An Exploratory Study of Data Quality and Participation in a Games-for-Science Game Community

Jason Radford
University of Chicago, Northeastern University
177 Huntington Ave
Boston, MA 021515
j.radford@neu.edu

David Lazer
Northeastern University
177 Huntington Ave
Boston, MA 02115
d.lazer@neu.edu

Abstract

In this paper we explore what predicts data quality and participation by human subjects in an online, games-for-science community. The advent of games for science, citizen science, and online laboratories represent a new world of possibilities for conducting scientific research with human subjects. However, many questions remain about the quality of participation across individual games. In this paper, we use data on user behavior from a unique dataset of 40,000 game sessions across 14 studies to explore what factors predict game completion, consent, and data quality among volunteer participants on a single platform called Volunteer Science.

1 Introduction

There is a growing community of scientists conducting research with human subjects online through the frame of games for science, citizen science, and the online lab. These online studies include human intelligence tasks like image tagging (Von Ahn et al., 2008; Raddick et al., 2010); gamified problem solving (Khatib et al., 2011; Srensen et al., 2016); and behavioral experiments (Radford et al., 2016; Reinecke and Gajos, 2015; Germine et al., 2012). The power of these studies is leading many to develop online platforms for recruiting subjects.

The most common is the creation of paid online subject pools like Mechanical Turk, Prolific, and Crowdflower (Paolacci and Chandler, 2014; Peer et al., 2017; Vakharia and Lease, 2015) as well as volunteer pools like Zooniverse (Simpson et al., 2014), Lab in the Wild (Reinecke and Gajos, 2015) and TestMyBrain (Germine et al., 2012). Second, researchers are building study development platforms – codebases which make developing scientific games easy (McKnight and Christakis, 2016; Mao et al., 2012). Finally, others are combining the two into a single online lab platform (Radford et al., 2016).

2 Problem Overview

The promise of these platforms is that they offer a large pool of low-cost or free subjects for online research. A variety of studies have now shown that paid and unpaid subjects can reproduce many of same findings as traditional subject recruitment methods (Radford et al., 2016; Germine et al., 2012; Reinecke and Gajos, 2015; Peer et al., 2017; Crump et al., 2013; Rand, 2012).

A number of challenges have arisen for conducting research using online subject pools. Paolacci and Chandler (2014) find that online samples may be becoming non-naive and dishonest, resulting in poor data with less statistical power (Chandler et al., 2015). Peer et al. (2017) find that workers on Mechanical Turk were more attentive, but less naive and more dishonest than workers on Crowdflower and Prolific. Finally, Zhou and Fishbach (2016) show that selective attrition – different subjects quitting different games at different times – can lead to spurious results.

3 Study

In this study, we use a unique dataset of 40,000 game sessions across 14 studies conducted with volunteers on Volunteer Science. These studies, originally published online (Radford et al., 2016), include canonical psychology, economics, sociology, and computer science experiments. We supplement this data with ongoing data collection as well as unpublished demographic and survey data.

We explore three questions: what predicts frequent participation in multiple studies (i.e. who are the power users?); what predicts data quality including completion, consent, cheating, and repeat players; and how do frequent participation

and high and low quality participation affect the original results.

4 Results

4.1 Participation and Quality

In all, 25,021 unique participants have played 41,872 games, consenting to 25,101 having games donated for science. 6,242 participants played more than one game and 1,104 returned for more than one session, which we define as playing a game after at least an hour break. The mean number of games a user plays is 1.67 with a variance of 3.8. The average participant engages in 1.05 sessions with a variance of .10, which we define returning to play a game after at least an hour break. This means that, as with most websites, most participants engage once and then never return.

Tests designed by Clauset et al. (2009) and implemented by Gillespie (2015) show the distribution of repeat game playing fit a log-normal distribution. Sessions fit a power-law distribution. Both patterns fit with activity on other online platforms like Wikipedia edits, online comments, and friend-making (Geiger and Halfaker, 2013; Wilkinson, 2008). The reason, we believe, game playing follows a log-normal curve is that the inclination to play many games begins to decrease over time as players run out of games to play, causing a dip in activity at higher levels of engagement. However, people returning for more sessions don't display this degradation in participation.

We divide our users into three groups: returners, explorers, and one-timers. Returners (n=1,104) are those who participate in more than one session. Explorers (n=5,138) are users who participate in only one session, but who play more than one gasme in that session. Finally, one-timers are players who play once and never return. Returners and explorers are both more likely to do the things we want our participants to do. First, they consent to their games more often: returners (81.2%), explorers (68.8%), one-timers (43%). They are also more likely to participate in more games per session: returners (2.37 games), explorers (3.33 games), and one-timers (1.0 game by definition). Finally, they play more types of games: returner (2.54 types), explorer (2.26 types), one-timer (1.0 by definition).

