Volunteer Science: An Online Laboratory for Experiments in Social Psychology

Jason Radford$^{1,2}$, Andy Pilny$^3$, Ashley Reichelmann$^2$, Brian Keegan$^4$, Brooke Foucault Welles$^2$, Jefferson Hoye$^5$, Katya Ognyanova$^6$, Waleed Meleis$^2$, and David Lazer$^2$

Abstract

Experimental research in traditional laboratories comes at a significant logistic and financial cost while drawing data from demographically narrow populations. The growth of online methods of research has resulted in effective means for social psychologists to collect large scale survey-based data in a cost-effective and timely manner. However, the same advancement has not occurred for social psychologists who rely on experimentation as their primary method of data collection. The aim of this paper is to provide an overview of one online laboratory for conducting experiments, Volunteer Science, and report the results of six studies which test canonical behaviors commonly captured in social psychological experiments. Our results show that the online laboratory is capable of performing a variety of studies with large numbers of diverse volunteers. We argue the online laboratory is a valid and cost-effective way to perform social psychological experiments with large numbers of diverse subjects.

Keywords

Online platform, experiments, replication, reliability

$^1$University of Chicago, Chicago, IL, USA  
$^2$Northeastern University, Boston, MA, USA  
$^3$University of Kentucky, Lexington, KY, USA  
$^4$University of Colorado, Boulder, Boulder, CO, USA  
$^5$Jefferson Hoye LLC, Arlington, VA  
$^6$Rutgers University, New Brunswick, NJ, USA

Corresponding Author:  
Jason Radford, Department of Sociology, University of Chicago, 5828 S. University Avenue, Chicago, IL 60637, USA.  
Email: jsradford@uchicago.edu
Social psychological experiments have relied upon brick-and-mortar laboratories to produce reliable results. However, some argue that the utility of these studies as an empirical check of general theoretical principles is constrained by narrow participant demographics, high costs, and low replicability (Ioannidis 2005, Open Science Collaboration 2015).

Two decades of research using the Internet to recruit subjects and deploy studies demonstrates that online methods improve subject recruitment by substantially expanding and diversifying our sample pool and allowing for standardized research designs, data collection, and data analyses that can more easily be shared, replicated, and extended (Reips 2000, Open Science Collaboration 2015).

The aim of this paper is to present Volunteer Science as an online laboratory for social and behavioral science experiments. This paper will describe our approach to the online laboratory and the methodological contribution it makes: bridging an online subject pool with shared code for experiments. Most importantly, we report the results of six studies, which we use to validate our approach by testing whether core social psychological experimental studies and results can be achieved by recruiting online volunteers into our online laboratory.

Background

Experiments are the hallmark of social psychology as a discipline, and have traditionally been used as a methodological tool of theory testing. Experiments are “an inquiry for which the investigator controls the phenomena of interest and sets the conditions under which they are observed and measured” (Willer and Walker 2007:2). The primary benefit of an experiment is the unique control the researcher has over condition, its artificiality (Webster and Sell 2007). By controlling known factors, experiments isolate the relationship between independent and dependent variable. Such control makes experiments fundamentally different than any other data collection format in the social sciences (Willer and Walker 2007), allowing a direct comparison between the presence of a condition and its absence (Webster and Sell 2007).

While the utility of artificiality remains the same, two forces have pushed researchers to improve experimental methods. First, studies demonstrating the validity and power of online research have pushed researchers to adapt paradigms to online contexts where large samples from many populations can be recruited effectively (Reips 2000; Gosling et al. 2010; Mason and Suri 2011; Kearns 2012; Crump, McDonnell, and Gureckis 2013).

The strength of large, diverse samples made possible by online methods lies not in their heterogeneity, but in their many homogenous samples. Larger and diverse samples provide the ability to test populations as moderating variables, therefore expanding our ability to assess the role that factors like culture and location play on the applicability of theory. Although experiments using large and diverse samples are still uncommon, some recent articles in SPQ have featured cross-societal experiments (Cook et al. 2005) and cross-national experiments (Kuwabara et al. 2007).
Second, the replication crisis in a range of fields has led to demands for higher methodological standards and reporting practices (Ioannidis 2005, Open Science Collaboration, 2015; Pashler and Wagenmakers 2012). The standards being put forward require significant investments in experimental methods which, we argue, can be met in part through the subject recruitment, technical standardization, and the transparent sharing enabled by online labs.

