Intrinsic Motivation at Scale: Online Volunteer Laboratories for Social Science Research[1]

Austin Strange[2] Ryan D. Enos[3] Mark Hill[4] Amy Lakeman[5]

This draft: October 24, 2018

[1]Funding for this research and the Harvard Digital Laboratory for the Social Sciences (DLABSS) is provided by the Pershing Square Venture Fund for Research on the Foundations of Human Behavior. DLABSS is a scientific program at the Harvard Institute for Quantitative Social Science (IQSS). We thank Angelo Dagonel, Yodahe Heramo, Zayan Faiyad, Jess Li, Angel Navidad, Juliet Pesner, Brandon Wang and Bryant Yang for their contributions to the Lab. For helpful feedback on earlier drafts, we thank Mark Brandt, Alexander Coppock, Jens Olav Dahlgaard, Gregory Huber, Connor Huff, Gary King, Thomas Leeper, Christopher Lucas, Aseem Mahajan, and Dustin Tingley.

[2]Institute for Quantitative Social Science and Department of Government, Harvard University; strange@g.harvard.edu

[3]Institute for Quantitative Social Science and Department of Government, Harvard University; renos@gov.harvard.edu

[4]Institute for Quantitative Social Science and Department of Government, Harvard University; markhill@g.harvard.edu

[5]Institute for Quantitative Social Science and Department of Government, Harvard University; alakeman@g.harvard.edu

Abstract

Once a fixture of empirical social science, volunteer subjects are now only rarely used in human-subjects research. Yet volunteers are a potentially valuable resource, especially for research conducted online. We argue that online volunteer laboratories, by relying on intrinsically motivated subjects, are able to produce high-quality data and may avoid some of the challenges associated with paid online crowd-sourcing. Furthermore, we argue that volunteer labs are scalable and able to produce large datasets for multiple researchers, while imposing little or no financial burden. Using a range of original tests, we show that volunteer and paid respondents differ in their motivations for participation yet have similar descriptive composition. Volunteer samples are able to replicate classic and contemporary social science findings, and produce high levels of overall response quality relative to paid subjects. Our results suggest that volunteer labs represent a potentially significant untapped source of human-subjects data in the social sciences.

1 Introduction

Social science human-subjects research has undergone a major transformation in the last two decades. Enabled by the Internet, large samples of paid subjects, including nationally-representative surveys, now dominate human-subjects research. In particular, as experimental social science has grown in popularity and influence, the use of low-cost convenience samples such as Amazon's Mechanical Turk (MTurk) has proliferated across political science and other social science disciplines. These platforms have dramatically lowered the costs of obtaining subjects, allowing for large samples to be recruited rapidly and inexpensively (Mullinix et al. 2015). Despite this shift towards low-cost online samples, a crucial aspect of human-subjects research remains unchanged: for their participation, subjects are rewarded with money, course credit, or other forms of extrinsic motivation.

Is material compensation the only way to attract high-quality subjects? In this manuscript we present the case for an alternative resource that allows for the collection of inexpensive, large, and diverse populations online: the volunteer. We introduce a scalable, shared online volunteer laboratory and show that volunteer subjects can generate similar results to those obtained from paid representative and convenience samples. Using the internet to gather sustainable pools of volunteer subjects represents a potential major change in human-subjects research by offering researchers high-quality data at little cost.

We discuss potential pitfalls of large, paid online convenience-samples and why using volunteer subjects may help avoid these. Then, using an extensive battery of tests, including replications of a variety of studies and direct tests of response quality, we demonstrate that volunteers can produce data of similar or better quality compared to paid respondents. Building on these results, we argue that volunteer laboratories can be scaled as sustainable sources of human-subjects data.

2 Volunteers: Quality Data at Scale

We define volunteers as unpaid research subjects motivated to participate in studies without receiving direct material compensation, distinguishing these from non-volunteer subjects who are financially or otherwise materially compensated.[1] While once a fixture of quantitative social science,[2] outside of some small scale surveys and field studies, volunteers now only rarely appear in quantitative research.

We argue that volunteer subjects are, on average, more motivated than paid subjects by intrinsic rewards. Drawing on research from psychology, we define intrinsic motivation as willingness to perform an activity that stems directly from the activity itself, or the "inherent satisfactions" it provides, rather than expectation of reward (Ryan and Deci 2000). Research suggests volunteering is intrinsically rewarding, and even leads to higher levels of happiness (Meier and Stutzer 2008). Moreover, intrinsic and extrinsic motivations can be negatively correlated. Indeed, the "Motivation Crowding Effect" posits that the introduction of financial rewards crowds out intrinsic motivation (Frey and Jegen 2001). As a seminal example, paying for blood tends to reduce peoples' incentive to donate it (Titmuss 1970).

As research has moved toward large web-based samples, higher levels of intrinsic motivation may enable volunteer samples to produce response data of quality as high as, and sometimes higher than, paid online samples. A major reason is simply that volunteers are less likely to induce biases rooted in the desire for material rewards. For example, when subjects are working at an hourly rate (sometimes as a major source of income, in the case of MTurk and other online labor pools (Williamson 2016)), they may have an incentive to rush through studies or provide fictitious information, potentially leading to low-quality

[1] We do not consider students who are participating in research for course credit to be volunteers. On the other hand, we do consider subjects who receive token compensation, for example stickers (Klar and Krupnikov 2016), to be volunteers. If subjects do not know they are participating in research, as in some field experiments, these subjects are not included

[2] See, for example, the discussion of psychology studies in Rosnow and Rosenthal (1997).

data. Moreover, individuals sometimes misrepresent their true identity in order to qualify for MTurk tasks that offer higher returns (Sharpe Wessling, Huber, and Netzer 2017). Relatedly, recent evidence suggests that significant amounts of paid online response data may be produced by automated bots (Dupuis, Meier, and Cuneo 2018). Another common concern about online labor markets is that subjects participate multiple times in the same or related experiments (Stewart et al. 2015). Though paid platforms often use terms-of-use agreements, membership fees, or software requirements to deter this behavior, such solutions create tension with these markets' ability to provide cheap and diverse subjects to researchers. Each of these behaviors raise concerns about data validity. In contrast, because volunteers are more insulated from extrinsic incentives, such concerns are minimal. Instead, there is evidence that intrinsically motivated volunteers contribute substantially to successful online services such as Wikipedia (Schroer and Hertel 2009).

Of course, paid online samples have a tremendous advantage of quickly providing inexpensive subjects to researchers. However, by leveraging similar technology, online volunteer pools are coming to complement these paid platforms. An example of an all-volunteer online laboratory is the Harvard Digital Lab for the Social Sciences (DLABSS),[3] a standing pool of over 13,000 volunteer subjects. As one of the first such online volunteer laboratories, DLABSS is ideal for our analysis because it can produce large volunteer samples needed for comparisons with online paid subject pools.[4] DLABSS is a public good in that it supports diverse substantive interests of a broad pool of researchers. The essential features of the laboratory are a website and an email list and therefore impose minimal costs. Volunteers are primarily recruited using social media and other free sources.

Volunteer laboratories provide many of the benefits of paid samples, but perhaps have additional advantages. Below we note four advantages specific to online volunteer laborato-

[3] http://dlabss.harvard.edu/

[4] There are other examples of volunteer digital labs. For example, a consortium of researchers have developed Volunteer Science (https://volunteerscience.com/). See Radford et al. (2016).

ries. First, in crowdsourced markets, communication between subjects cannot be controlled by the researcher. For example, MTurk workers can digitally interface on MTurk Forum and MTurk Crowd, raising concerns about subject crosstalk biases (Paolacci, Chandler, and Ipeirotis 2010), as well as violations of the Stable Unit Treatment Value Assumption (SUTVA) critical for valid causal inference.[5] In contrast, the subjects in volunteer labs have little incentive to create such platforms.

Second, subject attrition may cause bias in experimental research if subjects drop out of one treatment at higher rates than another. Selective attrition appears to be a concern for inference in MTurk-based studies: Zhou et al. (2016) replicate existing MTurk studies and find that every one had an attrition rate of greater than 20%, sometimes exceeding 50%. In contrast, selective attrition may be less severe in volunteer labs, perhaps due to the intrinsic motivation of subjects: the mean and median attrition rate for DLABSS studies is 22.5% and 20%, respectively (these numbers decrease considerably for surveys under 10 minutes in length; see Appendix Section A).

Third, paid crowd-sourced online research, while less expensive than many alternatives, can still be very costly. Volunteer laboratories impose no monetary costs on researchers other than the costs of hosting a website and recruiting subjects (which can be minimal and shared). For example, for a study we report later in this manuscript, we spent $840 to administer a simple 10-minute experiment on MTurk with 800 respondents. We conducted an identical study on DLABSS with a marginal cost of $0.[6]

Finally, volunteer laboratories can expand access to human-subjects research for both social scientists and volunteers. In addition to the lowered barriers to entry provided by monetary savings, the modular nature of a volunteer laboratory like DLABSS makes it

[5]Each of these fora have tens of thousands of members and thousands of discussions. Indeed, we found postings about our MTurk surveys on three fora.

[6] Relatedly, political scientists have also raised concerns about the ethics of low wages on MTurk resulting from cost-saving pressures felt by researchers (Williamson 2016).

accessible to a range of research programs, mitigating the need for researchers to set up platforms specific to a single project. As social science has moved toward "larger scale, collaborative, interdisciplinary" research (King 2014), this modularity is increasingly valuable. Moreover, from the perspective of volunteers, there are few barriers to entry, meaning individuals throughout the world can join a digital social science community, extending participation in social science to a wider audience.

Of course, volunteer subjects also have potential disadvantages. Though DLABSS has demonstrated the potential for large, readily available samples of volunteers, these subjects may be acutely unrepresentative, raising external validity concerns. For example, volunteers may be more motivated by academic interest than the general public (Rosenthal and Rosnow 1975). To address this possibility, we turn to comparing the demographic properties, representativeness, and relative quality of volunteer subjects.

3 Validity of Volunteer Labs as a Research Tool

To demonstrate the potential of volunteer subjects, we report four sets of findings. First, we compare the properties of volunteer and paid online subjects, including their stated motivations for participating in research as well as laboratory-wide sample properties. We then report replications of classic and contemporary studies using volunteers. Finally, we conduct several tests of relative response quality between volunteer and paid subjects.

3.1 Volunteer and Paid Subject Motivation

To compare subject motivation, we surveyed the motivations of 742 volunteers on DLABSS and 649 MTurk paid subjects. Volunteers and paid respondents were tasked with ranking, from most to least important on a 7 point scale, motivations for their own participation in online surveys. The motivations offered were earning money, learning about current affairs,

being part of an online community, helping researchers, experiencing studies, passing the time, and helping others. Respondents were also asked how likely they would be to participate in online studies even if there was no chance of compensation, and given the opportunity to describe in detail why they choose to participate in online studies.

Volunteers and paid subjects reported significantly differing motivations for participating in online surveys (see Appendix Figure A5). Those in paid subject pools are most motivated by earning money, while volunteers are least motivated by earning money and most motivated by the possibility of helping researchers. Additionally, when asked how likely they would be to participate in such studies even if there was no chance for compensation, only 15.2% of paid subjects said they would be very likely or somewhat likely to do so, while 85.7% of volunteers reported they would be likely to continue. When probed to explain their motivations in depth, 43.5% of paid subjects mentioned the word "money," compared to only 2.6% of volunteers.

