Social Psychology Quarterly
2016, Vol. 79(4) 376–396

DOI: 10.1177/0190272516675866
http://spq.sagepub.com

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], Katherine 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 article is to provide an overview of one online laboratory for conducting experiments, Volunteer Science, and report the results of six studies that 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 advocate for the use of the online laboratory as a valid and cost-effective way to perform social psychological experiments with large numbers of diverse subjects.

Keywords

online platform, experiments, replication, reliability

Social psychological experiments have relied on 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 sample pools and allowing for standardized research designs, data collection, and data analyses that can more easily

[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

be shared, replicated, and extended (Open Science Collaboration 2015; Reips 2000).

Heeding this call, some areas of social psychology research have already embraced and benefited from online methods. Survey-based research conducted using websites like Qualtrics and SurveyMonkey have been shown to be comparable to established industry standards like GfK (formerly Knowledge Networks) (Simmons and Bobo 2015; Weinberg, Freese, and McElhattan 2014). Similarly, the Time-Sharing Experiments for Social Scientists (TESS) program has been established to offer an outlet for survey-based experiments. However, the scientists who rely on experimental methods beyond surveys have not yet seen the same benefits from online research.

The aim of this article is to present Volunteer Science as an online laboratory for social and behavioral science experiments. This article 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 variables. 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 and diverse samples can be recruited effectively (Crump, McDonnell, and Gureckis 2013; Gosling et al. 2010; Kearns 2012; Mason and Suri 2011; Reips 2000). Large and diverse samples enable researchers to assess the generalizability of theoretical mechanisms through experimentation.

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 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 efforts to use online platforms and recruitment methods showed that most studies can be validly performed online (Reips 2000). The development of services like Qualtrics and GfK provide access to diverse, nationally representative subject pools (Simmons and Bobo 2015; Weinberg et al. 2014). And more recently, Amazon's Mechanical Turk service has proven to be a low-cost

source of engaged study participants (Mason and Suri 2011; Rand 2012).

In addition, researchers have used online platforms to develop new paradigms for research. Social scientists have developed online studies of markets, networks, and multi-team systems (Davison et al. 2012; Mason and Watts 2012; Salganik and Watts 2008). 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.

However, large and diverse samples are not necessarily a desirable feature for social psychological experiments. Experiments are intended to be deployed on homogenous samples in order to test theoretical nuances (Willer and Walker 2007). Sample homogeneity is one essential form of experimental control as diversity can complicate the isolation of the required condition. For example, one critique of the replication studies performed by the Open Science Collaboration (2015) is that many drew samples from different populations such as using Italians to replicate a study involving American attitudes toward African Americans (Gilbert et al. 2016).

There is a tradeoff however as the theoretical specificity allowed by homogeneity undermines its generalizability. As Henrich, Heine, and Norenzayan (2010) argue, findings based on experiments conducted with limited samples are improperly treated as broadly representative of human behavior. New avenues for large-scale experimentation have evolved to support general population (i.e., diverse sample) experiments as a response to these critiques. Advocates argue that these experiments are "particularly effective at documenting differences in the status of causal hypotheses between the type of people who are usually selected for laboratory experiments and those who are not" (TESS 2016).

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).

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. Recent reanalysis and replication studies in fields ranging from economics to cancer research have concluded that a large number of findings do not replicate (Begley and Ellis 2012; Chang and Li 2015; Lazer et al. 2014). Recommendations for addressing the replication crisis involve increasing sample sizes, sharing data and study materials, and performing independent verification (Begley and Ellis 2012; Ioannidis 2005; Pasher and Wagenmakers 2012).

The replication crisis has put all social science disciplines under scrutiny. But, 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 thus make these practices

more feasible, increasing the possibility that they will become standard in the field. However, the field of online experimentation is not yet capable of supporting these new standards.

The first function of replication is as a norm establishing the boundaries of scientific inquiry (Radder 1996; Schmidt 2009). Claims that are not replicable are generally regarded as unscientific. The other major function of replication is to establish stability in knowledge (Radder 1996; Schmidt 2009). Researchers seek to verify findings by fully or partially repeating the procedure that initially generated them. Replication serves as a control for chance results, lack of internal validity, or fraud. It can also demonstrate that results apply to a different or larger population. Conceptual replication can help validate hypotheses proposed by the initial research and corroborate its underlying theoretical framework.

At present, most online experiments are not easily replicable because they remain expensive to create and difficult to restage. Online experiments still require a great deal of technical expertise to create and significant investments in subject recruitment and management. And, existing experiments are typically created with customized code run in special-purpose computing environments and performed with a single-use sample specially recruited for that study. This makes them difficult to transfer to other researchers for independent replication. The present decentralized, ad hoc approach to infrastructure 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. We then 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 a platform for developing online experiments and a publicly accessible website for participating in research.[1] The experiment development platform provides researchers code and tools to reduce the costs of experiment development and enable code sharing for quick and faithful study replication. Volunteer Science is also a website where researchers can host their studies and recruit users to participate from anywhere in the world. The advantage of such a website is the ability to collectivize the recruitment process: recruiting subjects for one study makes those subjects available for other studies.

Volunteer Science is unique in combining experiment development tools with a pool of online volunteers. Current facilities for online research only provide one of these. Crowdwork platforms like Amazon's Mechanical Turk and Crowdflower offer access to pools of diverse, flexible labor. Programs like TESS provide access to a nationally representative panel. However, these pools do not come with their own tools for creating studies.

Conversely, researchers have built toolkits for creating online experiments of different kinds. Vecon Lab (Holt 2005) and Z-tree (Fischbacher 2007) offer a variety of economic experiments while Breadboard (McKnight and Christakis 2016) and Turkserver (Mao et al. 2012) offer support for studies of social networks. However, researchers must deploy these systems on their own and recruit their own users. Volunteer Science offers

[1]The technical details of the system will be published as a whitepaper on our website.

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) to reduce the costs of development. There are currently more than 20 experiment templates (including the studies reported in this article) that researchers can use to build their own experiments. Researchers can also use the API to create basic functionality like managing subject consent, subject randomization, and real-time communication between subjects. The API also enables new features to be developed and then made accessible to other researchers across the platform. By providing starter experiments and an API, Volunteer Science can significantly reduce the time, technical expertise, and cost associated with creating online experiments.