Computational technology has improved the effectiveness and efficiency of methods for collecting and analyzing data (Lazer et al. 2009). Early research using online platforms and recruitment methods showed that most studies can be validly performed online (Mason and Suri 2011; Rand 2012; Reips 2000; Weinberg, Freese and McElhattan 2014; ).

In addition, researchers have used online platforms to develop new paradigms. Social scientists have developed internet-based studies of markets, networks, and multi-team systems (Salganik and Watts 2008; Davison et al 2012; Mason and Watts 2012). Furthermore, researchers have used the Internet to attract thousands of volunteers through “citizen science” platforms to collect and analyze large scale data (Christian et al. 2012; Raddick et al. 2010; Sauermann and Franzoni 2015; Von Ahn et al. 2008). This body of work demonstrates that a wide variety of social science research can be validly conducted online for a fraction of the cost of traditional experiments and with more diverse samples of participants.

The second shift, brought about by the replication crisis, has been to increase the standards for performing experiments, reporting results, and sharing instruments and data. Recommendations for addressing the replication crisis involve increasing sample sizes, sharing data and study materials, and performing independent verification (Ioannidis 2005; Begley and Ellis 2012; Pasher and Wagenmakers 2012). Technological advances in online data collection can reduce the cost and logistical burden for recruiting larger sample sizes, provide transparency for methods, and ensure high fidelity access to study materials and data for validation and replication. Online methods make these practices more feasible, increasing the possibility that they will become standard in the field.

At present, online experiments still require a great deal of technical expertise to create in addition to significant investments in subject recruitment and management. This makes independent replication by other researchers difficult. Thus, the present decentralized, ad hoc approach to building online experiments furthers the replication crisis.

To solve these challenges, we created Volunteer Science in the mold of an online laboratory. In what follows, we describe how Volunteer Science reduces the cost of creating experiments and recruiting subjects, maximizes subject diversity, and promotes research material and data sharing. After that, we report the results of a wide-ranging series of studies we performed to test the validity of the online laboratory model.

**Volunteer Science: An Online Laboratory**

Volunteer Science (volunteerscience.com) is unique in that it combines a platform for developing online experiments with a website for recruiting subjects. Current facilities for online
research only provide one of these. Crowdwork platforms like Amazon’s Mechanical Turk and Crowdflower and programs like TESS provide access to subjects, but do not come with their own tools for creating experiments. Conversely, Vecon Lab (Holt 2005), Z-tree (Fischbacher 2007), Breadboard (McKnight and Christakis 2016), and Turkserver (Mao et al. 2012) offer code for developing experiments. However, researchers must deploy these systems and recruit subjects on their own. Volunteer Science offers a toolkit, study deployment, and subject recruitment all in the same system.

Research on Volunteer Science

For researchers, Volunteer Science provides experiment templates and an Application Programming Interface (API). There are currently more than twenty experiment templates (including the studies reported in this paper) researchers can use to build their own experiments. Researchers can also use the API to add functionality like collecting Facebook data, subject randomization, and creating a chatroom. By providing starter experiments and an API, Volunteer Science can significantly reduce the time, technical expertise, and cost associated with creating online experiments.

Second, Volunteer Science was designed to be a stable environment with open data policies which support study verification and replication. As a shared platform, Volunteer Science standardizes the environment, meaning a study can be shared, re-implemented, and re-run in Volunteer Science without any changes to the code. In addition, researchers are required to share their data and code once a study is completed. This enables other researchers on Volunteer Science to easily verify the original analysis, replicate a study, and extend the work of others in ways that remain faithful to the original design. In fact, all experiment code, data, and analytic code for this study is posted on Dataverse (Radford et al 2016).