Returners and explorers are also, unfortunately, more likely to exhibit what we call "adversarial behavior:" providing illogical answers, responding more quickly to surveys than is feasible (less than ten seconds), and filling out surveys with the same answer throughout (e.g. selecting "Agree Strongly" for all questions). In all, we observed 280 instances of adversarial behavior from 197 participants (0.7% of all participants). Of these, 127 were explorers and 48 were returners. Only 22 were one-timers. For those who exhibited adversarial behaviors, returners did so in only 19% of their games while explorers did so on 32% of their games. Those who provide bad data do not do so on all of their games, but a fairly small percentage of their games. This indicates that the few participants engaging in these behaviors are testing the system rather than attempting to undermine it.

4.2 Effect on Prior Results

Our second set of questions involved whether different kinds of users produced different kinds of data. We re-ran the analysis in (Radford et al., 2016) comparing data from repeaters, explorers, and one-timers. We find substantial differences in the behavior of explorers and repeaters versus one-timers.

One substantial difference we found is for users in social dilemma games. We did find what Chandler et al. (2015) report: the effect size for repeat participants is smaller than the effect size for one-timers. In these experiments, this reduction is not due to repeatedly playing the same game. It is more a feature of returners and explorers preferences, specifically returners and explorers are much more likely to defect than one-timers. This reduces the effect size by compressing the variance in returner and explorer behavior across experimental conditions. In other words, even though we vary the payoffs in prisoners dilemma and commons experiments substantially, returners and explorers are uniformly more likely to defect (use the commons or testify) than one-timers.

This willingness to defect may indicate a lack of seriousness on the part of returners and explorers. Perhaps the rewards and punishments of these dilemmas are less salient to repeaters? This does not seem to be the case. When we examined their behavior on the reaction time experiments, repeaters and explorers performed much faster than one-timers and, when repeating the experiment, improved their reaction time by almost 100 milliseconds on average. This indicates a substantial increase in quality engagement with the reaction

time experiments. More likely then, repeaters and explorers appear to have systematically different tastes in social dilemmas than one-timers.

We find no systematic differences in the remaining studies. Explorers, Returners, and one-timers perform the same on the anchoring effect and disease problem questions. On Timed risk reward, where we expect to see a negative correlation between the perceived risks of a technology (in our study we use bicycles, pesticides, chemical plants, and alcohol), the effects are slightly reduced for returners and explorers on some technologies (alcohol) but larger in others (pesticides).

We see a similar pattern of inconsistent results in the big five personality survey. In the original study, we independently validated the five dimensions. Only two items failed to load on their appropriate dimension (routine and inartistic). Using data for returners, two additional items failed to fit on their target dimension: planning and assertiveness. For explorers, two different items failed to load: depression and fault-finding. Finally, for one-timers, three items failed to load in addition to the two original items: quarrelsomeness, efficiency ,assertiveness.

The lack of a consistent pattern in the big five and the behavioral economic studies indicate the differences found are likely the result of error rather than any persistent differences between the three types of participants.

5 Discussion

Platforms for scientific games offer a promise to researchers that recruiting subjects collectively will reduce the costs and increase the quality of research with human participants. However, research using crowdwork platforms like Mechanical Turk indicate that participants may professionalize, becoming aware of the study conditions and either not engaging in them fully or learning to game the system.

In this study, we divided our participant pool into three distinct group: one-timers, returners, and explorers in order to determine the behavior of participants in each group and their effect on particular studies. Returners and explorers were both more likely to do the things scientist need them to do: participate in many studies from to beginning to end and sign the consent form. These participants were also more likely to provide poor quality data than one-timers. However only 0.7% of users ever behaved in such an adversarial way and those that did behave adversarially did so infrequently.

Thus, returners and explorers, those participants most likely to engage in many studies on a single platform are generally a well-behaved subject pool. However, there may be some problems with sharing these studies for some study types. While returners and explorers generally replicated the results found in behavioral economics and personality, they demonstrated increased engagement in the cognitive tasks (Flanker and Stroop) and were substantially more likely to defect in social dilemma games than one-timers.