Volunteer Science was designed to be a stable environment with open data policies that support study verification and replication. As a shared platform, Volunteer Science standardizes the environment, meaning a study can be shared, reimplemented, and restaged 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 who can participate 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 1,000 users per hour and 50 to 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. The vast majority of research involving the voluntary participation of non-scientists, called "citizen science," require subjects to invest substantial amounts of time and energy to participate (Sauermann and Franzoni 2015). However, projects like reCAPTCHA, in which individuals transcribe images to confirm they are not robots, demonstrate the power of harnessing a small piece of the massive amounts of activity individuals do every day (Von Ahn et al. 2008). In our case, the activity we harness is 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. We designed our experiments to be intrinsically rewarding and aesthetically pleasing.

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 to a particular study afterward. For example, a volunteer can fill out a personality survey. After finishing, we ask them in a pop-up window whether or not they want to donate that data to a particular study such as the one reported here. If they decide to donate their data, volunteers digitally sign the data donation consent form.

Volunteers are never participating blindly. Volunteers are still provided with information about the experiment before they start. And because volunteers have already generated the data when they consent to donating it, their consent is more robustly informed.

These subject protections come with some tradeoffs for researchers, however. Researchers can collect data from their research instruments but cannot use the data until volunteers have donated it to their study. In addition, we maintain a restriction against the use of deception because deception can erode the faith of the volunteer community and can be undermined by off-site discussions that are difficult to monitor.[2]

Finally, Volunteer Science does not facilitate financial transactions. However, participants can be compensated in three ways. First, researchers 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.

These rules and procedures around recruitment, consent, deception, and compensation were formulated to maximize public access to and participation in science while giving researchers as much flexibility as possible and still providing human subject protections to those who want to be research subjects. We designed Volunteer Science to be accessible to anyone, to be fun and engaging, and to empower volunteers with a data donation model of participation. Since deploying our first replication experiment in 2014, we recruited 27,333 volunteers from over a hundred countries to participate in 54,795 experiment sessions across these six studies.

[2]With regard to research involving robots, we ask researchers to avoid the use of bots where possible. But, if researchers do use bots, we require them to explicitly disclose the possibility that subjects could interact with bots or human players for that study or debrief participants on the nature of other players (bots vs. humans) prior to asking subjects to consent to donate data.

VALIDATION METHODOLOGY

We conducted a series of studies to test if Volunteer Science is capable of delivering on the promise of recruiting large numbers of diverse volunteers and producing valid experimental research. For this research, we selected a wide range of studies capable of eliciting patterns of behavior critical to social and behavioral research. We recruited tens of thousands of volunteers through a wide variety of online sources. The results provide evidence that experiments performed with volunteers on Volunteer Science can effectively and validly evoke patterns of behaviors that are comparable to brick-and-mortar laboratories.

Study Selection

We decided to replicate six studies capturing behaviors critical to different experimental traditions. The first study involves two experiments testing participants' reaction times, facilities that are essential for priming, memory, and implicit association research (Crump et al. 2013). Our second study replicates several experiments involving cognitive heuristics identified in behavioral economics. Replicating these experiments allows us to determine whether volunteers make the common yet counterintuitive decisions indicative of practical judgment (Kahneman 2003).

For our third study, we implement the big five personality survey and attempt to independently validate the five factors model of personality it is designed to capture. This helps determine whether or not

researchers are able to use survey-based, multidimensional inventories with volunteers on Volunteer Science. Fourth, we implement two studies that test two social forces: social influence (Nemeth 1986) and justice (Kay and Jost 2003). The question is to what extent online laboratories can deliver social information. The fifth type of behavior we test common to group experiments is problem solving, specifically the traveling salesperson problem. The final type of behavior we test is whether subjects respond to changes in incentives. We created experiments following the prisoner's dilemma, commons dilemma, and public goods paradigms. Individuals must decide whether to cooperate or defect from the collective good in exchange for systematically varying payoffs.

Each of these studies was created as a game or survey on Volunteer Science's public page (volunteerscience.com/experiments). 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.

Subject Recruitment

The advantage in creating an online pool of volunteers is the potential for a large-scale sample of free participants. We use a variety of outlets to reach volunteers, and the game-based design of Volunteer Science lends itself to high-reach, low-yield recruitment strategies that we pursued online and in traditional advertising. As a result of this strategy, we have been able to generally sustain growth, having run 68,402 experimental sessions in 2014 and 2015.

We use social media and online and offline advertising to recruit volunteers. We maintain a pipeline of volunteer participants through online and social media advertising, including Facebook advertising and links to experiments on the website Reddit.com. We also use social media to contact, build, and mobilize an online community of volunteer participants and researchers. This strategy allows us to maintain a large volunteer pool and share research results with participants.

Finally, we created the capacity for students to participate in these experiments and get credit for class. This enables us to recreate one of the primary modes of subject recruitment for offline social science laboratories: participation in experiments for class credit. To validate users' participation, we created a certification system that allows individuals to generate a PDF certificate that summarizes the number of experiments and time spent participating on Volunteer Science. Each certificate contains a hyperlink that faculty can visit to verify the information. Since August 2014, users have created 481 certificates.

Participants

The quality of our citizen science model for recruiting an engaged sample can be assessed by looking at volunteer participation. Demographic information for gender and age is taken from users who created accounts, while data on users' language and device type are collected from users' web browsers. For users who do not create accounts, we use cookies to track their participation across studies.

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 just over half the time.

For those who signed in with Facebook, we found no difference in the probability

of consenting by age ($t = -.52$, $p = .60$) or gender (77 percent of males donated vs. 75 percent of females, chi-square = .89, $p = .35$). We did find significant differences in those using English language browsers and those using other languages (44 vs. 58 percent, respectively, chi-square = 188.0, $p < .001$), and those only using desktop computers (47 percent) versus those using mobile devices (43 percent, chi-square = 18.5776, $p < .001$) are more likely to donate their data.