Participating in Volunteer Science

As a website, Volunteer Science is created to maximize the number and diversity of people participating in experiments. It is built on open source tools, including HTML5, Javascript, Django, and Bootstrap. This enables anyone in the world with modern Internet browsing technology to access and participate in Volunteer Science at any time. The site is deployed on an Amazon server that can support up to 1000 users per hour, and 50-75 concurrent users without system lag. With these specifications, the system can effectively handle millions of users per year.

The experience is designed to be light, engaging, and intrinsically rewarding. Building off the success of projects like reCAPTCHA (Von Ahn et al 2008), we try to harness a small piece of the massive amounts of activity individuals do every day: online gaming. Most studies are presented as games, often including awards and scores. In addition, our studies generally require less than a minute of training and typically last no more than five minutes.

One central design choice we made to encourage volunteer participation was implementing a post-hoc “data donation” consent paradigm whereby volunteers participate in
experiments and then consent to donate that data afterward. For example, when this study was running, after a volunteer filled out a personality survey, we opened a pop-up and ask them whether or not they want to donate that data to this study. Researchers can collect data from their research instruments, but cannot use the data until volunteers have donated it to their study.

In addition, we restrict the use of deception to special sections of Volunteer Science where volunteers know they may be deceived because deception can erode the trust of the volunteer community and can be undermined by off-site discussions which are difficult to monitor.

Finally, for studies involving compensation, researchers have three options. First, they can collect subjects’ email addresses and then pay them using an online service like PayPal. Researchers can also recruit local volunteers like students who can physically show up to collect their payment. Finally, Volunteer Science provides direct access to Mechanical Turk, enabling researchers to pay Turkers to complete a study.

Validation Methodology

We conducted several studies to validate that Volunteer Science can produce the promised volume and diversity of volunteers while reproducing well-regarded results from brick-and-mortar laboratory experiments.

Study Selection

We decided to replicate six foundational studies for capturing different aspects of human behavior. The first study involves two experiments testing participants’ reaction times, which are essential for priming, memory, and implicit association research (Crump et al 2013). Our second study replicates several behavioral economics experiments to show that volunteers make common yet counter-intuitive decisions indicative of practical judgment (Kahneman 2003). Our third study reproduces the big five personality survey which we use to determine whether or not researchers are able to validate surveys using volunteers on Volunteer Science. Fourth, we implement studies of social influence (Nemeth 1986) and justice (Kay and Jost 2003) to evaluate the extent to which online laboratories can deliver social information. Fifth, we test group dynamics through problem solving, specifically the travelling salesperson problem. Last, we test subjects’ susceptibility to change in incentives using the prisoner’s dilemma, commons dilemma, and public goods paradigms.

Subject Recruitment

Each of these studies was created as a game or survey on the Volunteer Science website. Subjects were recruited to the website to participate in experiments for social scientific research. Only those who participated in each study and donated their data are included in the analysis.

We used a variety of outlets to reach volunteers both online and offline. Online, we posted recruitment messages to Twitter, Facebook, and Reddit. We also ran ads on Facebook and
Twitter. Offline, we created a certification system such that students can participate in experiments for class credit. This recreates one of the primary modes of recruitment for offline laboratories studies. Faculty can see the experiments completed, time spent and validate students’ certificates. Since August 2014, users have created 481 certificates.

**Participants**

Volunteers are welcome to participate in studies with or without an account on Volunteer Science. A browser cookie tracks participation across studies for people without an account. People with an account additionally have demographic information such as gender and age. Browser’s languages and device type are recorded for all participants.

Overall, we recruited 15,915 individuals to participate in 26,216 experimental sessions. Half of our participants were female and the average age was 24 years old. Ninety-two percent of participants used English as their browser language and 95 percent of participants used desktop computers. The average person engaged in two experimental sessions, and consented to donating their data just over half the time.

For those who signed in with Facebook, we found no difference in the probability of consenting by age (t = -0.52, p = 0.60) or gender (77% of males donated vs 75% of females, chi-squared = 0.89, p = 0.35). We did find significant differences in those using English-language browsers and those using other languages (44 vs. 58 percent respectively, chi-squared = 188.0, p < .001), and those only using desktop computers (47 percent) vs those use mobile devices (43 percent chi-squared = 18.5776, p < .001) are more likely to donate their data.