Following the definition of intrinsic motivation provided earlier in the paper, of the seven answer choices, learning about current affairs, passing the time, and experiencing online studies are arguably the most unambiguously intrinsically motivated. The DLABSS panel ranked learning about current affairs and experiencing studies higher than the MTurk panel. Motivations that are not objectively intrinsic (but certainly non-material) include helping researchers and being part of an online community, both of which volunteers ranked as more important motivations than paid subjects.

3.2 Volunteer and Paid Subject Characteristics

We now turn to sample-wide analyses of volunteer and paid subjects. Drawing on the influential study of Berinsky, Huber, and Lenz (2012), which is often cited in research using MTurk, in Table 1 we compare the demographics, political behavior, and political knowledge

of DLABSS volunteers with a sample of MTurk paid subjects from December 2016[7] and with the online 2008-09 American National Elections Study Panel Study (ANESP), the 2012 American National Elections Study (ANES), and the 2012 Current Population Survey (CPS).

Generally speaking, DLABSS volunteers appear quite similar to those from both online and offline survey platforms, though important differences do exist.[8] For example, compared to paid MTurk subjects, volunteers appear more representative of the general population in terms of age. There are certain demographic characteristics that one might expect to be correlated with intrinsic motivation, for example leisure time that might be associated with education, income, and race. However, compared to MTurk subjects, volunteers are not substantially more educated, wealthy, or white – although, compared to the nationally representative samples, DLABSS volunteers tend to be more educated and have lower incomes. The volunteer sample expresses a higher level of political interest than the nationally representative samples, but this does not appear to translate into skewed political knowledge: volunteers more closely resemble the nationally representative surveys than do the MTurk subjects, demonstrating the usefulness of volunteers for political science studies, where such variables are central to the study of political behavior (e.g., Zaller (1992)).[9]

3.3 Replicating Paid Studies with Volunteer Subjects

We next report the replication of 16 classic and contemporary social science experiments using volunteer samples. We performed six of the replications ourselves and the rest were replications by other researchers using volunteer samples from DLABSS. The replications

[7] Following Berinsky, Huber, and Lenz (2012), our MTurk advertisement was for a "Survey of Public Affairs and Values."

[8] We report on subsets of the entire DLABSS volunteer population who took specific questionnaires with demographic and knowledge questions necessary for this comparison.

[9] In Tables A1 and A2 in the Appendix, we complete the replication of Berinsky, Huber, and Lenz (2012). Results from volunteers once again approximate other samples in a manner similar to MTurk.

Table 1: Comparing DLABSS sample demographics to internet and face-to-face samples

	Internet sample			*Face-to-face samples*	
	DLABSS	*MTurk*	*ANESP*	*CPS 2012*	*ANES 2012*
Female	57.2% (0.6)	48.0% (1.9)	57.6% (0.9)	51.9% (0.2)	52.0% (0.1)
Education (mean years)	15.1 (0.0)	14.9 (0.1)	16.2 (0.1)	13.4 (0.0)	13.6 (0.1)
Age (mean years)	43.3 (0.2)	37.8 (0.5)	49.7 (0.3)	46.7 (0.1)	47.3 (0.4)
Mean income	$48,203 ($480)	$43,592 ($1,168)	$69,043 ($749)	$61,977 ($138)	$63,199 ($1,274)
Median income	$37,500	$37,500	$67,500	$55,000	$32,500
Race					
White	74.4 (0.5)	78.3 (1.6)	83.0 (0.7)	79.5 (0.1)	74.5 (1.0)
Black	6.3 (0.3)	8.4 (1.0)	8.9 (0.7)	12.2 (0.1)	12.2 (0.7)
Hispanic	7.9 (0.3)	7.7 (1.0)	5.0 (0.4)	15.0 (0.1)	10.9 (0.7)
Marital Status					
Married	42.5 (1.5)	42.7 (1.0)	56.8 (0.9)	53.8 (0.2)	53.2 (1.1)
Housing status					
Own home	51.8 (1.5)	49.5 (1.9)	80.8 (0.8)		71.5 (1.0)
Religion					
None	43.9 (1.5)	40.0 (1.8)	13.1 (0.8)		21.3 (0.9)
Protestant	20.0 (1.2)	25.4 (1.6)	38.7 (1.4)		33.3 (1.1)
Catholic	16.8 (1.1)	20.4 (1.5)	22.9 (1.0)		22.7 (0.9)
Region of the US					
Northeast	22.8 (0.5)	21.5 (1.6)	16.9 (0.7)	18.2 (0.1)	18.2 (0.8)
Midwest	21.0 (0.5)	25.4 (1.7)	28.3 (0.9)	21.6 (0.1)	22.6 (0.9)
South	31.1 (0.6)	38.2 (1.9)	31.4 (0.9)	37.0 (0.2)	37.2 (1.1)
West	25.2 (0.6)	14.9 (1.4)	23.4 (0.8)	23.2 (0.1)	22.1 (0.9)
Party Identification					
Democrat	47.9 (0.6)	44.3 (1.9)			
Independent/Other	30.3 (0.5)	30.1 (1.7)			
Republican	21.8 (0.5)	22.8 (1.6)			
Ideology					
Liberal	59.9 (0.5)	62.6 (1.9)			
Conservative	32.9 (0.5)	37.4 (1.9)			
Registration/turnout					
Registered	89.3 (0.9)	91.6 (1.0)	92.0 (0.7)	71.2 (0.1)	72.8 (1.0)
Voted in 2008	80.1 (1.4)	73.8 (1.7)	89.8 (0.5)	61.8* (0.2)	70.2* (1.0)
Political Interest	3.84 (0.03)	3.62 (0.04)	2.71 (0.02)		3.34 (0.03)
Political Knowledge (% correct)					
Presidential succession after Vice President	67.7 (1.4)	60.3 (1.9)	65.2 (2.0)		
House vote percentage to override veto	75.8 (1.3)	87.9 (1.2)	73.6 (1.3)		
Number of terms an individual can be elected president	89.4 (0.9)	97.4 (0.6)	92.8 (0.7)		
Length of a US Senate term	51.8 (1.5)	62.5 (1.8)	37.5 (1.3)		
Number of Senators per state	78.5 (1.2)	83.2 (1.4)	73.2 (1.2)		
Length of a US House term	51.3 (1.5)	49.3 (1.9)	38.9 (1.3)		
Average	69.1	73.5	63.5		
N	909-8,122	673-705	2,727-3,003	92,311-102,011	2,004-2,054

*Standard errors are in parentheses. N is a range because of differing missingness across survey questions. * indicates turnout in 2012. Political interest is on a 5-point scale with 5 indicating high interest.*

have the same approximate magnitude, directionality, and level of statistical significance as in the original studies. They include studies using a range of other online platforms, including MTurk, and more expensive samples including the General Social Survey (GSS), Knowledge Networks (KN), Qualtrics, and the Danish firm Epinion. Table 2 summarizes these replications, with additional descriptions below and in Appendix B.

We begin by showing that all three of the experiments replicated by Berinsky, Huber, and Lenz (2012) were replicable using volunteers. First, we replicate Rasinski (1989) (Appendix Table A3), who using data from the GSS found that framing policy choices dramatically changes stated preferences for redistributive policies. Second, we replicate the well-known Asian Disease Problem popularized by Tversky and Kahneman (1981) (Appendix Table A4), who found that framing policy options in terms of losses (deaths) rather than gains (lives saved) leads to stronger preferences for probabilistic outcomes. Third, we replicate Kam and Simas (2010) (Appendix Table A5), who demonstrate that individuals willing to accept higher amounts of risk are more likely to support probabilistic policy outcomes.

In addition, we used volunteers to replicate three recent, prominent studies arguably involving more complex sociopolitical attitudes. First, we replicate Tomz (2007)'s study on the microfoundations of audience costs in international relations (Appendix Table A6). Second, we use a volunteer sample to replicate Hainmueller and Hiscox (2010)'s experiment that challenges traditional political economy theories about preferences for and against immigration (Appendix Figure A4). Third, we replicate Gadarian and Albertson (2014)'s finding that different levels of anxiety affect how individuals search for information about immigration.

In addition, Table 2 reports other research that has used volunteers to replicate findings across a range of topics. We provide additional details in the Appendix to provide a broader sense of the variety of research volunteer laboratories can reproduce.

Table 2: Experimental social science replicated on DLABSS

Replicated Study	Dependent Variable	N	MTurk	N	Other	N
Tversky and Kahneman (1981)	Risk acceptance	539	✓	450	students	307
Rasinski (1989)	Support for government spending	788	✓	329	GSS	1,470
Tomz (2007)	Audience costs	495			KN	1,127
Hainmueller and Hiscox (2010)	Immigration attitudes	736	✓*	833	KN	1,601
Kam and Simas (2010)	Policy acceptance	752	✓	699	KN	752
Gadarian and Albertson (2014)	Information seeking	668	✓*	736	KN	384
Krosch et al. (2013)	Perceptions of race	204	✓	31	Qualtrics	708
Enos and Carney (2015)	Racism scales	1,478	✓	4,488	TESS	733
Enos and Celaya (2015)	Perceptions of race	365	✓	716		
Mahler (2016)	Voting outcomes	400			Epinion	2,000
Hankinson (2017)	Housing preferences	655	✓	803		
Bonikowski and Zhang (2017)	Populism	642	✓	421	Qualtrics	1,035
Kaufman (2018)	Survey bias	272	✓	524		
Kaufman, King, and Komisarchik (2018)	District compactness	373	✓	764		
Saha and Weeks (2018)	Candidate ambition	550			SSI	1200
Mozer et al. (2018)	Article similarity	226	✓	336		

*The "Other" column indicates the first sample, of which we are aware, other than MTurk or DLABSS, on which the study was carried out and is not an exhaustive list of replications. The first N column is the number of subjects on DLABSS, the second N is the number of subjects on MTurk, and the third N is the number of subjects on different platforms, where applicable. A * next to the ✓ for MTurk indicates that we carried out the MTurk replication ourselves.*

3.4 Testing Volunteer and Paid Subject Response Quality

Our final tests explore response quality across volunteer and paid subjects. In order to establish that any observed differences are not idiosyncratic to subject matter, we used two different surveys, each fielded on both DLABSS and MTurk. The two surveys vary only in substantive issue area: the first focused on religion and secularism in the United States (N= 557 DLABSS, 459 MTurk), and the second on foreign economic policies (N = 519 DLABSS, 482 Mturk). In Table A8 we compare the demographic composition of the volunteer and paid samples and find balance across a number of covariates.[10]

As there are a number of approaches to measuring response quality, we adopt eleven tests across seven quality dimensions. Drawing on relevant literature (e.g. Singer and Maher 2000, James and Bolstein 1990, Glalesic and Bosnjak 2009), we measure response quality based on subjects' propensity to 1) invest time in reading a prompt and answering questions; 2) answer grid-style questions without engaging in straightlining; 3) invest effort into open-ended responses; 4) offer committal answers; 5) answer opinion questions consistently at different points in the survey; 6) avoid skipping questions; and 7) catch an embedded attention check. Table 3 summarizes each test. Appendix Section D.3 further describes our measurement strategies, presenting detailed results of each test.[11]

Figure 1 presents a summary of the results for each test across both surveys.[12] Our quantity of interest is the average difference in response between volunteer and paid subjects. We measure this difference from an OLS regression of each dependent variable on a dummy variable for whether the subject was a volunteer or not, with positive values representing higher quality. The coefficient on volunteer is thus the added quality of a volunteer subject. We

[10]Due to missingness on some covariates and attrition, effective sample sizes vary across the tests of response quality we present below. See Table A7 for details.