Those who consented were more likely to participate in multiple experiments than those who never consented (2.6 vs. 1.6 experiments, respectively, $t = -25.5$, $p < .001$). There were no differences in participation by gender ($t = -1.38$, $p = .17$) or age ($t = 1.06$, $p = .29$). However, volunteers 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 ($t = 4.01$, $p < .001$).

Finally, as will be reported individually for each experiment, users who do participate contribute high-quality data. The proportion of incomplete participation varies widely and reaches as high as 25 percent on some surveys. However, the average rate of noncompletion for consenting subjects is 8 percent. In the tests where we are able to assess participant quality, we find that those who complete the study and donate their data provide usable data in 99 percent of cases. Thus, while only 56 percent of all experiments or measures are donated, of those, 91 percent are complete and valid.

RESULTS

Study 1: Reaction Times

First, we replicate two reaction-time-based studies that elicit the Stroop and flanker effects (Eriksen 1995; MacLeod 1991). Measures of human reaction time are essential to a range of social psychological studies, including measures of implicit association, working memory, and perception. However, there is a question of whether delays in computational processing and communication as well as subjects' attention span will allow for an online experiment to detect the small reaction time differences. The advantage of using these two tests is that they differ in time sensitivity. In traditional laboratory studies, the Stroop effect produces a 100- to 200-millisecond delay in reaction while the flanker effect produces a 50- to 60-millisecond delay (Crump et al. 2013). By replicating both, we test how precisely the Volunteer Science system can validly measure reaction time.

The Stroop experiment. We implement the Stroop experiment according to Crump et al. (2013). The Stroop experiment tests the effect of cognitive interference generated by incongruent contextual information. 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. An example of an incongruent prompt is the word *yellow* displayed in the color red (MacLeod 1991). The hypothesis is that subjects will show a significant delay in identifying the target information in the incongruent condition (i.e., *yellow* displayed in red).

At the time of writing, 1,310 unique individuals had participated in 1,674 sessions of the Stroop experiment. Of these, 1,333 sessions were donated (80 percent), and 1,306 (98 percent) of those donated were complete. In total, 286 sessions were excluded because they were not the subject's first session, although including their data did not affect the final results. Following Crump et al. (2013), we excluded users who got less than 65 percent of the items correct in Stroop (2.0 percent). As a result, the total number of sessions in the final analysis is 970.

For Stroop, the mean response time is 951.3 milliseconds for congruent and 1,141.4 milliseconds for incongruent stimuli ($t = -29.41$ $p < .001$). 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, the mean response times are slightly higher than found in traditional laboratory settings. For example, Logan and Zbrodoff (1998) report a mean of 809 milliseconds for congruent stimuli and 1,023 milliseconds for incongruent stimuli. The 150-millisecond difference is roughly a tenth of a second and may be accounted for by delays induced by technology.

The flanker experiment. The flanker experiment tests the same type of effect as the Stroop experiment. In the flanker test, 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). Like the Stroop experiment, the hypothesis is that subjects will show a significant delay in identifying the target information in the incongruent condition (i.e., *hhfhh* and *ffhff*).

At the time of writing, 1,310 unique individuals participated in 1,721 sessions of the flanker experiment. Of these, 1,458 sessions were donated (85 percent). Of donated experiments, 1,433 (98 percent) were complete. In addition, 342 sessions were excluded because they were not the subject's first session, although including their data did not affect the final results. Finally, we excluded 28 sessions where users got less than 65 percent of the items correct (1.9 percent). As a result, 1,049 experiment sessions were included in the final analysis.

For flanker, the mean response time is 689.6 milliseconds for congruent and 752.7 milliseconds for incongruent stimuli ($t = -10.13$, $p < .001$). These were also slower than reported in physical laboratories. Wendt and Kiesel (2011) found mean response times of 604 milliseconds and 647 milliseconds for congruent and incongruent stimuli, respectively. 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 10 percent.

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 analysis. Our purpose is to examine whether or not volunteers make counterintuitive decisions indicative of practical judgment.

The disease problem experiment. First, we implemented Tversky and Kahneman's (1981) disease problem. This experiment involves asking subjects to make one of two choices: first, in the positive frame, subjects choose to save 200 out of 600 people or to have a one-third probability of saving 600 people. In the negative frame, subjects choose to let 400 out of 600 people die or to have a one-third probability that no one will die. Prior research shows subjects choose certainty in the positive frame (saving 200) condition and will take risks in the negative frame (one-third probability no one will die).