Consenting participants were more likely to participate in multiple experiments than non-consents (2.6 vs. 1.6 experiments respectively, t = -25.5, p < .001). There were no differences in participation by gender (t = -1.38, p = 0.17) or age (t = 1.06, p = 0.29). However, users using languages other than English or mobile devices donated more data than those who were using English-language browsers (t = 4.18, p < .001) and desktop computers (t = 4.01, p < .001).

**Results**

**Study 1: Reaction Times**

First, we replicate two reaction-time studies which elicit the Stroop and flanker effects (MacLeod 1991; Eriksen 1995). Measures of human reaction time are essential to a range of studies including implicit association, working memory, and perception. However, there is a question of whether online experiments can detect small reaction time differences given delays in computational processing and communication and subjects’ attention-span. The advantage of using these two tests is that they differ in time sensitivity. In traditional laboratory studies, the Stroop effect produces a 100-200ms delay in reaction while the flanker effect produces a 50-60ms delay (Crump et al. 2013). By replicating both, we test how precisely the Volunteer Science system can validly measure reaction time.
The Stroop and flanker experiments both test the effect of cognitive interference generated by incongruent contextual information. In Stroop, subjects are asked to identify the color of a word; however, the words themselves are colors. For example, in a congruent prompt, the word "blue" would be colored blue while, in an incongruent prompt it is displayed in another color like red (MacLeod 1991). In the flanker experiment subjects are asked to identify the letter in the middle of a string of five letters. An example of a congruent prompt would be the letter ‘h’ flanked by ‘h’ (i.e. “hhhhh”) while an incongruent prompt would be ‘f’ flanked by ‘h’ (i.e. “hhfhh”) (Eriksen 1995). In both experiments, the hypothesis is that subjects will show a significantly delayed reaction when given incongruent information.

In total, volunteers participated in 1,674 sessions of Stroop and 1,721 sessions of flanker. Of these 970 Stroop sessions and 1,049 flanker sessions were donated to science, were the subjects’ first session, and met our basic data quality requirements for completeness and accuracy.

The results show a significant delay in incongruent conditions for Stroop \((t = -29.41 \ p < .001)\) and flanker \((t = -10.13, p < .001)\). For Stroop, the mean response time was 951.3ms for congruent and 1141.4ms for incongruent stimuli \((t = -29.41 \ p < .001)\). For flanker, the mean response time is 689.6ms for congruent and 752.7ms for incongruent stimuli.

This represents a direct replication of prior experimental results and suggests that the Volunteer Science system can support reaction-time tests to the tens of milliseconds. However, there is a uniform increase in reaction times of about fifteen percent across all conditions for both experiments than found in traditional laboratory settings. For example, Logan and Zbrodoff (1998: 982) report a mean of 809ms for congruent stimuli and 1,023ms for incongruent stimuli.

**Study 2: Cognitive Biases and Heuristics**

Studies of biases and heuristics pioneered by social psychologists and behavioral economists examine how humans make decisions. Empirical studies of human decision-making have been critical to understanding the role factors like social identity, emotion, and intuition play in everyday life (Bechara and Damasio 2005; Kahneman 2003; Stangor et al 1992). We implement four studies taken from Stanovich and West’s (2008) recent comprehensive review. Our purpose is to examine whether or not volunteers make counter-intuitive decisions indicative of practical judgment.