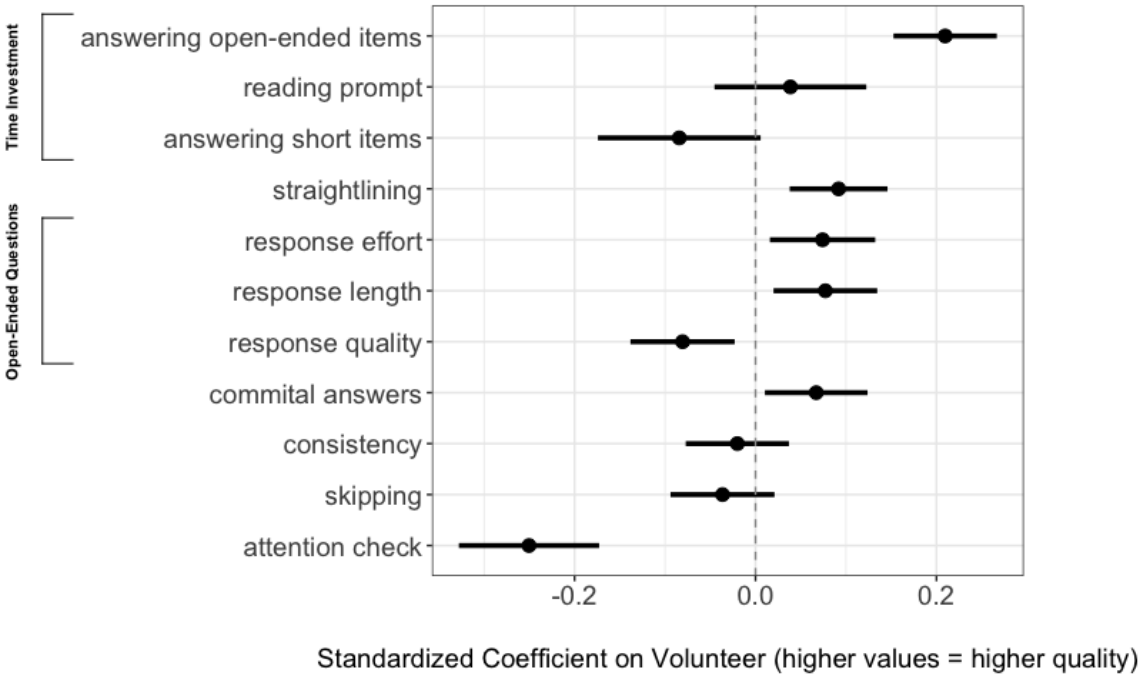
[11]We employ a range of tests since some of these measures, such as straightlining, are clear measures of response quality, but others, such as question skipping (which might be due to uncertainty and caution rather than rushing) are more ambiguous.

[12]Appendix Figure A6 includes disaggregated multivariate results for the secularism and foreign policy surveys. Figure A7 includes bivariate results.

Table 3: Tests of Response Quality

Quality Dimension	**Test**	**Measurement**
Time Investment	Time answering open-ended items	Seconds spent on response page for each of two open-ended survey items (presented as an average and separately)
	Time reading prompt	Seconds viewing an article of about 400 words
	Time answering short items	Seconds spent on response page with five short-answer/multiple choice items pertaining to article
Straightlining	Straightlining in matrix-style question grids	Share of three bidirectional question matrices with entirely uniform answers
Open-Ended Investment	Subjective effort	Effort score out of five (with five being most perceived effort) given by two human coders for one open-ended survey item
	Response length	Number of characters in response to each of two open-ended survey items (presented as an average and separately)
	Subjective response quality	Summary of three dimensions coded 0 or 1 by two human coders for one open-ended survey item: whether response is long, topical, and complete
Committal Answers	"Don't Know" answers regarding commitment to action	Whether respondent answers "Don't know" to each of two committal questions (presented as an average and separately)
Consistency	Contradicting previous responses	Whether subject direction of subjects' response to a multiple-choice question contradicts previous response in grid-style question
Skipping	Skipping questions when given opportunity	Whether subject picks the "Skip" option in a multiple choice opinion question about policy opinions
Attention Check	Noticing embedded "attention check"	Whether subject answers a factual short-item question with "yes," rather than the true answer, as directed in a reading prompt

Figure 1: Standardized "Volunteer" Coefficients for All Response Quality Tests



Coefficients on "reading prompt," "answering short items," and "attention check" represent results for a single substantive survey, because they were not included in both questionnaires.

control for individual demographic covariates that have some unbalance across the samples (See Appendix Section D.3).

Volunteers tended to perform as well as, and in many cases better than, paid subjects. On five tests of quality, including time invested reading a prompt, propensity for straightlining, open-ended response length, and committal, volunteers offered statistically significantly higher-quality responses. On four other tests, volunteers and paid subjects were statistically indistinguishable. In two tests, paid subjects' responses were higher quality: first, open-ended responses, for which paid subjects scored higher on a subjective measure of overall quality, while volunteers scored higher on response effort and length. Second, the attention check, or "screener," embedded in a text block in the secularism survey, on which paid subjects were more likely to correctly address the check. This indicates that paid subjects may

have been more carefully reading the survey, although some of the difference may have come from familiarity with these checks resulting from the frequency with which paid subjects participate in studies.

The overall results from this series of original tests strengthen our confidence in the benefits of using a large-scale volunteer panel. Volunteers, relative to paid subjects, appear to produce higher-quality survey responses across a number of dimensions and differ little from paid subjects on others.

4 Conclusion

Building on the innovation of online crowd-sourcing, we argue that volunteer laboratories extend the accessibility and quality frontiers of online human-subjects research. A large body of literature has demonstrated the viability of paid online subjects for social science survey-based research (Buhrmester, Kwang, and Gosling 2011, Berinsky, Huber, and Lenz 2012). Volunteer subjects resembling these paid populations add another source of data for human-subjects researchers. To show this potential, we replicated a series of diverse studies using a volunteer laboratory and highlighted advantages of volunteer subjects through a series of tests.

The quality and scalability of online volunteer pools is encouraging. Of course, volunteer laboratories, like all data sources, have advantages and disadvantages that will allow them occupy a particular niche in human-subjects research, and will likely emerge as complements, rather than replacements, to other survey platforms. Coordination across institutions would significantly increase the efficiency of volunteer laboratories, in much the same way that economies of scale in designs like the Cooperative Congressional Election Study (CCES) have greatly improved the efficiency of survey research. We urge researchers to undertake such collaborations.

References

Berinsky, Adam J, Gregory A Huber, and Gabriel S Lenz. 2012. "Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk." *Political Analysis* 20(3): 351–368.

Bonikowski, Bart, and Yueran Zhang. 2017. "Populism as Dog-Whistle Politics: Anti-Elite Discourse and Sentiments toward Minorities in the 2016 Presidential Election.".

Buhrmester, Michael, Tracy Kwang, and Samuel D Gosling. 2011. "Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data?" *Perspectives on Psychological Science* 6(1): 3–5.

Dupuis, Marc, Emanuele Meier, and Félix Cuneo. 2018. "Detecting Computer-generated Random Responding in Questionnaire-based Data: A Comparison of Seven Indices." *Behavior research methods* pp. 1–10.

Enos, Ryan D., and Christopher Celaya. 2015. Segregation Directly Affects Human Perception and Intergroup Bias. In *American Political Science Association, Annual Meeting*. San Francisco: .

Enos, Ryan D., and Riley Carney. 2015. "Is Modern Racism Caused by Anti-Black Affect?: An Experimental Investigation of the Attitudes Measured by Modern Racism Scales.".

Frey, Bruno S, and Reto Jegen. 2001. "Motivation crowding theory." *Journal of economic surveys* 15(5): 589–611.

Gadarian, Shana Kushner, and Bethany Albertson. 2014. "Anxiety, immigration, and the search for information." *Political Psychology* 35(2): 133–164.

Glalesic, Mirta, and Michael Bosnjak. 2009. "Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey." *Public Opinion Quarterly* 73(2): 349–360.

Hainmueller, Jens, and Michael J Hiscox. 2010. "Attitudes toward highly skilled and low-skilled immigration: Evidence from a survey experiment." *American Political Science Review* 104(01): 61–84.

Hankinson, Michael. 2017. "Do NIMBYs Think Outside of Their Neighborhood? Free-Riding and Fairness in Collective Action.".

James, Jeannine M., and Richard Bolstein. 1990. "The Effect of Monetary Incentives and Follow-Up Mailings on the Response Rate and Response Quality in Mail Surveys." *Public Opinion Quarterly* 54(3): 346–361.

Kam, Cindy D, and Elizabeth N Simas. 2010. "Risk orientations and policy frames." *The*

Journal of Politics 72(02): 381–396.

Kaufman, Aaron. 2018. "An Automated Method to Estimate Bias in Survey Questions.".

Kaufman, Aaron, Gary King, and Mayya Komisarchik. 2018. "How to Measure Legislative District Compactness If You Only Know it When You See It.".

King, Gary. 2014. "Restructuring the Social Sciences: Reflections from Harvard's Institute for Quantitative Social Science." *PS: Political Science & Politics* 47(01): 165–172.

Klar, Samara, and Yanna Krupnikov. 2016. . *Independent Politics: How American Disdain for Parties Leads to Political Inaction*. New York: Cambridge University Press.

Krosch, Amy R, Leslie Berntsen, David M Amodio, John T Jost, and Jay J Van Bavel. 2013. "On the ideology of hypodescent: Political conservatism predicts categorization of racially ambiguous faces as Black." *Journal of Experimental Social Psychology* 49(6): 1196–1203.

Mahler, Daniel. 2016. "Do Altruistic Preferences Matter for Voting Outcomes?".

Meier, Stephan, and Alois Stutzer. 2008. "Is Volunteering Rewarding in Itself?" *Economica* 75(297): 39–59.

Mozer, Reagan, Luke Miratrix, Aaron R. Kaufman, and L. Jason Anastasopoulos. 2018. "Matching with Text Data: An Experimental Evaluation of Methods for Matching Documents and of Measuring Match Quality.".

Mullinix, Kevin J, Thomas J Leeper, James N Druckman, and Jeremy Freese. 2015. "The generalizability of survey experiments." *Journal of Experimental Political Science* 2(02): 109–138.

Paolacci, Gabriele, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. "Running experiments on amazon mechanical turk." *Judgment and Decision making* 5(5): 411–419.

Radford, Jason, Andy Pilny, Katya Ognyanova, Luke Horgan, Stefan Wojcik, and David Lazer. 2016. Gaming for Science: A Demo of Online Experiments on VolunteerScience. com. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*. ACM pp. 86–89.

Rasinski, Kenneth A. 1989. "The effect of question wording on public support for government spending." *Public Opinion Quarterly* 53(3): 388–394.

Rosenthal, Robert, and Ralph L Rosnow. 1975. "The volunteer subject." *Australian Journal of Psychology* 28(2): 97–108.

Rosnow, Ralph L., and Robert Rosenthal. 1997. *People Studying People*. New York: W.H. Freeman and Company.

Ryan, Richard M, and Edward L Deci. 2000. "Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions." *Contemporary Educational Psychology* 25(1): 54–67.

Saha, Sparsha, and Ana Catalano Weeks. 2018. "Ambitious Women: Gender and Voter Perceptions of Candidate Ambition.".

Schroer, Joachim, and Guido Hertel. 2009. "Voluntary Engagement in an Open Web-based Encyclopedia: Wikipedians and Why They Do It." *Media Psychology* 12(1): 96–120.