In total, 688 experimental sessions were completed, 535 (78 percent) were

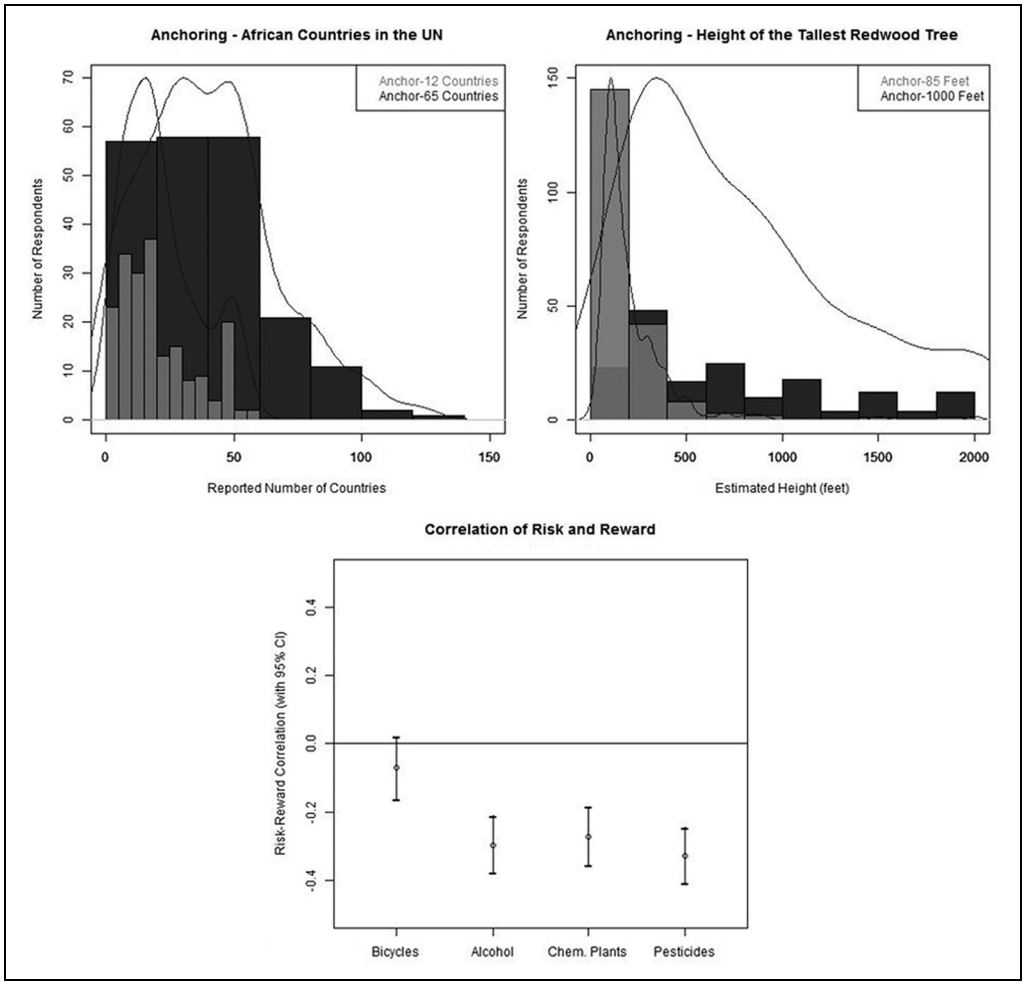

Figure 1. Cognitive Bias Study Results

donated, and 455 (85 percent) were complete, valid, and the participant's first experiment. The results are shown in Figure 1. In the positive, "lives saved" condition, individuals chose the certain outcome of saving 200 lives 60 percent of the time. The opposite occurred, as expected, in the "deaths prevented" condition as 61 percent chose the probabilistic outcome (one-third chance of no one dying) (odds = 2.28, $p < .001$ in Fisher's exact test). These results reflect the direction of Tversky and Kahneman's (1981) finding but represent a weaker difference. Tversky and Kahneman found that subjects in the "lives saved" condition chose the certain outcome 72 percent of the time and chose the probabilistic outcome in the "deaths prevented" condition 78 percent of the time.

Anchoring experiments. We implemented two anchoring effect experiments used by Stanovich and West (2008) that involve asking individuals to estimate a quantity after asking them whether a small or large quantity is the answer. In the first version, we ask "How many African countries are in the United Nations?" and ask whether the answer

is 12 countries (small prompt) or 80 countries (large prompt). The second version asks "How tall is the tallest redwood tree in feet?" The small anchor suggests "85 feet" while the large anchor suggests "1,000 feet." Individuals are randomly assigned to either the small or large anchor and then asked to estimate a response value to the initial question. In both cases, the anchoring hypothesis would predict that participants will give smaller estimates following a small anchor and larger estimates following a large anchor.

At the time of writing, 733 experiments using the African frame were taken, and 689 of the redwood version were taken. Five hundred forty-three Africa experiments (74 percent) and 519 redwood experiments (75 percent) were donated, and of those, 424 (78 percent) and 390 (75 percent) were complete and valid and the participant's first experiment. The results are shown in Figure 1.

For the African countries anchor, the mean estimates in the small and large prompts (12 and 80) were 22 and 41 countries, respectively, $F(1, 178) = 71.0$, Mean Square Error (MSE) = 37,053, $p < .001$. 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 = 3,4307,016, $p < .001$. These generally align with Stanovich and West's (2008) results, which were 14.9 and 42.6 countries and 127 and 989 feet.

Timed risk-reward experiment. Finally, we examined the tendency for individuals to spuriously associate risk and reward. 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).

In all, 1,076 experimental sessions were completed, 808 sessions were donated (75 percent), and 457 were complete, valid, and the participant's first experiment (57 percent). The results are shown in Figure 1. The coefficient for each item was negative and statistically significant except for bicycles, which has repeatedly been shown to not demonstrate the negative correlation (Finucane et al. 2000; Stanovich and West 2008). Comparing our results to Finucane et al. (2000), the correlation coefficients were −.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 scales central to studies of personality, motivation, and culture. These scales can be difficult to create and test because they often need large numbers of subjects to generate reliable estimates of the dimensions being measured. For this study, we attempted to independently validate the 44-question version of the five-factor model of personality. The five-factor model was chosen because it has proven to be robust over a number of samples drawn from diverse populations (McCrae and Terracciano 2005; Schmitt et al. 2007).

At the time of analysis, the survey had been taken 852 times and donated 701 times (82.3 percent). Fifty-four users had taken the survey more than once (7.7 percent), 77 users had missing data (11 percent), and 40 people either entered the same response for every question or finished the survey in under a minute (5.7 percent). No responses could be considered illogical. Of these, 584 surveys

were complete, valid, and the participant's first completion.

We used Cronbach's alpha to assess the consistency of users' responses and exploratory factor analysis to determine the extent to which we could produce a high-quality replication (Lang et al. 2011). The internal consistency of items meant to measure each factor is acceptable in all cases. The Cronbach's alpha is .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 first factor explains 10 percent of the variance, the next three 8 percent each, and the last 7 percent. The result replicates a big five structure, with high positive loadings for almost all items on the corresponding factor and no strong cross-loading patterns. Only two items failed to load strongly on the expected factor: routine (openness) and unartistic (openness).

Study 4: Justice and Group Influence

Complementary justice. Our fourth study looks to induce two essential forces studied by social psychologists: social influence and individuals' sense of justice. We implemented a replication of one of the four studies from Kay and Jost (2003) to investigate whether Volunteer Science could activate participants' sense of justice and whether those priming effects would be detectable through implicit and explicit measures.

