiment to satisfy their inherent interests. Different incentive mechanisms—extrinsic versus intrinsic— are likely to attract participants driven by different motivations. However, we know little about how such differences influence data reliability in online experiments.

Thus, we conducted a study comparing the data reliability between MTurk and LabintheWild on a subjective preference task. These tasks are challenging to conduct online because they are especially prone to low data reliability given the lack of possibilities to control the accuracy of participants' answers [2].

Our findings contribute to the literature by showing that unpaid experiments conducted on a platform that provides feedback instead of financial compensation can elicit higher data quality in subjective preference tasks than experiments conducted on paid crowdsourcing platforms. In addition, our findings demonstrate that providing an intrinsic incentive to participants on a platform that primarily appeals to participants' extrinsic motivation (i.e., on MTurk) does not lead to higher data quality.

Hypotheses

While possibly motivated by both intrinsic and extrinsic factors, Turkers have been found to be primarily motivated by the financial return: about 75% of US Turkers reported that MTurk was their primary or secondary source of income, compared to 40% US Turkers reported to participate for fun and 30% to kill time [4]. As a result, Turkers who are primarily driven by the monetary reward would have an interest in minimizing time and effort needed to perform an experiment. In contrast, LabintheWild participants complete experiments in order to compare themselves to others [5]. Seeing that providing truthful responses is a precondition to receiving such feedback based on their true preferences, we assume that LabintheWild participants provide more reliable responses in a subjective task:

Hypothesis 1: The data reliability from LabintheWild will be higher than that from Amazon's Mechanical Turk.

Experiments on MTurk usually do not provide personalized performance feedback, or the ability to compare oneself to others. However, if participants are provided with such feedback in addition to financial compensation, they may be more motivated to perform the task with greater attention and effort. This, in turn, should improve data reliability. For example, Turkers produced more accurate results when they were told that they were performing a task for a non-profit (increasing intrinsic motivation) than for a for-profit organization [6]. Therefore:

Hypothesis 2: The data reliability from Amazon's Mechanical Turk participants is higher when Turkers are provided with personalized feedback in addition to their financial compensation.

Method

To test our hypotheses, we replicated an experiment that has been used to test data reliability from the literature [3,5] on both LabintheWild and MTurk. In the task, participants were asked to provide subjective ratings on the appeal of websites.

Implementation

Procedure: After reading and accepting the informed consent, participants were presented a demographics questionnaire, as well as a screen containing instructions on the task. The instructions emphasized that the short stimulus exposure time of 500ms per website would require extra attention. Participants were able to try the task in a practice run with five websites. The main experiment consisted of two phases. In each phase, the participants were shown the same 30 website screenshots in a randomized order followed by a 9-point Likert-type scale ranging from "not visually appealing at all" to "very visually appealing." The full experiment took around 8 minutes to complete.

Measure: The data reliability was measured by a participant's consistency between two ratings of the same website. Higher consistency indicated higher data reliability.

Condition	# Participants	% Female	Age	Incentive
LabintheWild (Feedback only)	29,999	53.99%	18-69	feedback only
MTurk Overall	209	43.75%	18-69	$0.80 (+ feedback)
General Turkers (Money with Feedback)	53	38.00%	19-51	$0.80 + feedback
General Turkers (Money only)	53	49.06%	22-62	$0.80
Master Turkers (Money with Feedback)	52	42.31%	22-69	$0.80 + feedback
Master Turkers (Money Only)	51	45.28%	18-68	$0.80

Table 1: Participant demographics and incentives in the five experiment conditions

The experiment used a between-subjects design. Incentives provided and demographic details of each condition can be seen in Table 1. In all conditions with feedback, the feedback, a comparison of one's aesthetic website preferences to other participants, was provided at the end of the experiment. In all conditions with money, participants received $0.80 (roughly $6/hour for the 8-minute task) for their participation. The MTurk experiment was open for any participant from the US[1] (we call these participants *general Turkers*). We expected general Turkers to be most comparable to the population on LabintheWild that is open for anyone to participate. The demographics of the two populations were later matched for analysis to ensure comparability.

To test the effect of feedback on Turkers with high approval ratings and qualifications, we additionally conducted the experiment with *Master Turkers*, who came from the United States, had an approval rate greater than 98%, and had completed at least 1000 tasks.