Sharpe Wessling, Kathryn, Joel Huber, and Oded Netzer. 2017. "MTurk Character Misrepresentation: Assessment and Solutions." *Journal of Consumer Research* 44(1): 211–230.

Singer, Eleanor, Van Hoewyk John, and Mary P. Maher. 2000. "Experiments with incentives in telephone surveys." *Public Opinion Quarterly* 64(2): 171–188.

Stewart, Neil, Christoph Ungemach, Adam JL Harris, Daniel M Bartels, Ben R Newell, Gabriele Paolacci, and Jesse Chandler. 2015. "The Average Laboratory Samples a Population of 7,300 Amazon Mechanical Turk workers." *Judgment and Decision Making* 10(5): 479.

Titmuss, Richard M. 1970. "The gift relationship." *London* 19: 70.

Tomz, Michael. 2007. "Domestic audience costs in international relations: An experimental approach." *International Organization* 61(04): 821–840.

Tversky, Amos, and Daniel Kahneman. 1981. "The framing of decisions and the psychology of choice." *Science* 211(4481): 453–458.

Williamson, Vanessa. 2016. "On the Ethics of Crowdsourced Research." *PS: Political Science & Politics* 49(01): 77–81.

Zaller, John R. 1992. *The Nature and Origins of Mass Opinion*. New York: Cambridge University Press.

Zhou, Haotian, Ayelet Fishbach, Franklin Shaddy, Janina Steinmetz, Jessica Bregant, Juliana Schroeder, Kaitlin Woolley, Natalie Wheeler, Oliver Sheldon, Sarah Molouki et al. 2016. "The Pitfall of Experimenting on the Web: How Unattended Selective Attrition Leads to Surprising (yet False) Research Conclusions." *Journal of Personality and Social Psychology, forthcoming* .

Appendices for *Intrinsic Motivation at Scale: Online Volunteer Laboratories for Social Science Research*

Contents

A Additional Background Information on DLABSS **2**
A.1 DLABSS Operations 2
A.2 DLABSS Subject Attributes 3

B Additional Replications Using Volunteers **8**

C Additional Analysis on Subject Motivations **13**

D Additional Analysis on Response Quality **15**
D.1 Sample Composition 15
D.2 Response Quality by Survey Topic 16
D.3 Design and Results Details for All Tests 16

A Additional Background Information on DLABSS

This appendix includes additional information on the operations and properties of the volunteer lab used in this paer, the Harvard Digital Lab for the Social Sciences (DLABSS).

A.1 DLABSS Operations

DLABSS was created in 2014 with the principle that volunteer online subjects should be a public good to researchers with diverse substantive interests. DLABSS is staffed by several part-time researchers, including a faculty member and graduate and undergraduate students. The basic structure is a website that serves as a clearinghouse for experimental and other survey-based social science studies. The research instruments are not directly hosted on DLABSS, rather staff curate links to the instruments along with basic descriptions on the website. This simple structure allows researchers to host their instruments on any other convenient platform, such as Qualtrics or an application built specifically for the research. Using Qualtrics, the lab collects basic demographic and attitudinal data on all subjects in the lab. This simple structure of the lab allows researchers to collect response data for virtually any type of experimental social science research question.

Subjects are recruited using a variety of methods, including organic web search, paid search, social media, and email. Thus far, Craigslist, Reddit, social media, and targeted ad campaigns on email newsletters for specific demographic groups have been the primary channels for recruiting new subjects.

Specific studies are often used to attract subjects and, as such, subjects for a given study usually consist of two types: 1) new subjects that enter an experiment after being recruited from the web and 2) subjects from the existing subject pool who enter an experiment after being directly solicited, usually by email. DLABSS staff email a fraction of the active subject pool each week to invite them to participate in the latest studies. Subjects are emailed

biweekly on average. These email solicitations can allow a research to shape her subject pool, if desired, by targeting certain populations based on known covariates. In this manner, DLABSS shares this targeting ability with other opt-in internet panels, but maintains the low-cost structure of crowd-sourced research.

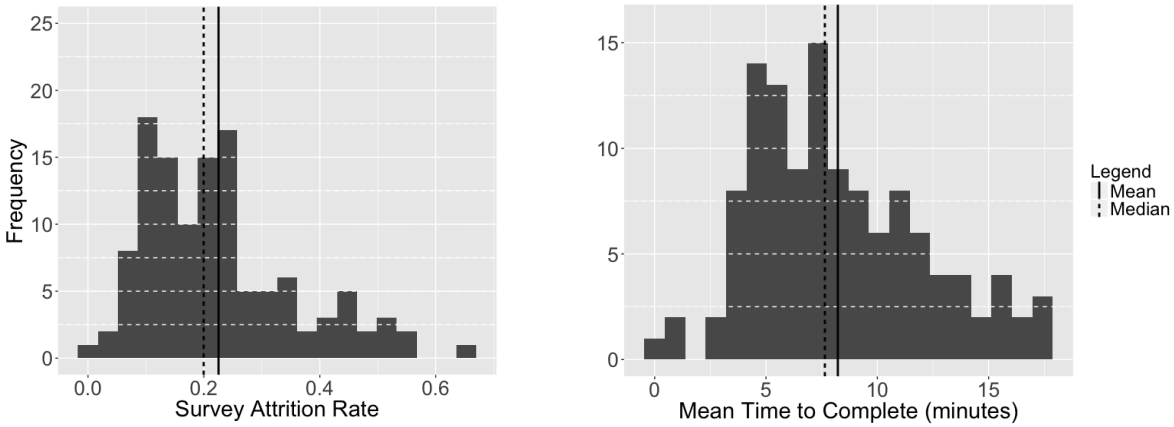
As of October 2018, the active DLABSS volunteer pool includes over 13,000 subjects. This means that since its inception, DLABSS has attracted almost 9 new subjects per day. Notably, these descriptive figures include weekends and summer and other academic holidays, and subject recruitment is significantly higher during the periods of the academic calendar where DLABSS staff are working. Subject acquisition discussed above is linked to the nature and diversity of research studies hosted on DLABSS. Since its inception, DLABSS has hosted over 100 social science studies. 35% of these were Harvard faculty projects, 60% were graduate student projects, and 5% were undergraduate student projects.

Prior to beginning operation, DLABSS secured approval from an institutional revie board to recruit subjects and collect demographic data. Researchers using the lab secure human subjects approval for their individual studies. The use of deception in studies is left to the discretion of the researcher. Studies have come from a variety of disciplines, including business, economics, sociology, political science, public policy, and psychology. The majority of studies on DLABSS have been collaborative and have involved two or more social scientists.

A.2 DLABSS Subject Attributes

In the manuscript, we discuss the way that volunteer subjects potentially exhibit less propensity for attrition than paid subjects. Figure A1 displays the mean and median attrition rates and completion times for DLABSS studies. We also explore the representativeness of volunteer samples. We expand on this with Figure A3, which depicts the geographic locations of DLABSS volunteers, and Table A1, which compares the DLABSS volunteer pool with samples from online convenience samples and nationally-representative survey samples.

Figure A1: DLABSS Study Attrition Rates and Completion Times



Left figure is total attrition by study and right figure is mean time to complete by study for 120 studies hosted by DLABSS.

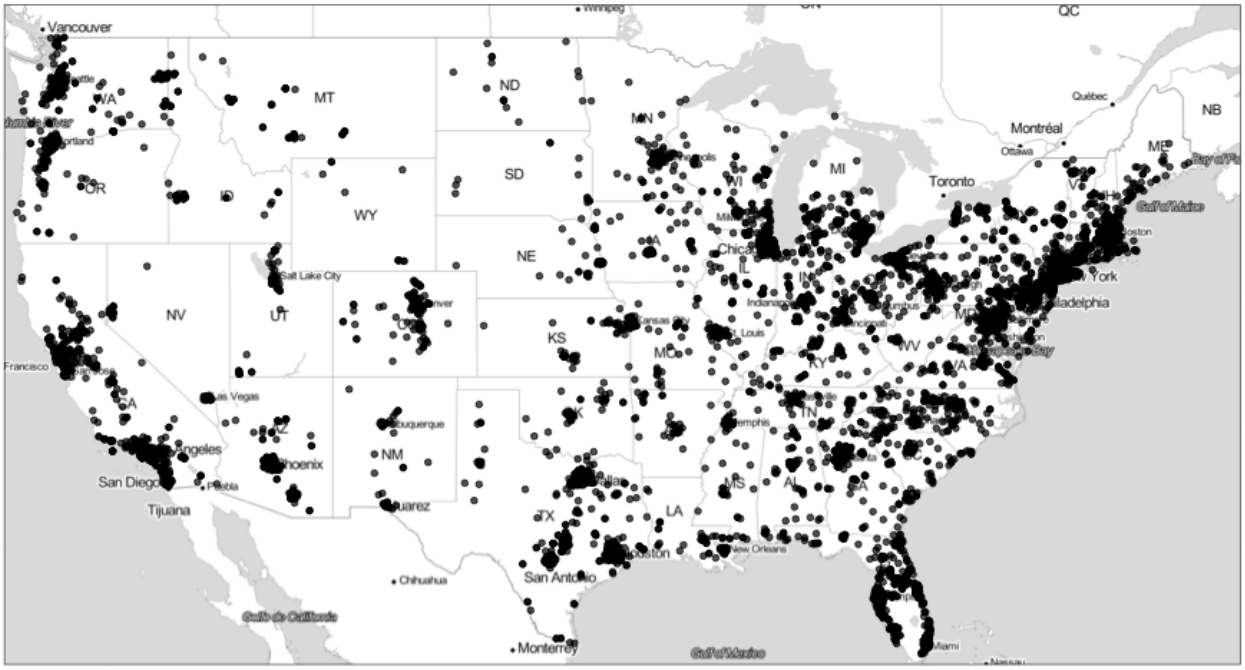
Despite some differences in sample characteristics, volunteers report relatively similar policy attitudes to other survey samples in Table A2. The exceptions to this are immigration and gay marriage attitudes, on which the DLABSS sample reports more liberal attitudes than some other platforms. However, on immigration DLABSS subjects report attitudes similar to those of a face-to-face sample from the ANES in 2012. On gay marriage, DLABSS subjects report attitudes relatively similar to MTurk subjects, who are also more liberal in their attitudes than nationally representative samples.

Table A1: Comparing Subject Properties across DLABSS and Other Convenience Samples

			Convenience samples			
					Adult samples (Berinsky and Kinder 2006)	
Demographics	*DLABSS*	*MTurk*	*Student samples (Kam et al. 2007)*	*Adult sample (Kam et al 2007)*	*Experiment 1: Ann Arbor, MI*	*Experiment 2: Princeton, NJ*
Female	55.6% (0.7)	48.0% (1.9)	56.7% (1.3)	75.7% (4.1)	66.0%	57.1 %
Age (mean years)	44.1 (0.2)	37.8 (0.5)	20.3 (8.2)	45.5 (.916)	42.5	45.3
Education (mean years)	15.2 (0.0)	14.9 (0.1)	–	5.48 (1.29)	15.1	14.9
White	74.7 (0.5)	78.3 (1.6)	42.5	82.2 (3.7)	81.4	72.4
Black	6.4 (0.3)	8.4 (1.0)			12.9	22.7
Party identification						
Democrat	46.9 (0.6)	44.3 (1.9)			46.1	46.5
Independent/Other	29.6 (0.6)	30.1 (1.7)			37.6	27.7
Republican	23.5 (0.5)	22.8 (1.6)			16.3	25.8
N	807-6,280	673-705	277-1428	109	141	163

DLABSS and MTurk results from December 2016, all other results from Berinsky, Huber, and Lenz (2012). Note that the Education in the Kam et al 2007 sample is a ordinal indicator, rather than years, and is as reported in Berinsky, Huber, and Lenz (2012).