In their study, Kay and Jost (2003) present students with a vignette about two friends, Joseph and Mitchell, one of whom eventually becomes wealthy and the other poor. The justice prime comes from connecting wealth and happiness. In the noncomplementary version, Joseph "has it all" while Mitchell becomes "that broke, miserable guy." In the complementary version, Joseph is "rich but miserable" and Mitchell is "broke but happy." Kay and Jost found that subjects who were exposed to the noncomplementary scenario (i.e., "has it all" and "broke, miserable guy") responded more readily to justice-related words in a lexical decision task and had higher scores on a system justification inventory.

We implemented the vignette, lexical task, the Protestant Work Ethic (PWE) scale, and system justification (SJ) inventory described by Kay and Jost (2003). Subjects were randomly assigned to either the complementary or noncomplementary vignettes and then continued to participate in the subsequent three tasks. At the time of writing, individuals had started the vignette 1,691 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 perform the same 2×2 ANOVA predicting SJ scores using the interaction of the experimental condition (complementary and noncomplementary condition) with a dummy for people who scored above or below the median on the PWE scale. 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 the system justification score directly, $F(1, 133) = .37$, MSE = .291, $p = .54$, or in interaction with the PWE, $F(1, 113) = 1.81$, MSE = 1.81, $p = .131$. There was also no effect on participants' (logged) reaction time for justice-related words, $F(1, 133) = .02$, MSE = .008, $p = .89$, indicating that our vignette failed to prime participants' sense of justice.

Group influence experiment. We also implemented a replication of Nemeth's

(1986) group influence study to investigate whether subjects would respond to simulated social influence online. In the original study, individuals are placed in a group of six with either two or four confederates and two or four subjects; they were then 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) who give correct or incorrect responses. The result is 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 the five confederates. In the minority condition, we choose two simulated confederates who give right or wrong answers while having the remaining simulated subjects give the easy, correct answer only. In the majority condition, the only difference is that we have four simulated confederates who give the same right or wrong answers.

At the time of writing, 1,188 influence studies had been taken, 866 (73 percent) had been donated, and 515 experiments (80 percent) were complete, valid, and the participant's first experiment. As a test of validity, we found that participants exposed to correct answers, whether or not they were in the majority or minority condition, were more likely to include those answers in the second round than subjects who did not see the correct answers, $F(1, 384) = 9.59$, MSE = 3.02, $p < .01$.

Contrary to the original result, individuals in the majority condition were no more likely to converge to the majority opinion than those in the minority condition converged to the minority opinion, irrespective of whether the minority or majority were right or wrong, $F(1, 384) = .64$, MSE = .09, $p = .42$. Additionally, there was no evidence that subjects in the minority condition found more unique, correct solutions than subjects in the majority condition, $F(1, 201) = .57$, MSE = .08, $p = .45$.

Study 5: Problem Solving

The fifth test involved using a classic puzzle originally created in computer science but increasingly used to test human cooperation in groups: the traveling salesperson (TSP). 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. The TSP is one of several problems in computer science in which humans traditionally perform much more efficiently than computers. As such, it is a commonly studied problem in which we know how humans behave (MacGregor and Chu 2011; Shore, Bernstein, and Lazer 2015).

In our implementation of the traveling salesperson problem, we provide users with a two-dimensional Cartesian plane with 20 points ("cities"). Users are asked to connect the points in a way that minimizes the total distance "traveled between cities." Users are given 10 rounds to try and minimize their distance. Existing research shows that the most difficult maps (those with the highest error) are those with more cities inside the interior convex hull of the cities (MacGregor and Chu 2011). The more clustered cities are in the middle of the space, the more difficult we expect it to be for users to minimize the total distance.

In total, 6,280 subjects had participated in 7,366 sessions that attempted to solve maps with between 9 and 15 cities inside the interior hull. Of these, 3,651 (45 percent) were donated. We excluded 142

Table 1. Payoff Matrices for Social Dilemmas

Prisoner's Dilemma Payoffs					
Condition	Prediction	All Testify	Ratted Out	Rat Out	None Testify
1	Not testify	3 years	5 years	0 years	1 year
2	Testify	3 years	10 years	0 years	3 years
Commons Payoffs					
Condition	Prediction	Barn Feed	One Commons	Two Commons	All Commons
1	Barn	.75 points	1 point	0 points	−1 points
2	Lean barn	.25 points	1 point	0 points	−1 points
3	Lean commons	.25 points	3 points	0 points	−1 points
4	Commons	.25 points	3 points	0 points	0 points

participants (3.4 percent) who participated via a mobile device because they were given smaller maps. We also excluded 323 (8.8 percent) incomplete cases.

All players played the same sequence of maps, and since the vast majority of players participated in only one round, there are many more data for some maps than others. To accurately measure our effects in an imbalanced sample, we bootstrapped the estimate of the correlation, drawing random samples of data and calculating the Pearson coefficient for each sample. We then performed a one-sample t test on the average of these coefficients to determine whether or not the average effect was different from zero. The results show that the average estimated correlation coefficient is −.09 ($p < .001$), meaning as the number of cities inside the convex hull increases, the number of edges guessed correctly decreases. These results conform to existing research showing that TSP problems with more cities inside the interior hull are more difficult for subjects to solve (MacGregor and Chu 2011).

Study 6: Social Dilemmas

For our sixth study, we implemented three canonical social dilemmas: the prisoner's dilemma, a commons dilemma, and the public goods dilemma. 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 of research in this tradition is that participants are sensitive to incentives. 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 within each experiment if we randomly assigned them to different payoff systems.

Prisoner's dilemma experiment. The prisoner's dilemma (PD) involves choosing to cooperate or defect from a partner. Subjects are rewarded based on the combination of their choice and the choice of other players. For example, in Condition 1 in Table 1, if the subject cooperates and the other player defects ("ratted out"), Player 1 receives five years in prison. Table 1 outlines the three payoff conditions. The ideal strategy for whether or not to cooperate depends on the payoffs offered, and in all games the Pure Strategy Nash Equilibrium (PSNE) corresponds to

both participants testifying. By keeping the PSNE constant across conditions and only varying the scale of the payoff, we are able to isolate the effect of changing the quantity of the payoff on subject behavior.