[1] We restricted access to Turkers from the US to ensure that participants understood the instructions provided in English.

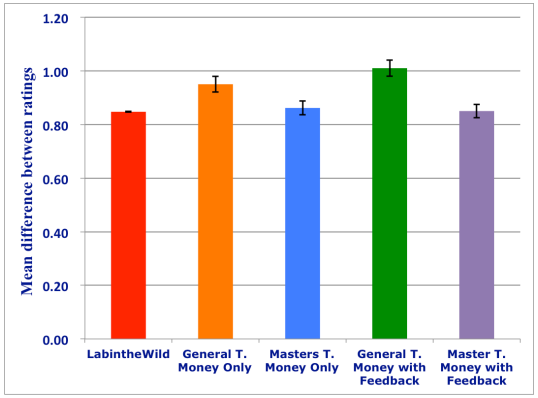

Figure 1: Mean differences in subjective ratings between test phases 1 and 2 on a 9-point scale. Error bars show the standard error.

Results

Figure 1 provides an overview of the results. *Hypothesis 1*: The results of ANOVA showed that differences in ratings between the two phases by LabintheWild participants were significantly smaller ($m = 0.85, sd = 0.95$ on a 9-point Likert-type scale) than those provided by Turkers ($m = 0.95, sd = 1.17, F_{(1, 1526480)} = 16.984, p < .001$). An additional bootstrapping analysis with equal sample sizes (randomly selecting 50

participants from the MTurk and 50 from the LabintheWild dataset and resampling the data 1000 times) confirmed the result (*observed coefficient = 16.98, bootstrap SE = 5.32, 95% CI= [6.56, 27.40], p< .001*). Both analyses supported H1.

We additionally compared the reliability of participants' responses between LabintheWild participants (*m = .85, sd = 0.95*) and Master Turkers (*m = 0.86, sd = 1.00*), but there was no significant difference ($F_{(1,1526420)} = 0.151$, $p = .70$), showing that Master Turkers provided equally reliable subjective ratings as LabintheWild participants.

Hypothesis 2: The results of an ANOVA showed no significant difference between the reliability of ratings provided by general Turkers in the money-with-feedback (*m = 1.01, sd = 1.21*) vs. money-only condition (*m = 0.95, sd = 1.17*) *with* $F_{(1,3178)} = 2.001$, $p = .16$). Hypothesis 2 was therefore not supported.

Conclusion

Our results show that uncompensated experiments conducted on platforms that provide personalized feedback instead of financial compensation can support researchers in collecting more reliable subjective data than on MTurk when not restricting access to the experiment. We hope that our work opens up new opportunities for researchers to conduct their experiments on a variety of platforms using different incentive mechanisms.

Acknowledgements

The authors thank participants for their participation, our friends for their helpful feedback, and Trevor Croxson for assistance in setting up the experiments.

References

1. 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: 847–857.
2. Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing User Studies with Mechanical Turk. CHI '08, 453–456. https://doi.org/10.1145/1357054.1357127
3. Gitte Lindgaard, Cathy Dudek, Devjani Sen, Livia Sumegi, and Patrick Noonan. 2011. An exploration of relations between visual appeal, trustworthiness and perceived usability of homepages. *ACM Transactions on Computer-Human Interaction (TOCHI)* 18, 1: 1.
4. Gabriele Paolacci, Jesse Chandler, and Panagiotis G. Ipeirotis. 2010. Running experiments on amazon mechanical turk. *Judgment and Decision making* 5, 5: 411–419.
5. Katharina Reinecke and Krzysztof Z. Gajos. 2015. LabintheWild: Conducting Large-Scale Online Experiments With Uncompensated Samples. CSCW '15, 1364–1378. https://doi.org/10.1145/2675133.2675246
6. Jakob Rogstadius, Vassilis Kostakos, Aniket Kittur, Boris Smus, Jim Laredo, and Maja Vukovic. 2011. An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets. *ICWSM* 11: 17–21.
7. Aaron D. Shaw, John J. Horton, and Daniel L. Chen. 2011. Designing Incentives for Inexpert Human Raters. CSCW '11, 275–284. https://doi.org/10.1145/1958824.1958865