Figure A2: United States Distribution of DLABSS Volunteer Subjects



A total of 10,506 points are plotted on this map, each one representing a single DLABSS participant.

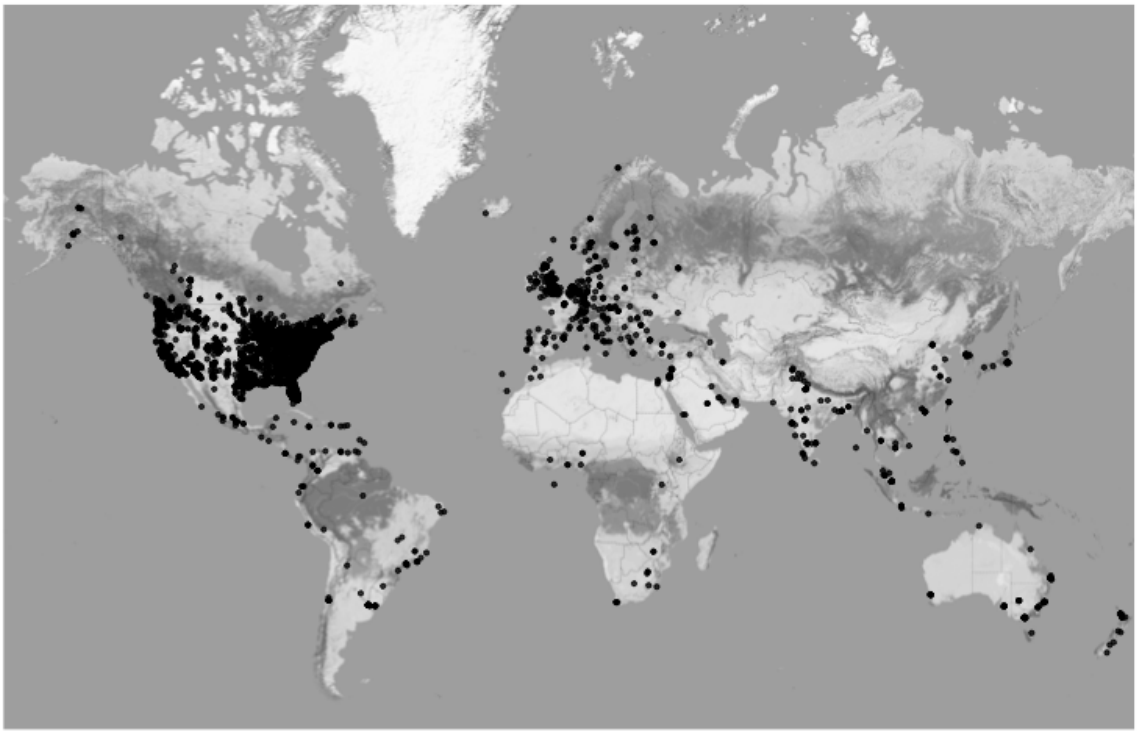
Table A2: Comparing DLABSS sample policy attitudes

	DLABSS	*MTurk*	*Internet sample ANESP*	*Face-to-face samples ANES 2012*
Favor prescription drug benefit for seniors	75.7% (1.3)	71.6% (1.7)	74.8% (1.1)	
Favor universal health care	54.5 (1.5)	58.8 (1.9)	41.7 (1.2)	
Favor citizenship process for illegals	62.9 (1.5)	48.1 (1.9)	42.7 (1.2)	63.7 (1.1)
Favor a constitutional amendment banning gay marriage	10.2 (0.9)	21.3 (1.5)	55.4 (1.2)	57.1 (1.1)*
Favor raising taxes on people making more than $200,000	69.9 (1.4)	67.2 (1.8)	55.4 (1.2)	79.3 (0.9)**
Favor raising taxes on people making less than $200,000	7.5 (0.8)	8.9 (0.1)	7.1 (0.6)	
N	1,041-1,044	703-705	1,614-1,618	1,995-2,026

** Gay and lesbian couples should be allowed to form civil unions but not legally marry, or there should be no legal recognition of a gay or lesbian couple's relationship*

*** Increasing income taxes on people making over one million dollars per year.*

Figure A3: World-wide Distribution of DLABSS Volunteer Subjects



A total of 11,717 points are plotted on this map, each one representing a single DLABSS participant.

B Additional Replications Using Volunteers

In this appendix, we offer additional information regarding replications conducted on DLABSS. In Tables A3, A4, A5, and A6, and in Figure A4 we display results for the replications of three well-known studies.

Next, we provide additional information on studies conducted by researchers aside from the authors that have replicated results on DLABSS in the course of their research.

Several studies hosted on DLABSS have explored racial politics. One study (Enos 2017), using both DLABSS and Qualtrics ' proprietary survey panel, replicated findings from prominent recent studies that, using small MTurk samples, found significant links between political ideology and visual perceptions of race (Krosch et al. 2013, Krosch and Amodio 2014). Another study tested how spatial segregation affects perceptions of similarity in human faces across DLABSS and MTurk (Enos and Celaya 2015). Several DLABSS studies also investigated the properties of Modern Racism Scales (Sears and Kinder 1971), finding similar distributions of racial attitudes as those in the Cooperative Campaign Analysis Project (CCAP) survey and replicating experimental results on the nationally representative Time Sharing for Experimental Social Science (TESS) panel and MTurk (Enos and Carney 2015).

Another researcher used DLABSS to study populism. DLABSS and MTurk samples produced similar results, while a Qualtrics panel, which was manipulated to be disproportionately conservative, produced larger effects (Bonikowski and Zhang 2017).

In the context of studying blocked randomization designs, a researcher studied a vari-

Table A3: Replication of Rasinski (1989) in DLABSS

	Platform	Poor	Welfare	Difference	p	n
1	DLABSS	64	39	25	<.001	788
2	General Social Surveys (GSS)	64	23	37	<.001	1470
3	MTurk (Berinsky et al. 2012)	55	17	38	<.001	329

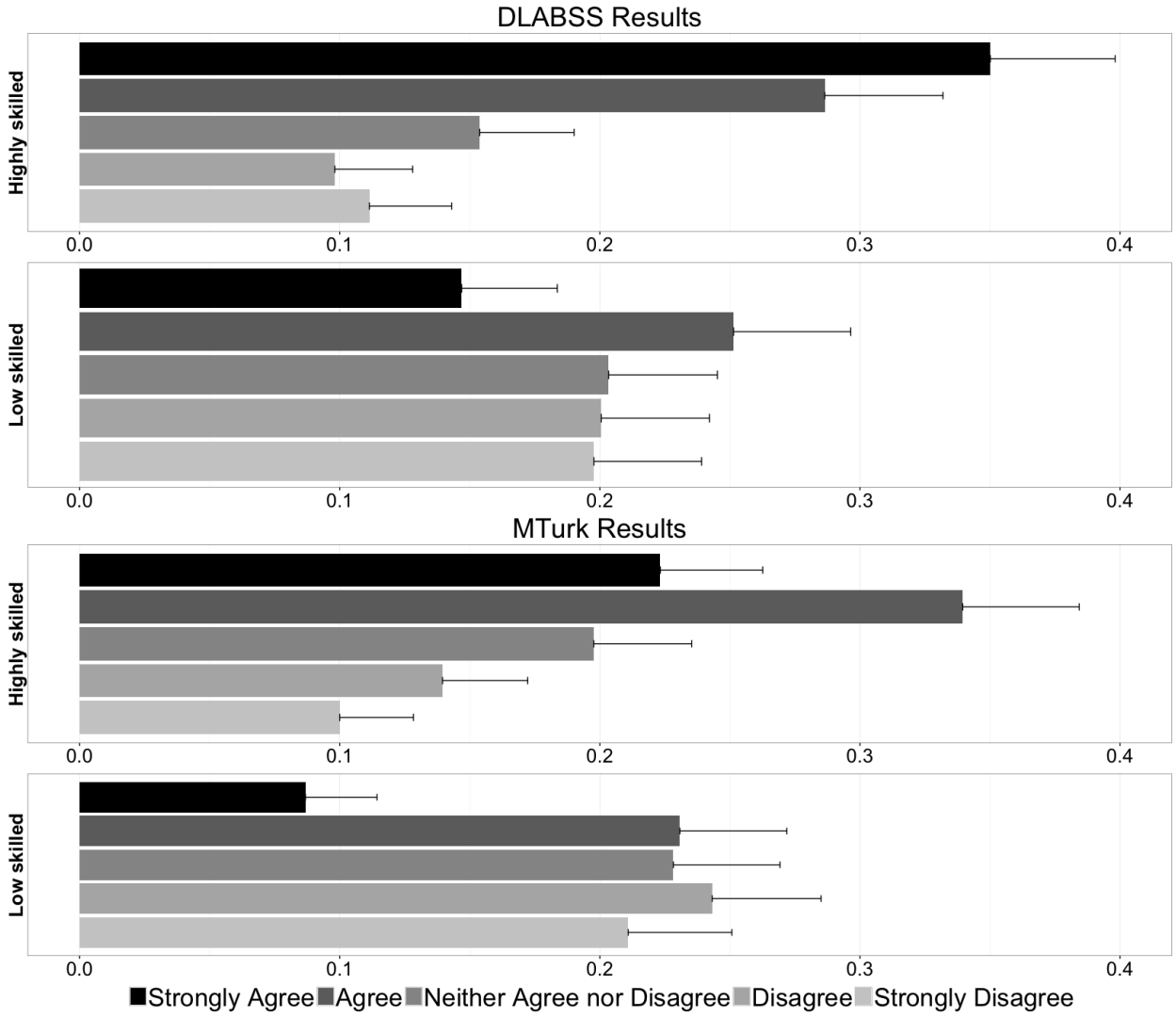
Cells represent percent of respondents favoring a policy with each frame. P values are from a T-test of difference of means.

Table A4: Replication of Tversky and Kahneman (1981) in DLABSS

	Platform	Lives Saved	Lives Lost	Difference	p	n
1	DLABSS	63	34	29	<.001	539
2	MTurk (Berinsky et al. 2012)	74	38	36	<.001	450
3	Tversky and Kahneman 1981	72	22	50	<.001	307

Cells are percent of respondents choosing non-probabilistic (certain) outcome with each frame.

Figure A4: Replication of Hainmueller and Hiscox (2010): Support for Highly- and Low-skilled Immigration among DLABSS Respondents



Whiskers are the upper bounds of 95% confidence intervals for proportions. Respondents in the "highly-skilled" group were asked "Do you agree or disagree that the US should allow more highly skilled immigrants from other countries to come and live here? (emphasis added)?" Respondents in the " low-skilled" group were asked "Do you agree or disagree that the US should allow more low-skilled immigrants from other countries to come and live here? (emphasis added)?"