At the time of analysis, 663 unique individuals had participated in 825 sessions of the prisoner's dilemma. Of these, 340 PD sessions (41 percent) were donated. Of the 340 donated, 236 sessions (69 percent) were complete and the subject's first experimental session. In each session, subjects were randomly assigned to one of the experimental conditions, and the session ends at the completion of that one condition. That is, subjects did not participate in all conditions.[3]

The results of a pairwise tests show significant differences in subjects' average choice across Conditions 1 and 2 (t = 2.42, p = .016). Because we are not explicitly replicating a prior study, there is no prior established rate of cooperation or defection against which to compare these results. Instead, what the results show is simply that subjects respond to differing incentives in the expected direction. Thus, volunteer participants on Volunteer Science are responsive to incentives within experiments, even in the absence of monetary reward (Amir, Rand, and Gal 2012).

Commons experiment. The commons dilemma involves choosing to use a common resource or a private resource (a cow pasture or barn in our case). The private resource provides fewer but certain benefits to the player, whereas the common resource provides potentially more but uncertain benefits. In the commons dilemma, there are three players, one human and two randomly playing computer agents ("bots"), and subjects are punished if too many people use the commons simultaneously. Table 1 outlines the conditional payoffs for choosing the private (barn) or public (commons) resource. All four games resolve to a Pure Strategy Nash Equilibrium in which only one player plays "Commons" and the others play "Barn," creating three total PSNEs: {Barn, Barn, Commons}, {Barn, Commons, Barn}, and {Commons, Barn, Barn}.

We collected data from December 12, 2015 through December 31, 2015. In this time, 3,189 unique individuals participated in 4,145 sessions of the commons experiment. Of these, 3,008 sessions were donated (68 percent). Of those donated, 1,786 (59 percent) were complete and the subject's first session.

Pairwise tests of subjects' average choice between neighboring conditions 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, $p < .001$), and Conditions 3 and 4 (t = 2.24, p = .025). As in the PD experiment, there is no earlier experiment to compare the rates of private and commons use against. Instead, this is another demonstration that volunteers respond to incentives in the expected (i.e., monotonic) way.

Public goods investment experiment. The final experiment we ran was based on the public goods paradigm. There is a long tradition of using experiments to gain insight into what is known as the public goods problem (Van Laerhoven and Ostrom, 2007). The public goods dilemma (PG) asks individuals to either cooperate or defect in collective dilemmas. Following Suri and Watts (2011), we create an economic version where users must decide how much money to contribute to a "group investment program."

The game begins with an animated video that details how the game works

[3]Following the feedback from one reviewer, we eliminated one experimental condition because the condition altered the Pure Strategy Nash Equilibrium. This reduced our sample from 236 to 167 individuals. Thus, the number of cases in each cell is 81 for Condition 1 and 83 for Condition 2.

and the pros and cons of cooperation and defection. The more money put into the pot, the more money the pot will have at the end of the round. After viewing the video, the game begins with a user and four bots. The user has a choice to invest a minimum of \$0 and a maximum of \$10. When a round concludes, the whole group splits the pot evenly regardless of an individual's contribution level. Therefore, subjects are incentivized to maximize the total amount of money contributed while minimizing their individual contribution.

For the PG experiment, a total of 532 subjects participated. After removing 40 individuals that did not complete all three rounds (7.5 percent) and 26 individuals who did not consent (4.9 percent), the final sample size is 466. One of the issues for the current experiment is that subjects will not respond to contributing or free-riding in the same way than if the experiment was done offline, where they may even receive financial contributions based on how well they participated in the experiment (e.g., Andreoni and Petrie 2004). By comparing the contribution distribution to offline PGGs, we can get an initial glimpse into whether or not the online PGG "maps" with offline PGGs

To compare distributions, we looked at (1) overall average contributions and (2) distribution of "free-riders" and "contributors" in the first round. Subjects typically contribute 40 to 60 percent in the first round. This decreases with each round but remains above zero (Ostrom 2000). In the current experiment, volunteers donated 46.5 percent of their endowment in the initial round. In subsequent rounds, subjects differed in the amount they contributed, $F(1.92, 927.64) = 6.71$, $p < .01$, $\eta^2 = .014$, contributing less ($t = 2.28$, $p = .02$) in the final round (M = 4.21) than they did in the first (M = 4.65) but still remained above zero, reflective of Ostrom's (2000) summary of offline PGGs.

For the second point, we follow Gunnthorsdottir, Houser, and McCabe's (2007:308) classification of a free-rider as "someone who contributes 30 percent or less of his endowment in the first round" and the remaining part of the cohort as contributors. Using a similar return percentage from the collective donation (Gunnthorsdottir et al. used 50 percent while we used 40 percent) and similar random contributions from other participants, Gunnthorsdottir and colleagues found a distribution of 30 percent (n = 18) free-riders and 70 percent (n = 42) contributors. Our distribution was very similar, with 32.6 percent free-riders and 67.4 percent contributors. As such, we find some support that individuals played the PGG online in a similar fashion as they would have played it offline.

DISCUSSION

On the whole, the findings from each of these experiments supports 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 thousands of experiment results. We are able to induce and measure very low-latency reaction times. Using questionnaires, we can validate multidimensional inventories and elicit behaviorally realistic responses to tests of cognitive bias. And, our participants engage in economic tradeoffs and puzzle solving in ways found in a variety of other types of 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. And, the direction of the results held but were not statistically significant. This suggests 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. This leads us to believe our vignette did not elicit the priming effect found by Kay and Jost (2003). Thus, both of these results point to the need to create stronger social signaling to activate the justice primes or the sense of peer pressure in online settings.

While this and a range of other studies have demonstrated that a variety of research can be performed online, there are several shortcomings to our approach. First, some experiments are much more popular than others. On Volunteer Science, game-based experiments attract much more attention 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.

Further, online platforms are inherently limited to the types of studies that can be implemented, executed online, and administered with standard computing equipment. Although a great number of studies can be meaningfully implemented online, 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. Finally, much of the work we have done with Volunteer Science to date either relies on single-person experiments or the use of computer agents (bots) in multiperson 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. As people continue to use Internet-based technologies in their daily lives, social science will benefit from collecting these data and integrating them into our research (Lazer et al. 2009). 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 that 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 that 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 that we are currently working to solve.