This variability in the differences between one-timers and explorers and returners indicates that the advantages or disadvantages of sharing subjects are not uniform. It may benefit some study types and hurt others. More research is needed to understand the differences observed here. For example, it is possible that the decreased reaction time is the effect of practice or concentration which should be higher among returners and explorers. It could be that returners and explorers are more likely to be younger than one-timers. For the social dilemmas, a similar practice effect may be occurring. Alternatively, returners and explorers may be more aware of the fact that they are playing with bots and so may not perceive a social stigma against defecting.

These results indicate that collectivizing subject recruitment through shared platforms does lead to repeat participation from high quality participants. With appropriate data quality controls, it is possible to recruit subjects openly on these platforms. However, there are some cases where sharing subjects may lead to highly skewed results and more research is needed to understand the conditions in which this occurs.

References

Jesse Chandler, Gabriele Paolacci, Eyal Peer, Pam Mueller, and Kate A. Ratliff. 2015. Using nonnaive participants can reduce effect sizes. *Psychological science*, 26(7):1131–1139.

Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. 2009. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703.

Matthew J. C. Crump, John V. McDonnell, and Todd M. Gureckis. 2013. Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE*, 8(3):e57410, March.

R. Stuart Geiger and Aaron Halfaker. 2013. Using Edit Sessions to Measure Participation in Wikipedia. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pages 861–870, New York, NY, USA. ACM.

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, October.

Colin S. Gillespie. 2015. Fitting Heavy Tailed Distributions: The poweRlaw Package. *Journal of Statistical Software*, 64(2):1–16.

Firas Khatib, Seth Cooper, Michael D Tyka, Kefan Xu, Ilya Makedon, Zoran Popovi, David Baker, and Foldit Players. 2011. Algorithm discovery by protein folding game players. *Proceedings of the National Academy of Sciences*, 108(47):18949–18953.

Andrew Mao, Yiling Chen, Krzysztof Z Gajos, David C. Parkes, Ariel Procaccia, and Haoqi Zhang. 2012. Turkserver: Enabling synchronous and longitudinal online experiments. In *Proceedings of the Fourth Workshop on Human Computation (HCOMP'12)*. AAAI Press.

Mark E McKnight and Nicholas Christakis. 2016. Breadboard: Software for Online Social Experiments, May.

G. Paolacci and J. Chandler. 2014. Inside the Turk: Understanding Mechanical Turk as a Participant Pool. *Current Directions in Psychological Science*, 23(3):184–188, June.

Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70:153–163, May.

M. Jordan Raddick, Georgia Bracey, Pamela L. Gay, Chris J. Lintott, Phil Murray, Kevin Schawinski, Alexander S. Szalay, and Jan Vandenberg. 2010. Galaxy Zoo: Exploring the Motivations of Citizen Science Volunteers. *Astronomy Education Review*, 9(1):010103.

Jason Radford, Andy Pilny, Ashley Reichelmann, Brian Keegan, Brooke Foucault Welles, Jefferson Hoye, Katherine Ognyanova, Waleed Meleis, and David Lazer. 2016. Volunteer Science. *Social Psychology Quarterly*, 79(4):376–396.

David G. Rand. 2012. The Promise of Mechanical Turk: How Online Labor Markets can help Theorists run Behavioral Experiments. *Journal of Theoretical Biology*, 299:172–179, April.

Katharina Reinecke and Krzysztof Z. Gajos. 2015. LabintheWild: Conducting Large-Scale Online Experiments With Uncompensated Samples. pages 1364–1378. ACM Press.

Robert Simpson, Kevin R. Page, and David De Roure. 2014. Zooniverse: observing the world's largest citizen science platform. In *Proceedings of the 23rd international conference on world wide web*, pages 1049–1054. ACM Press.

Jens Jakob WH Srensen, Mads Kock Pedersen, Michael Munch, Pinja Haikka, Jesper Halkj\a er Jensen, Tilo Planke, Morten Ginnerup Andreasen, Miroslav Gajdacz, Klaus Mlmer, Andreas Lieberoth, and others. 2016. Exploring the quantum speed limit with computer games. *Nature*, 532(7598):210–213.

Donna Vakharia and Matthew Lease. 2015. Beyond Mechanical Turk: An analysis of paid crowd work platforms. *Proceedings of the iConference*.

Luis Von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. 2008. reCAPTCHA: Human-based character recognition via web security measures. *Science*, 321(5895):1465–1468.

Dennis M. Wilkinson. 2008. Strong Regularities in Online Peer Production. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, EC '08, pages 302–309, New York, NY, USA. ACM.

Haotian Zhou and Ayelet Fishbach. 2016. The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, 111(4):493–504.