First, we implemented Tversky and Kahneman’s Disease Problem (1981) which asks subjects to choose between a certain or probabilistic outcome. In the “positive” frame, the certain outcome is posed as “saving the lives of 200 people” from a disease out of a total of 600 people or having a one-third probability of saving all 600 people. In the “negative” frame, the certain outcome is “letting 400 people die” and the probabilistic outcome is a one-third probability “no one will die.” Tversky and Kahneman find that people choose the certain outcome in the positive condition and the probabilistic outcome in the negative frame, even though they are equivalent dilemmas.
Second, we implemented two experiments which elicit anchoring effects whereby people’s judgements are biased based on prior information. In one version, we ask “How many African countries are in the United Nations?” In the second, we ask “How tall is the tallest redwood tree in feet?” Users are anchored by our suggestions. In the small condition, we suggest there are 12 countries or that the tallest redwood is 85 feet. In the large anchor condition, we suggest there are eighty countries and that the tallest tree is 1,000 feet. For each question, individuals are randomly assigned to either the small or large anchor, and then asked to estimate a response value to the initial question. Prior work shows that participants will give smaller estimates following a small anchor, and larger estimates following a large anchor.

Third, implemented the timed risk-reward experiment. Finucane et al. (2000) show that, under time pressure, people tend to judge activities they perceive to be highly rewarding to have low risk and, conversely, those that are highly risky to have low reward. Following their methods, we give respondents six seconds to rate the risks and benefits of four items on a seven point Likert scale (bicycles, alcoholic beverages, chemical plants, and pesticides).

Subjects participated in these individually. In total, volunteers participated in 688 sessions of the Disease Problem and 455 met our consent and data quality inclusion requirements. Volunteers participated in 1,076 sessions of risk-reward and 457 met the same requirements. Finally, there were 1,422 anchoring sessions, 710 of the country version and 689 of the tree versions, and 814 met our requirements.

![Cognitive Bias Study Results](image)

**Figure 1: Cognitive Bias Study Results**

The results, shown in Figure 1, replicate each of the three tests. For the disease experiment, people chose the certain outcome 60 percent of the time when given the positive frame, but only 39 percent given the negative frame (Odds = 2.28, p < .001 in Fisher’s Exact Test). These results are weaker than Tversky and Kahneman’s original findings of a switch from 72 percent to 22 percent (1981: 453).

For the African countries anchor, the mean estimates in the small and large prompts (twelve and eighty) were 22 and 41 countries respectively (F(1, 178) = 71.0, MSE = 37053, p < .001). For the redwood anchor, the mean estimates in the small and large prompts (85 and 1,000 feet) were 212 and 813 feet (F(1,179) = 158.6, MSE = 34307016, p < .001). These generally
align with Stanovich and West’s results which were 14.9 and 42.6 countries and 127 and 989 feet (2008: 676).

Finally, for risk-reward, the correlation between risk and reward was negative and statistically significant for every item except bicycles (Finucane et al. 2000; Stanovich and West 2008). Again, our results are weaker than Finucane et al. (2000: 7): -.07 and .02 for bicycles, -.30 and -.71 for alcohol, -.27 and -.62 for chemical plants, and -.33 and -.47 for pesticides, respectively.

Study 3: Validating the Big Five Personality Survey

Our third study investigates the viability of using Volunteer Science to develop multidimensional survey-based measures of individual characteristics like personality, motivation, and culture. For this study, we attempted to independently validate the forty-four item version of the five-factor model of personality, called “the big five.” The five-factor model was chosen because it has proven to be robust over a number samples drawn from diverse populations (McCrae and Terracciano, 2005; Schmitt et al., 2007).

The survey had been taken 852 times and 584 surveys fit our inclusion requirements of being complete, valid, and the participant’s first completion. The Cronbach's alpha values, which measure the consistency of subjects’ responses across items within each factor, were acceptable: .78 for Openness, .83 for Neuroticism, .87 for Extraversion, .78 for Agreeableness, and .84 for Conscientiousness. We also ran an exploratory factor analysis with varimax rotation and five factors. The result replicates a big five structure, with high positive loadings for all but two items, routine (Openness) and unartistic (Openness), on the predicted factor.