Table A5: Replication of Kam and Simas (2010) in DLABSS

	Kam and Simas (2010)			*Berinsky et al. MTurk Replication*			*DLABSS Replication*		
	(H1a) Mortality frame and risk acceptance	*(H1b) Adding controls*	*(H2) Frame x Risk acceptance*	*(H1a)*	*(H1b)*	*(H2)*	*(H1a)*	*(H1b)*	*(H2)*
Mortality frame	1.068	1.082	1.058	1.180	1.180	1.410	1.011	1.026	1.437
in Trial 1	(0.10)	(0.10)	(0.29)	(0.10)	(0.10)	(0.31)	(0.11)	(0.11)	(0.36)
Risk	0.521	0.628	0.507	0.760	0.780	0.990	1.024	1.029	1.424
acceptance	(0.31)	(0.32)	(0.48)	(0.29)	(0.31)	(0.42)	(0.33)	(0.34)	(0.46)
Female		0.105			-0.018			-0.013	
		(0.10)			(0.11)			(0.11)	
Age		0.262			0.110			0.443	
		(0.22)			(0.31)			(0.24)	
Education		-0.214			0.025			-0.056	
		(0.20)			(0.23)			(0.23)	
Income		0.205			-0.024			-0.022	
		(0.23)			(0.23)			(0.21)	
Partisan		0.038			0.006			0.013	
ideology		(0.19)			(0.15)			(0.13)	
Risk acceptance x			0.023			-0.450			-0.827
Mortality frame			(0.62)			(0.58)			(0.66)
Intercept	-0.706	-0.933	-0.700	-1.060	-1.100	-1.190	-1.098	-1.256	-1.309
	(0.155)	(0.259)	(0.227)	(0.170)	(0.290)	(0.230)	(0.187)	(0.257)	(0.255)
N	752	750	752	699	699	699	634	597	634

Cells are signs and p-values for probit regressions of individual-level acceptance of probabilistic policy outcomes on risk acceptance attitudes (top row) and other covariates.

Table A6: Replication of Tomz (2007) in DLABSS

	DLABSS Replication of Tomz (2007) Table 1				Tomz (2007 Table 1)			
	Public reaction to empty threat (%)	*Public reaction to staying out (%)*	*Difference in opinion (%)*	*Summary of differences (%)*	*Public reaction to empty threat (%)*	*Public reaction to staying out (%)*	*Difference in opinion (%)*	*Summary of differences (%)*
Disapprove								
Disapprove very strongly	27 (21 to 32)	14 (10 to 19)	12 (6 to 20)	14 (6 to 22)	31 (27 to 35)	20 (17 to 23)	11 (6 to 17)	16 (10 to 22)
Disapprove somewhat	30 (25 to 36)	15 (11 to 20)	15 (8 to 23)		18 (14 to 21)	13 (10 to 16)	5 (0 to 9)	
Neither								
Lean toward disapproving	5 (3 to 9)	14 (10 to 19)	-8 (-14 to -3)	-3 (-12 to 5)	8 (6 to 11)	9 (7 to 11)	0 (-3 to 3)	-4 (-9 to 2)
Don't lean either way	10 (7 to 14)	13 (9 to 17)	-2 (-8 to 4)		21 (17 to 24)	21 (18 to 24)	0 (-5 to 4)	
Lean toward approving	12 (9 to 17)	12 (8 to 16)	0 (-5 to 7)		8 (6 to 11)	11 (9 to 14)	-3 (-6 to 0)	
Approve								
Approve somewhat	10 (6 to 13)	21 (16 to 27)	-12 (-18 to -5)	-9 (-17 to -2)	8 (5 to 10)	13 (11 to 16)	-6 (-9 to -2)	-12 (-17 to -8)
Approve very strongly	5 (3 to 8)	11 (8 to 16)	-6 (-11 to -1)		6 (4 to 9)	13 (10 to 16)	-7 (-10 to -3)	

The table gives the percentage of respondents who expressed each opinion. Bayesian 95 percent credible intervals appear in parentheses.

ant of the Tomz (2007) study referenced above and replicated the results on MTurk and DLABSS (Kaufman and Kim 2017). Another team crowdsourced perceptions of the compactness of legislative districts on both MTurk and DLABSS with similar results between the two platforms (Kaufman, King, and Komisarchik 2018). And researchers used MTurk and DLABSS to validate a computational model of sentiment analysis of survey questions with similar results across the platforms (Kaufman 2017).

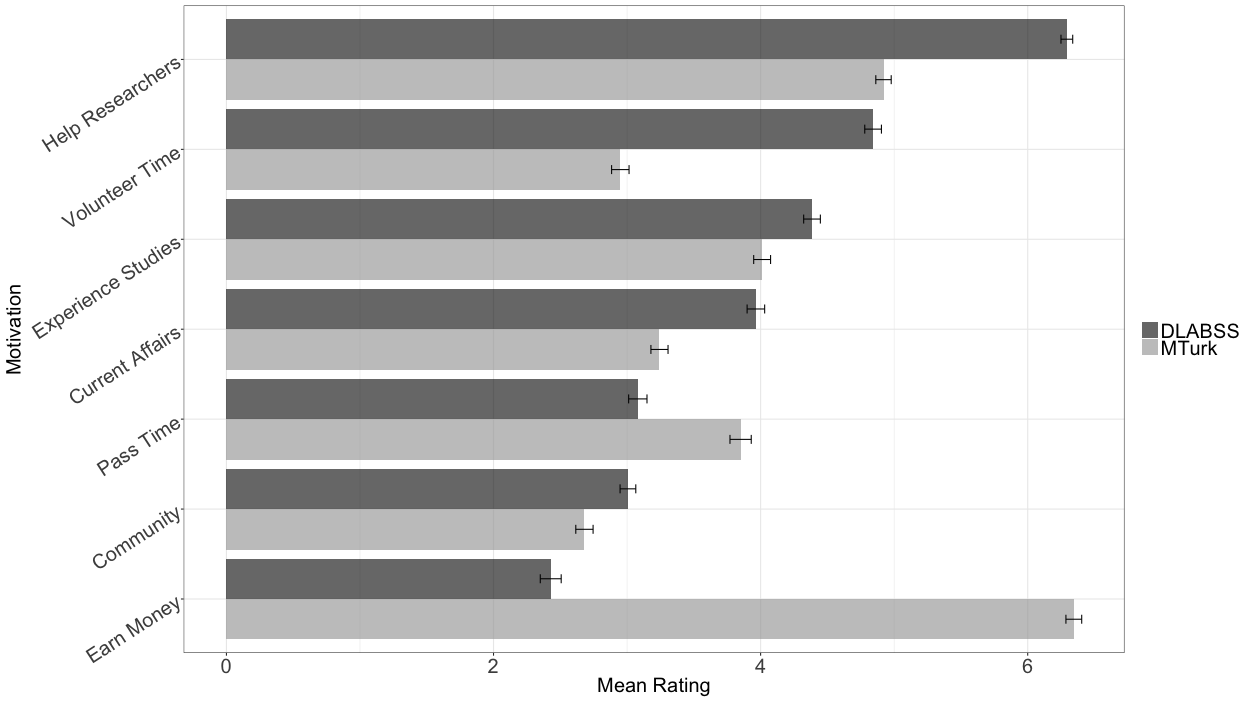
In an intriguing finding on the effects of altruistic voting behavior on voting outcomes, a researcher used a representative Danish sample from Epinion and replicated the result on DLABSS with U.S. subjects (Mahler 2016). Finally, another researcher replicated a survey experiment from MTurk on preferences for housing allocation based on the geographic location of the housing (Hankinson 2017).

C Additional Analysis on Subject Motivations

In this appendix, we provide detailed results from our original survey of paid and volunteer subjects related to their motivations for participating in survey research. Figure A5 depicts the mean rating of each of seven possible motivations among paid and unpaid subjects.

Figure A5 illustrates the results of our test of respondent motivation across DLABSS and MTurk participants mentioned in the main text. Volunteer and paid subjects clearly report fundamentally different motivations for participating in online survey research. Paid subjects disporportionately value earning money, and also appear to place relative currency in helping researchers, experiencing new studies and simply passing time. In contrast, earning money is the least important factor reported by volunteer respondents, who instead emphasize helping researchers, simply contributing their time as a volunteer, experiencing new studies and learning about current affairs.

Figure A5: Comparing DLABSS and MTurk Self-reported Motivations



Black bars represent standard errors of the mean.

D Additional Analysis on Response Quality

This appendix includes additional information on the design and results of individual response quality tests.

D.1 Sample Composition

We provide an overview of the sample sizes for each of the tests of response quality discussed in the manuscript. Due to a combination of missingness in demographic covariates, plus some attrition throughout the survey, the sample size for each test ranges from around 1,400 to 1,700 when combined across survey topics and volunteer status (Table A7). The table suggests even balance across paid and volunteer subjects in missingness and attrition.

Table A7: Sample Sizes by Response Quality Test, Survey Topic, and Volunteer Status.

	Sample Size	By Survey Topic		By Volunteer Status	
	Combined	FEP	Secular	MTurk	Volunteer
Time answering open-ended 2	1,477	719	758	755	722
Time answering open-ended 1	1,411	664	747	732	679
Time reading prompt		750		806	741
Time answering short items		697		740	698
Straightlining	1,685	822	863	871	814
Open-ended Response Effort	1,466	708	758	755	711
Open-ended question 2 length	1,475	718	757	753	722
Open-ended question 1 length	1,484	725	759	761	723
Open-ended Response Quality	1,466	708	758	755	711
Commital question 2	1,494	730	764	771	723
Commital question 1	1,493	730	763	771	722
Consistency	1,499	734	765	773	726
Skipping	1,509	741	768	774	735
Attention Check			773	384	389

We compare the demographic balance across the paid and unpaid samples that took the response quality surveys in Table A8. We show the samples are balanced on many variables, but that DLABSS tends to be older, richer, whiter, more religious, and more politically conservative. As noted in the body of the paper, many of these differences actually make DLABSS more similar to the larger US population and are the result of intentional targeting of certain populations in DLABSS recruitment efforts. Notably however, DLABSS does contain more missingness in demographic variables.

D.2 Response Quality by Survey Topic

In the body of the paper, we present plots including the standardized coefficients for volunteer response quality across all tests, including demographic covariates such as age, income, education, race, frequency of religious service attendance, religious tradition, political ideology, and party identification and dummy variable for the survey topic. We report results using the combined survey data across studies. Here, we provide additional background information on test design and present coefficient plots separated by survey (Figure A6). We also present the same coefficient plot for the bivariate regression of each quality test on volunteer status, without controlling for demographic covariates (Figure A7). Due to the embedded attention check in the secularism version of the survey, we can only present results for time investment into reading the article and answering subsequent questions for the FEP survey. Likewise, results for the attention check are limited to the secularism survey.

D.3 Design and Results Details for All Tests

Time Investment: One measure of response quality is time investment in answering survey questions. In general, we consider longer time investment to be a reflection of more careful and higher quality responses. Our primary test involves a reading prompt followed

Table A8: Demographics in DLABSS and MTurk.

	DLABSS	*MTurk*
Female	49.0% (1.8) [5.1%]	50.3% (1.8) [0.1%]
Education (mean years)	16.0 (0.1) [3.7]	14.9 (0.1) [0.0]
Age (mean years)	56.0 (0.6) [2.4]	38.7 (0.5) [0.0]
Mean income	$60,717 ($1,594) [9.1]	$44,727 ($1,192) [0.7]
Median income	$55,000	$37,500
Race		
White	88.9 (1.1) [2.1]	77.0 (1.5) [0.0]
Black	2.4 (0.5)	6.8 (0.9)
Hispanic	4.5 (0.7)	7.4 (0.9)
Attend Religious Service Weekly	33.6 (2.2) [2.5]	15.0 (1.8) [0.0]
Religion		
Protestant	45.7 (2.3) [0.4]	34.4 (2.4) [0.0]
Other Christian	4.9 (1.0)	2.1 (0.7)
Other	14.8 (1.6)	19.9 (2.0)
None	34.6 (2.2)	43.7 (2.5)
Party Identification		
Democrat	37.4 (1.7) [7.9]	50.2 (1.8) [3.2]
Independent	30.5 (1.6)	23.8 (1.5)
Republican	32.1 (1.7)	26.0 (1.6)
Liberal	54.4 (2.3) [4.7]	60.5 (2.5) [0.0]
Speak English at Home	99.8 (0.2) [2.0]	100.0 (0.0) [0.0]
N	476-843	387-782

Standard errors are in parentheses. Percent missing in brackets. N varies across questions due to missingness and the religious questions only appearing on one survey.