We introduce Volunteer Science as an online laboratory that 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 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.

ACKNOWLEDGMENTS

The authors would like to thank the editors of the special issue and anonymous reviewers for their incisive feedback. We would also like to thank the many collaborators who helped develop Volunteer Science, including Alan Mislove, Burak Aslan, Ceyhun Karbeyaz, Christo Wilson, Christoph Riedl, Dan Calacci, Kyla Ryan, Luke Horgan, Mikal Khoso Hussein, Nisha Shah, Qaish Kanchwala, and Skyler Place.

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

Amir, Ofra, David G. Rand, and Ya'akov Kobi Gal. 2012. "Economic Games on the Internet: The Effect of $1 Stakes." *PLoS ONE* 7(2):e31461.

Andreoni, James, and Ragan Petrie 2004. "Public Goods Experiments without Confidentiality: A Glimpse into Fund-raising." *Journal of Public Economics* 88(7):605–1623.

Bechara, Antoine, and Antonio R. Damasio. 2005. "The Somatic Marker Hypothesis: A Neural Theory of Economic Decision." *Games and Economic Behavior* 52(2):336–72.

Begley, C. Glenn, and Lee M. Ellis. 2012. "Drug Development: Raise Standards for Preclinical Cancer Research." *Nature* 483(7391):531–33.

Chang, Andrew C., and Phillip Li. 2015. "Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say 'Usually Not.'" *Finance and Economics Discussion Series* 2015(83):1–26.

Christian, Carol, Chris Lintott, Arfon Smith, Lucy Fortson, and Steven Bamford. 2012. "Citizen Science: Contributions to Astronomy Research." Retrieved October 31, 2013 (http://arxiv.org/abs/1202.2577).

Cook, Karen S., and Eric Rice. 2006. "Social Exchange Theory." Pp. 53–76 in *Handbook of Social Psychology, Handbooks of Sociology and Social Research*, edited by J. Delamater. New York: Springer.

Cook, Karen S., Toshio Yamagishi, Coye Cheshire, Robin Cooper, Masafumi Matsuda, and Rie Mashima. 2005. "Trust Building via Risk Taking: A Cross-Societal Experiment." *Social Psychology Quarterly* 68(2):121–42.

Crump, Matthew J. C., 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.

Davison, Robert B., John R. Hollenbeck, Christopher M. Barnes, Dustin J. Sleesman, and Daniel R. Ilgen. 2012. "Coordinated Action in Multiteam Systems." *Journal of Applied Psychology* 97(4):808–24.

Eriksen, Charles W. 1995. "The Flankers Task and Response Competition: A Useful Tool for Investigating a Variety of Cognitive Problems." *Visual Cognition* 2(2–3):101–18.

Finucane, Melissa L., Ali Alhakami, Paul Slovic, and Stephen M. Johnson. 2000. "The Affect Heuristic in Judgments of Risks and Benefits." *Journal of Behavioral Decision Making* 13(1):1–17.

Fischbacher, Urs. 2007. "Z-Tree: Zurich Toolbox for Ready-Made Economic Experiments." *Experimental Economics* 10(2):171–78.

Gilbert, Daniel, Gary King, Stephen Pettigrew, and Timothy Wilson. 2016.

"Comment on 'Estimating the Reproducibility of Psychological Science.'" *Science* 351(6277):1037a–1037b.

Gosling, Samuel D., Carson J. Sandy, Oliver P. John, and Jeff Potter. 2010. "Wired but Not WEIRD: The Promise of the Internet in Reaching More Diverse Samples." *Behavioral and Brain Sciences* 33(2–3):94–95.

Gunnthorsdottir, Aanna, Daniel Houser, and Kevin McCabe. 2007. "Disposition, History and Contributions in Public Goods Experiments." *Journal of Economic Behavior & Organization* 62(2):304–15.

Hackman, J. Richard, and Nancy Katz. 2010. "Group Behavior and Performance." Pp. 1208–51 in *Handbook of Social Psychology*, edited by S. Fiske, D. Gilbert, and G. Lindzey. New York: Wiley.

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. "The Weirdest People in the World?" *Behavioral and Brain Sciences* 33(2–3):61–83.

Holt, Charles. 2005. *Vecon Lab*. Retrieved August 16, 2016 (http://veconlab.econ.virginia.edu/guide.php).

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLoS Med* 2(8):e124.

Kahneman, Daniel. 2003. "A Perspective on Judgment and Choice: Mapping Bounded Rationality." *American Psychologist* 58(9):697–720.

Kay, Aaron C., and John T. Jost. 2003. "Complementary Justice: Effects of 'Poor but Happy' and 'Poor but Honest' Stereotype Exemplars on System Justification and Implicit Activation of the Justice Motive." *Journal of Personality and Social Psychology* 85(5):823–37.

Kearns, Michael. 2012. "Experiments in Social Computation." *Communications of the ACM* 55(10): 56–67.

Kuwabara, Ko, Robb Willer, Michael W. Macy, Rie Mashima, Shigeru, Terai, and Toshio Yamagishi. 2007. "Culture, Identity, and Structure in Social Exchange: A Web-Based Trust Experiment in the United States and Japan." *Social Psychology Quarterly* 70(4):461–79.

Lang, Frieder R., Dennis John, Oliver Lüdtke, Jürgen Schupp, and Gert G. Wagner. 2011. "Short Assessment of the Big Five: Robust across Survey Methods Except Telephone Interviewing." *Behavior Research Methods* 43(2):548–67.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. "The Parable of Google Flu: Traps in Big Data Analysis." *Science* 343(6176):1203–205.

Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. 2009. "Computational Social Science." *Science* 323(5915):721–23.

Logan, Gordon D., and N. Jane Zbrodoff. 1998. "Stroop-Type Interference: Congruity Effects in Color Naming with Typewritten Responses." *Journal of Experimental Psychology: Human Perception and Performance* 24(3):978–92.