Study 4: Justice and Group Influence

Complementary Justice

Our fourth study looks to induce two essential forces studied by social psychologists: individual’s sense of justice and group influence. First, we implemented study three from Kay and Jost (2003: 830-31) to investigate whether Volunteer Science could prime participants' sense of justice and measure the detect the prime through implicit and explicit measures. In the study, students are presented with a vignette about two friends named Mitchell and Joseph. In one version, Joseph “has it all” while Mitchell becomes “that broke, miserable guy.” In the other version, Joseph is “rich, but miserable” and Mitchell is “broke but happy.” Kay and Jost found that subjects who were exposed to the first scenario responded more quickly to words related to justice in a subsequent lexical task and had higher scores on a system justification inventory conditional on their having a high score on the Protestant Work Ethic scale.

We implemented the vignette, lexical task, the Protestant Work Ethic (PWE) scale, and system justification (SJ) inventory described by Kay and Jost. Subjects were randomly assigned to either the complementary or non-complementary vignettes and then continued to participate in the subsequent three tasks.
Volunteers started the vignette 1691 times and 540 unique individuals completed all four tasks in the Kay and Jost protocol on Volunteer Science. In total, 464 (85.8 percent) were complete, valid, done on desktops, and the participant’s first experiment. We replicated the main effect of the protestant work ethic on system justification (F(1,133) = 37.4, MSE 29.3, p < .001). However, we found no evidence that the experimental condition affected participants’ reaction time for justice-related words (F(1,133) = .02, MSE = 0.008, p = 0.89) or their system justification score (F(1,113) = 1.81, MSE = 1.81, p = .131). This indicates we were unable to prime participants’ sense of justice.

*Group Influence Experiment*

We also implemented a version of Nemeth’s group influence study (1986) to investigate whether subjects would respond to simulated group influence. In the original study, individuals are placed in a group of six with either two or four confederates and two or four subjects and asked to solve a graphical problem. After solving the problem and sharing the results, participants are given the chance to solve the problem again. The experimental manipulation involves having four or two confederates (the “majority” and “minority” conditions) give correct or incorrect responses. Nemeth showed that subjects in the minority correct condition tend to increase the number of correct responses in the second round, while subjects in the majority condition tend to follow the majority. In our version, we simulate the responses of all five participants and have the non-confederate subjects only give the easy, correct answer.

Volunteers participated in 1,188 sessions and 515 experiments met our inclusion requirements. As a test of validity, we found that participants exposed to correct answers were more likely to include those answers in the second round (F(1,384) = 9.59, MSE=3.02, p < .01). Contrary to the original result however, individuals in the majority condition were no more likely to converge to the majority opinion (F(1,384)=0.64 MSE = .09, p = .42). And, there was no evidence that subjects in the minority condition found more unique solutions than subjects in the majority condition (F(1,201) = .57, MSE = .08, p = .45).

*Study 5: Problem Solving*

Experiments based on collective problem-solving are essential to studies of group behavior in social psychology (Hackman and Katz 2010). However, problem solving is a complex task, making it difficult to train subjects in online settings. We test whether such research can be done with volunteers by examining how they solve a commonly used puzzle, the traveling salesperson (TSP) (MacGregor and Chu 2011).

In our implementation, we provide users with a two-dimensional Cartesian plane with 20 dots (‘cities’) and ask users to connect the dots in a way that minimizes the distance “travelled between cities.” Users are given ten rounds to minimize their distance. Existing research shows that the most difficult maps are those with more dots clustered in the middle of the space inside the interior convex hull (MacGregor and Chu 2011).
Volunteers participated in 7,366 sessions with maps containing between nine and fifteen dots inside the interior hull. Of these, 3,107 met our inclusion requirements. Consistent with prior results, we estimate the correlation between the number of cities and number of correct edges to be -0.09 (p < .001), meaning the number of edges guessed correctly decreases as the number of cities inside the convex hull increases.

**Study 6: Social Dilemmas**

Studying individual decision making and collective bargaining are central to research on social exchange and the development of social norms (Cook and Rice 2006, Suri and Watts 2011). The central premise is that participants are sensitive to incentives. However, the challenge for online research with volunteers is that the lack of payment may make subjects insensitive to incentives. We used the prisoner’s dilemma, commons dilemma, and public goods dilemma to test whether subjects would behave differently if we randomly assigned them to different incentive schemes.