Figure A6: Standardized "Volunteer" Coefficients for All Response Quality Tests, by survey with covariates

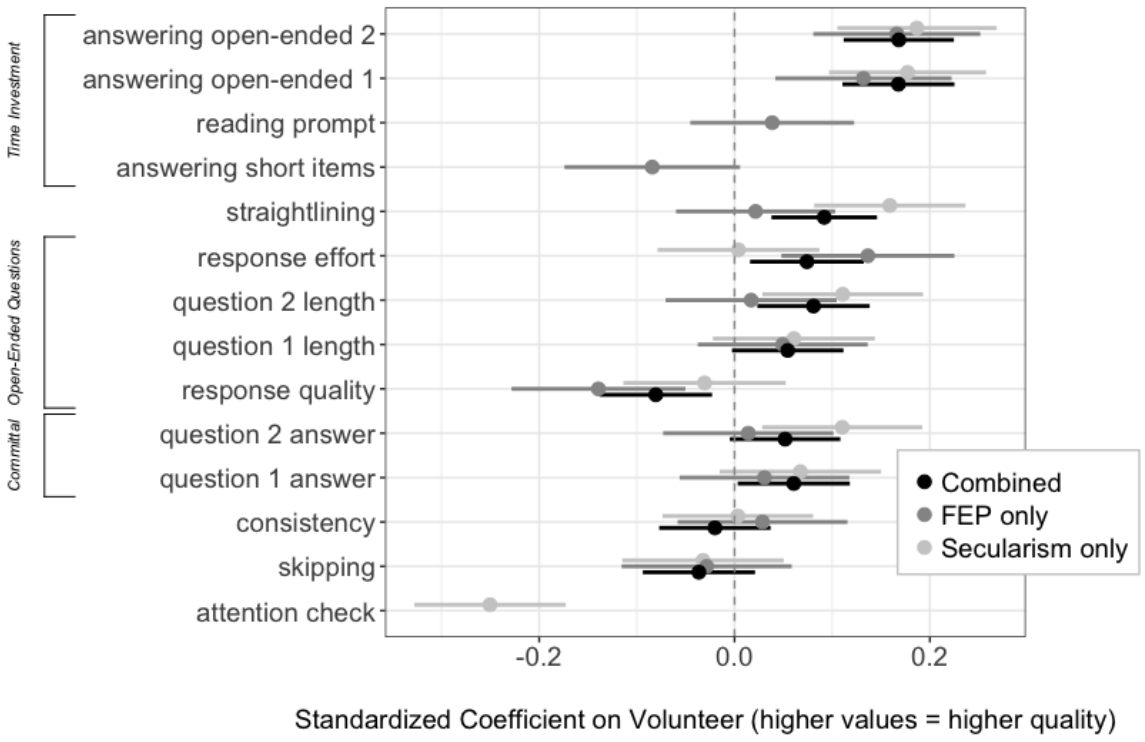
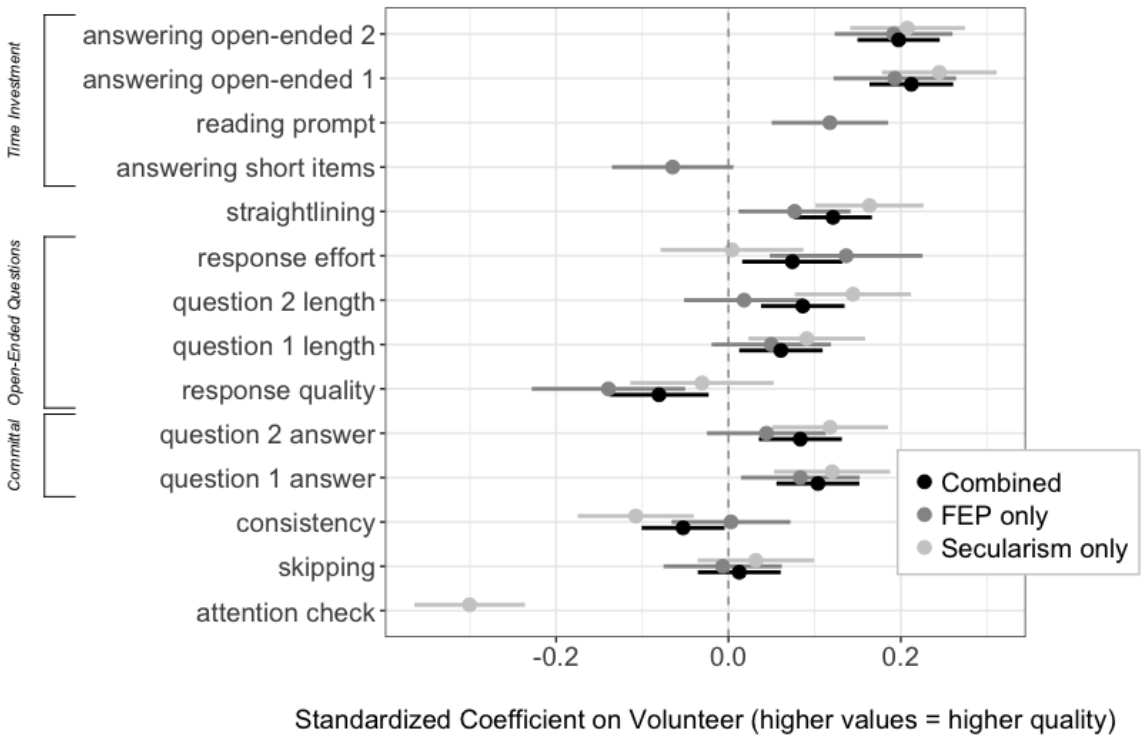


Figure A7: Standardized "Volunteer" Coefficients for All Response Quality Tests, by survey without covariates



by several short questions. We chose this format because reading prompts take more time than standard survey questions and subjects have no restrictions on how much time they need to spend reading the prompt. This test thus enables us to identify potential variation in time spent reading the prompt and answering questions, the latter of which are not available until a respondent confirms they have finished reading.

For the secularism questionnaire the prompt is an excerpt of a *New York Times* op-ed about the rise of secularism and atheism in American society. For the foreign economic policy study, respondents were invited to read an excerpt from *Foreign Policy Magazine* on Chinese economic statecraft. The prompts are structurally similar in terms of their number and length of paragraphs, their emphasis on providing objective facts on the topic, and their overall length. Each prompt is about 400 words. While we include a time investment test in both surveys, we also embed an attention check in the secularism survey, invalidating comparisons across the two survey instruments for time investment. Thus, we present results only for the FEP survey.

For our primary test of subjects' investment of time into their survey responses, we measure the number of seconds respondents spend reading the article prompt before clicking "next" to answer questions about it. We also measure the number of seconds spent responding to five short-answer or multiple choice questions about the article. Two additional tests of time investment are the number of seconds respondents spend answering two open-ended questions later in the survey. For all of our time investment tests, the dependent variable is the number of seconds spent before clicking to the next page. Positive coefficients would mean volunteers spend more time on the task, and we interpret more seconds spent as an indication of higher response quality. For this analysis, we trim the outer 5th percentile of time to eliminate extreme outliers.

Both with and without controls, volunteers spend more time than paid subjects on the

reading prompts, though the difference becomes statistically insignificant with controls.[1] Volunteers spent slightly less time responding to the block of questions about the article they had read. Volunteers spent more time responding to open-ended questions on both surveys than paid subjects, controlling for subject characteristics.

Straightlining: We examine whether paid and volunteer subjects have significantly different propensities to engage in straightlining. Straightlining is a well-known phenomenon in survey research in which respondents rush through a survey and provide the same response for many questions without actually reading and considering the question content.[2] With well-designed survey questions that naturally induce variation, less straightlining signals a higher quality response.

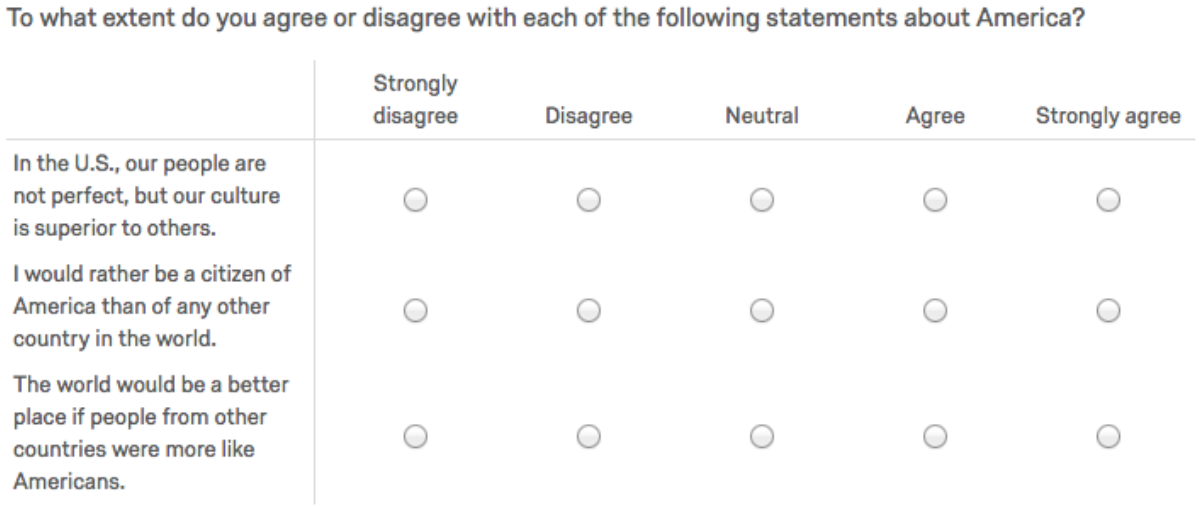
We design a test for straightlining that presents respondents with several matrix-style question blocks. This type of questioning is arguably especially vulnerable to straightlining. This is because several grid-style questions are presented on a single page, and answer choices to each question are located in close proximity to each other. Figure A8 offers an example of a question block from the foreign economic policy survey. We include seven of these blocks in each survey.

Question blocks within each survey vary in terms of their directionality, that is, the extent to which choosing the same answer for each question would reflect consistent, logical attitudes. Including bidirectionality in some of these question blocks allows us to detect straightlining in instances where a respondent with consistent preferences should not choose the same answer category. For example, in one of the question blocks in the foreign economic policy survey, we ask respondents whether they thought both "border walls" and "more

[1] For a random half of subjects, we also included a reminder to respondents to take their time reading, as they would not be permitted to click "back" to view the article. As yet, we have not tested whether volunteer versus paid subjects were influenced to spend more time reading given the presence of such a prompt, but in the future we will explore this.

[2] More complex forms of straight lining include patterned or random responses, which are considerably more difficult for investigators to detect.