MacLeod, Colin M. 1991. "Half a Century of Research on the Stroop Effect: An Integrative Review." *Psychological Bulletin* 109(2):163–203.

Mao, Andrew, Yiling Chen, Krzysztof Z. Gajos, David C. Parkes, Ariel D. Procaccia, and Haoqi Zhang. 2012. "Turkserver: Enabling Synchronous and Longitudinal Online Experiments." Retrieved October 21, 2016 (http://www.eecs.harvard.edu/~kgajos/papers/2012/mao12-turkserver.pdf).

MacGregor, James N., and Yun Chu. 2011. "Human Performance on the Traveling Salesman and Related Problems: A Review." *The Journal of Problem Solving* 3(2):1–29.

Mason, Winter and Siddharth Suri. 2011. "Conducting Behavioral Research on Amazon's Mechanical Turk." *Behavior Research Methods* 44(1):1–23.

McCrae, Robert R., and Antonio Terracciano. 2005. "Universal Features of Personality Traits from the Observer's Perspective: Data from 50 Cultures." *Journal of Personality and Social Psychology* 88(3):547–61.

McKnight, Mark E., and Nicholas A. Christakis. *Breadboard: Software for Online Social Experiments*. Vers. 2. Cambridge, MA: Yale University.

Nemeth, Charlan J. 1986. "Differential Contributions of Majority and Minority Influence." *Psychological Review* 93(1):23.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349(6251):aac4716–aac4716.

Ostrom, Elinor. 2000." Collective Action and the Evolution of Social Norms." *The Journal of Economic Perspectives* 14(3): 137–58.

Pashler, H., and E. J. Wagenmakers. 2012. "Editors' Introduction to the Special Section on Replicability in Psychological Science: A

Crisis of Confidence?" *Perspectives on Psychological Science* 7(6):528–30.

Radder, Hans. 1996. *In and about the World: Philosophical Studies of Science and Technology*. New York, NY: SUNY Press.

Raddick, M. Jordan, 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.

Radford, Jason, Andy Pilny, Ashley Reichelman, Brian Keegan, Brooke Foucault Welles, Jefferson Hoye, Katherine Ognyanova, Waleed Meleis, David Lazer. 2016. "Volunteer Science Validation Study." V1. Harvard Dataverse. Retrieved October 21, 2016 (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/MYRDQC).

Rand, David G. 2012. "The Promise of Mechanical Turk: How Online Labor Markets Can Help Theorists Run Behavioral Experiments." *Journal of Theoretical Biology* 299:172–79.

Reips, Ulf-Dietrich. 2000. "The Web Experiment Method: Advantages, Disadvantages, and Solutions." Pp. 89–117 in *Psychological Experiments on the Internet*, edited by M. H. Birnbaum. San Diego: Academic Press.

Salganik, Matthew J., and Duncan J. Watts. 2008. "Leading the Herd Astray: An Experimental Study of Self-Fulfilling Prophecies in an Artificial Cultural Market." *Social Psychology Quarterly* 71(4):338–55.

Sauermann, Henry, and Chiara Franzoni. 2015. "Crowd Science User Contribution Patterns and Their Implications." Retrieved October 21, 2016 (http://www.pnas.org/content/112/3/679.full.pdf).

Schmidt, Stefan. 2009. "Shall We Really Do It Again? The Powerful Concept of Replication Is Neglected in the Social Sciences." *Review of General Psychology* 13(2):90–100.

Schmitt, D. P., J. Allik, R. R. McCrae, and V. Benet-Martinez. 2007. "The Geographic Distribution of Big Five Personality Traits: Patterns and Profiles of Human Self-Description across 56 Nations." *Journal of Cross-Cultural Psychology* 38(2):173–212.

Shore, Jesse, Ethan Bernstein, and David Lazer. 2015. "Facts and Figuring: An Experimental Investigation of Network Structure and Performance in Information and Solution Spaces." *Organization Science* 26(5):1432–46.

Simmons, Alice D., and Lawrence D. Bobo. 2015. "Can Non-Full-Probability Internet Surveys Yield Useful Data? A Comparison with Full-Probability Face-to-Face Surveys in the Domain of Race and Social Inequality Attitudes." *Sociological Methodology* 45(1):357–87.

Stangor, Charles, Laure Lynch, Changming Duan, and Beth Glas. 1992. "Categorization of Individuals on the Basis of Multiple Social Features." *Journal of Personality and Social Psychology* 62(2):207–18.

Stanovich, Keith E., and Richard F. West. 2008. "On the Relative Independence of Thinking Biases and Cognitive Ability." *Journal of Personality and Social Psychology* 94(4):672–95.

Suri, Siddharth, and Duncan J. Watts. 2011. "Cooperation and Contagion in Web-Based, Networked Public Goods Experiments." *PLoS One* 6(3):e16836.

TESS. 2016. "Introducing TESS." Accessed April 26, 2016 (http://www.tessexperiments.org/introduction.html).

Tversky, Amos, and Daniel. Kahneman. 1981. "The Framing of Decisions and the Psychology of Choice." *Science* 211(4481):453–58.

Van Laerhoven, Frank, and Elinor Ostrom. 2007. "Traditions and Trends in the Study of the Commons." *International Journal of the Commons* 1(1):3–28

Von Ahn, Luis, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. 2008. "reCAPTCHA: Human-Based Character Recognition via Web Security Measures." *Science* 321(5895):1465–68.

Webster, Murray, and Jane Sell. 2007. *Laboratory Experiments in the Social Sciences.* Boston, MA: Academic Press.

Weinberg, Jill, Jeremy Freese, and David McElhattan. 2014. "Comparing Data Characteristics and Results of an Online Factorial Survey between a Population-Based and a Crowdsource-Recruited Sample." *Sociological Science* 1:292–310.

Wendt, Mike, and Andrea Kiesel. 2011. "Conflict Adaptation in Time: Foreperiods as Contextual Cues for Attentional Adjustment." *Psychonomic Bulletin & Review* 18(5):910–16.

Willer, David, and Henry A. Walker. 2007. *Building Experiments: Testing Social Theory.* Stanford, CA: Stanford University Press.

BIOS

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.

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.

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.

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

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.

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.