In the each of these dilemmas, users must choose to cooperate or defect from a partner and are rewarded based on the combination of their choice and the choice of other players. In the prisoner’s dilemma (PD), individuals must choose between testifying against their partner or not. In the commons dilemma (CD), individuals choose to either use a private resource providing fewer but certain benefits or a common resource providing more but uncertain benefits. In our case, users are deciding whether to feed their cows from their barn or from a common pasture. For PD and CD, the incentives are such that individuals should always defect while the collective good is maximized only when everyone cooperates. In our experiments, we maintain this dilemma structure, but change the size of the trade-offs for cooperating or defecting (see Table 1). We are not explicitly replicating a prior study. Instead, we attempt to test whether subjects respond to differing incentives in the expected ways.

**Table 1.**

<table>
<thead>
<tr>
<th>Payoff Matrices for Social Dilemmas</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prisoner’s Dilemma Payoffs</strong></td>
</tr>
<tr>
<td>Condition</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td><strong>Commons Payoffs</strong></td>
</tr>
<tr>
<td>Condition</td>
</tr>
<tr>
<td>-----------</td>
</tr>
</tbody>
</table>
With the public goods game (PGG), we do look to replicate prior findings. In PGG, individuals receive a set amount of (simulated) money each round and must contribute a portion of it to common pot. At the end of the round, they receive a percentage of interest based on how much money is put into the pot. In this study, we wanted to see whether subjects would replicate prior findings regarding the overall average contributions and distribution of “free-riders.”

Volunteers participated in 825 sessions of PD of which 236 sessions met our inclusion requirements, 4,145 sessions of CD with 3,008 meeting our requirements, and 532 sessions of PGG with 466 meeting our requirements.

For the PD, the results of a pairwise tests show significant differences in subjects’ average choice to cooperate or defect across conditions (t = 2.42 p = .016). For the CD, pairwise tests of subjects’ average choice shows that each is significantly different from the other: conditions 1 and 2 (t = 9.43 p < .001), conditions 2 and 3 (t = 4.40 p < .001), conditions 3 and 4 (t = 2.24 p = .025). These results show that volunteers respond to incentives in the expected (i.e. monotonic) way.

In the public goods game, we find volunteers donated 46.5 percent of their endowment in the initial round and contributed less (t = 2.28, p = .02) in the final round (M = 4.21) than they did in the first (M = 4.65). This was consistent with Ostrom (2000: 140). And consistent with Gunnthorsdottir, Houser, and McCabe (2007: 308) we also found that 32.6 percent of volunteers were free-riders and 67.4 percent contributors. As such, we find some support that individuals played the PGG online in a similar fashion as if they would have played it offline.

**Discussion**

On the whole, the findings from each of these experiments support the validity of using an online laboratory to conduct research in social psychology. We are able to recruit thousands of volunteers from around the world to participate in and donate experiment results. Using questionnaires, we can validate multidimensional inventories and elicit behaviorally realistic responses to tests of cognitive bias as well as induce and measure low latency reaction times. And, our participants engage in economic trade-offs and puzzle solving in ways found in a variety of other research. We were unable, however, to prime users’ sense of justice using a complementary justice vignette or deliver simulated group influences.
Validation and secondary analysis on the group influence experiment indicated that subjects were learning from their simulated group. The direction of the results held, but was not statistically significant, suggesting that the underlying effect may be weaker than first reported or that we failed to sufficiently simulate group influence. Similarly, in the justice study, we validly measured subjects’ explicit justice-related beliefs and the reaction time study demonstrated that we can detect valid reaction time differences. Yet, our vignette did not elicit the priming effect found by Kay and Jost. These results point to the potential need to create stronger social signaling in online contexts to activate the justice primes or the sense of peer pressure in online settings.

Overall, we found that game-based experiments attract much more participants than survey-based experiments. Therefore, social psychologists may experience more success with “gamified” online experiments than with experiments of other types. Studies on Volunteer Science work best when they are quick and engaging, and thus, experiments that require lengthy protocols may not be appropriate.

For this reason, it would be difficult to execute any experiment that is predicated on face-to-face interaction, nonverbal behavior, or the use of physical bodies and/or environments as experimental stimuli or data. Much of the work we have done with Volunteer Science to date either relies on single-person experiments, or on the use of computer agents (bots) in multi-person experiments. Although the Volunteer Science system can technically support experiments involving tens or even hundreds of participants in a single session, the logistics of recruiting and coordinating more than a few simultaneous participants have proven challenging to date.

In the future, we will continue to expand the kinds of research possible on Volunteer Science. For example, we are creating the capacity for users to donate social media data, browser data, and mobile phone data. In addition, we are in the process of developing a panel of participants among our volunteers to provide demographic control over the subjects recruited for new studies. A panel also enables us to link data across studies potentially providing the most comprehensive portrait of experimental participation available.

Finally, the future of this model rests on making it available as a common good for researchers. This entails creating a model of collaboration and openness which minimizes the barriers to entry while protecting users and their data and ensuring the transparency of scientific research. Collaboration is the heart of science, and deploying Volunteer Science as a common good requires developing systems which enable social scientists with limited technical training to access and contribute to the system. However, such openness has to be balanced with the requirements to meet standards for human subject protection, security, and usability. How this balance should be struck is itself an experiment which we are currently working to solve.

We introduce Volunteer Science as an online laboratory which can advance the social psychological research agenda by diversifying the sample pool, decreasing the cost of running online experiments, and easing replication by making protocol and data shareable and open. We have validated the system by reproducing a number of behavioral patterns observed in traditional social psychology research. Although Volunteer Science cannot entirely replace brick-and-
mortar laboratories, it can may allow researchers to achieve generalizable experimental results at a reasonable cost. Volunteer Science answers the call for researchers who are looking for a reasonable, valid, and efficient alternative to the brick-and-mortar lab.

**Funding**

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053 (the ARL Network Science CTA) and in part, by a grant from the US Army Research Office (PI Foucault Welles, W911NF-14-1-0672). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

**References**


Author Biographies

**Jason Radford** is a graduate student in sociology at the University of Chicago and the project lead for Volunteer Science. He is interested in the intersection of computational social science and organizational sociology. His dissertation examines processes of change and innovation in a charter school.

**Andrew Pilny** is an assistant professor at the University of Kentucky. He studies communication, social networks, and team science. He is also interested in computational approaches to social science.

**Ashley Reichelmann** is a PhD candidate in the Sociology Department at Northeastern University, focusing on race and ethnic relations, conflict and violence, and social psychology. She uses mixed methods to study collective memory, identity, and violence. Recently, her coauthored work on hate crimes and group threat was published in *American Behavioral Scientist*. Her dissertation project is an original survey-based experiment that explores how white Americans react to representations of slavery, for which she was awarded the Social Psychology Section’s Graduate Student Investigator Award.

**Brooke Foucault Welles** is an assistant professor in the Department of Communication Studies at Northeastern University. Using a variety of quantitative, qualitative, and computational methods, she studies how social networks provide resources to advance the achievement of individual, group, and social goals.

**Brian Keegan** is an assistant professor in the Department of Information Science at the University of Colorado, Boulder. He uses quantitative methods from computational social science to understand the structure and dynamics of online collaborations.

**Katherine Ognyanova** is an assistant professor at the School of Communication and Information, Rutgers University. She does work in the areas of computational social science and network analysis. Her research has a broad focus on the impact of technology on social structures, political and civic engagement, and the media system.

**Jeff Hoye** is a professional software engineer. He specializes in design and development of distributed systems, computer graphics, and online multiplayer computer games.

**Waleed Meleis** is an associate professor of electrical and computer engineering at Northeastern University and is associate chair of the department. His research is on applications of combinatorial optimization and machine learning to diverse engineering problems, including cloud computing, spectrum management, high-performance compilers, computer networks, instruction scheduling, and parallel programming.
David Lazer is Distinguished Professor of Political Science and Computer and Information Science, Northeastern University, and Co-Director, NULab for Texts, Maps, and Networks. His research focuses on computational social science, network science, and collective intelligence.