Figure A8: Sample Question Matrix used in Straightlining Tests



open borders" are beneficial for American interests. In this case, vertical straightlining behavior, in which a respondent chooses the same preference for every question, would be strong evidence that a subject is rushing through a survey. While respondents see multiple blocks, some with bidirectional response items and others with unidirectional response items, our measure of straightlining is constructed including only those blocks with bidirectional responses, i.e., those on which a respondent paying attention would *not* be expected to straightline naturally.

Arguably the most straight-forward test of straightlining is to simply create a variable that measures whether a subject engaged in vertical straightlining—meaning he or she answered the same response category for every question—for each question block. This type of answer behavior, if detected, is perhaps the most egregious form of straightlining and thus represents a useful first step. Thus, we use a simple binary operationalization of whether a respondent engaged in vertical straightlining by offering a uniform response category for every item within a straightlining block. Since each of our surveys had three total bidirec-

tional grid-style question matrices, our dependent variable in this test is the proportion of these three questions in which the subject did offer a single uniform response category.

As depicted in Figure A7, the volunteer coefficient for both the economic policy and secularism surveys is statistically distinguishable from the null in the bivariate regressions, in the direction of a higher quality response. In a multivariate setup (Figure A6), volunteers were less likely to engage in straightlining in the secularism survey than paid respondents, but there was no significant difference between volunteer and paid respondents in the survey related to economic policy.

To explain this difference, we consider differences between the two substantive surveys' straightlining blocks. By merit of the two substantive topics' different emphases, the foreign economic policy's straightlining questions tended to rely more on complex political knowledge than those of the secularism survey, which instead placed greater emphasis on subjects' personal experiences and normative values. These distinctions suggest that further exploration around paid and volunteer distinctions related to complex political knowledge would be beneficial. Nevertheless, on the whole volunteers fare at least as well as paid subjects in a test of their propensity to engage in straightlining.

Open-Ended Investment: We next test the possibility that subjects motivated by different incentives vary in the quality of their open-ended survey answers. We include multiple open-ended response questions in each survey, which provide subjects with the opportunity to expand on their other answers, leave feedback for the research team, or otherwise write additional content. Researchers relying on open-ended response data may perceive higher quality responses as those which are longer and include more interesting content. Researchers might also value open-ended responses that provide suggestions for improving the study.

In the body of the manuscript and Figure A6, we present results on open-ended response quality based on a human-coded, subjective, composite quality score with several dimensions.

Two undergraduate students coded each open-ended response. They provided a subjective 1-5 ranking of the amount of effort they perceived the subjects invested in the item. They also provide binary 0-1 score for whether each response was long, topical, and complete. These last three dimensions of quality are aggregated into a subjective quality score. One of the authors mediated all human-coded responses on which there was a difference of more than 2 points (in the 5-point) effort scale or a disagreement on the binary score for any of the three quality dimensions. For both effort and the composite quality score, higher values represent higher quality responses.

We also test open-ended response quality by checking whether respondents significantly differ in the number of characters written in response to open-answer prompts. We argue that increases in this measure on average indicate a higher quality response or, at least, greater participant investment in the survey.

For all open-ended tests, we differentiate between open-ended questions left blank due to skipping (in which a subject viewed an open-ended question but did not write anything) and those left blank because the subject had attrited prior to viewing the open-ended question. We include skipped questions as zeros in our analysis, but exclude attrited respondents from the analysis.

Figure A6 provides mixed quantitative support for the idea that respondent motivation impacts responses to open-ended questions. On both open-ended questions, volunteers wrote more characters than paid respondents for the secularism survey, though not the foreign policy survey. For the human-coded scores, on average across the surveys, volunteers scored higher on effort, but lower on response quality overall.

Non-committal responses: Our surveys also include tests of commitment. In this context, commitment refers to subjects' willingness to signal intensity of attitudes in their reported responses by stating their intent to support (or oppose) a cause with behavior beyond simply reported survey responses. We employ multiple measures of commitment

in each survey. For example, in the secularization survey we ask whether respondents are willing to 1) sign a petition and 2) confront an individual about inappropriate conduct. Subjects can report their intent to partake or abstain from either behavior, or can choose a less committal answer such as "It depends" or "Unsure."

For our tests of commitment, we code a noncommittal answer as one in which a respondent does not express intent to participate or abstain from participation. Choosing "unsure" or "it depends" was taken as a lower-quality response, though we address ambiguity about interpreting these results in our discussion of findings below.

Results on these items were mixed. In the absence of demographic covariates, volunteers were somewhat less likely to choose the noncommittal response for all but the second committal item on the economic policy survey. However, after the inclusion of demographic covariates, these distinctions are no longer significant in many cases.

We note that a response of "unsure" can be interpreted in various ways with respect to response quality. Although one interpretation of quality is to expect higher-quality responses to include fewer non-committal responses, alternative interpretations are possible. For example, more unsure responses might be an expression that respondents are less likely to engage in cheap talk. Future research could work to examine this distinction in greater detail.

Inconsistency: We examine consistency across a subject's answers within a survey. In general we perceive higher consistency as a measure of higher response quality: if respondents report unstable or illogical opinions, preferences, or other responses in a short study, there is certainly reason to doubt whether such responses reflect actual attitudes (Achen 1975). Inconsistency alternatively may simply be a proxy for lower levels of attention paid by subjects to the study content, an equally worrisome possibility.

To explore the possibility that paid and unpaid subjects differ in their cognitive invested in the content of a study, we design a test that asks subjects the same question twice. The

first version of the question is embedded in one of the straightlining blocks discussed above. The second version is a the same question phrased differently and presented in a different format, in this case as a standalone multiple choice question later in the survey.[3]

We score this metric by creating a variable that measures whether there is directional consistency across the two questions, that is, if a respondent's reported preferences are in the same direction, if not to the same degree. To require exact equality in both direction and degree seemed to be a test so stringent that it was inconsistent with what the literature would anticipate as reasonably high-quality and consistent (Ansolabehere, Rodden, and Snyder 2008).

Figure A6 depicts the likelihood of volunteer versus paid subjects responding in a consistent direction on the two items. More consistency is seen as an indicator of higher-quality responses. For the most part, paid and unpaid subjects did not statistically differ on these items. As one exception, volunteer survey responses were arguably of lower quality in relation to response consistency on the secularism version of the survey in the bivariate framework. However, this distinction did not hold once control variables were introduced. Anecdotal evidence from the secularism survey's open-ended responses points to ambiguity in the question wording, which may have influenced subjects to provide inconsistent results. The negative result was driven by this survey item, and volunteers in the foreign policy survey fared no worse than their paid counterparts. Overall, volunteers were slightly less likely to offer a consistent response on one version of the survey, but this distinction went away with the inclusion of control demographic covariates.

Skipping: We design a simple test to detect a subject's propensity to skip questions. Because question skipping creates missing data, which can create bias if not corrected (King et al. 2001), subjects who skip fewer questions are typically more desirable than those who

[3]We randomize whether or not a respondent is first reminded that he or she has already been asked about this issue. We have not yet tested the impact of these randomized prompts.

skip more. We design this test by creating a question about individual's preferences for a certain policy. For example, in the secularization survey we ask a question related to churches' rights to engage in political activities, and for the foreign economic policy survey we ask about whether the government should encourage more free trade. For these questions respondents could choose, "I'm not sure. Skip." as an answer choice beyond the standard support-oppose scale. To measure skipping, we simply create a variable that receives the value of "1" if a respondent chose to skip the question.[4]

It is worth noting that the choice to skip in survey questions can be interpreted in multiple directions in relation to response quality. On the one hand, skipping may represent taking an "easy way out," wherein subjects avoid engaging with a cognitively challenging question or are simply rushing to finish and, thus, may represent low quality responses. On the other hand, skipping may genuinely represent uncertainty among subjects and may be more desirable than other response strategies, such as choosing an answer at random. For subjects who are aware that they are uninformed on a particular policy issue, skipping may represent a reasonable, high-quality choice. Future work could explore this distinction in greater detail. Results in this manuscript (such as in Figure A6) are presented such that more skipping indicates lower quality, consistent with our original hypothesis. There are no significant differences between paid and unpaid respondents in this test of skipping.

Attention Check: We also embed an attention check, or "screener" in our secularism survey instrument. As noted above, we task subjects with reading an article and responding to questions about that article. In order to ensure respondents were actually reading the article, not merely opening the prompt and then doing something else before proceeding, we included a sentence in the middle of the article that asked respondents to reply "yes" to an open-ended question in the question block following the article, rather than answering the

[4] While this variable is our primary measure of skipping, it is also worth noting that we do not force answers on the majority of questions in either questionnaire. As a result, we can also explore in future research whether paid subjects or volunteers are more likely to skip other types of questions.

question. We include this test only in the text of the article in the secularism survey.

Contrary to our hypothesis, a higher percentage of paid than unpaid subjects correctly addressed this attention check. This indicates that paid subjects may have been more carefully reading the survey. On the other hand, some of this may have come from familiarity with these checks, resulting from the frequency with which paid subjects participate in studies. Arguable, this finding helps alleviate the concern that volunteers become overly familiar with surveys to the extent that they operate as "professional" survey takers.

References

Achen, Christopher H. 1975. "Mass Political Attitudes and the Survey Response." *American Political Science Review* 69(04): 1218–1231.

Ansolabehere, Stephen, Jonathan Rodden, and James M Snyder. 2008. "The strength of issues: Using multiple measures to gauge preference stability, ideological constraint, and issue voting." *American Political Science Review* 102(2): 215–232.

Berinsky, Adam J, Gregory A Huber, and Gabriel S Lenz. 2012. "Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk." *Political Analysis* 20(3): 351–368.

Bonikowski, Bart, and Yueran Zhang. 2017. "Populism as Dog-Whistle Politics: Anti-Elite Discourse and Sentiments toward Minorities in the 2016 Presidential Election.".

Enos, Ryan D. 2017. *The Space Between Us: Social Geography and Politics*. New York: Cambridge University Press.

Enos, Ryan D., and Christopher Celaya. 2015. Segregation Directly Affects Human Perception and Intergroup Bias. In *American Political Science Association, Annual Meeting*. San Francisco: .

Enos, Ryan D., and Riley Carney. 2015. "Is Modern Racism Caused by Anti-Black Affect?: An Experimental Investigation of the Attitudes Measured by Modern Racism Scales.".

Hainmueller, Jens, and Michael J Hiscox. 2010. "Attitudes toward highly skilled and low-skilled immigration: Evidence from a survey experiment." *American Political Science Review* 104(01): 61–84.

Hankinson, Michael. 2017. "Do NIMBYs Think Outside of Their Neighborhood? Free-Riding and Fairness in Collective Action.".

Kam, Cindy D, and Elizabeth N Simas. 2010. "Risk orientations and policy frames." *The Journal of Politics* 72(02): 381–396.

Kaufman, Aaron. 2017. "An Automated Method to Estimate Bias in Survey Questions.".

Kaufman, Aaron, and Matthew Kim. 2017. "Sequential Blocked Randomization for Internet-Based Survey Experiments.".

Kaufman, Aaron, Gary King, and Mayya Komisarchik. 2018. "How to Measure Legislative District Compactness If You Only Know it When You See It.".

King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve. 2001. "Analyzing incomplete political science data: An alternative algorithm for multiple imputation." *American political science review* 95(1): 49–69.

Krosch, Amy R, and David M Amodio. 2014. "Economic scarcity alters the perception of race." *Proceedings of the National Academy of Sciences* 111(25): 9079–9084.

Krosch, Amy R, Leslie Berntsen, David M Amodio, John T Jost, and Jay J Van Bavel. 2013. "On the ideology of hypodescent: Political conservatism predicts categorization of racially ambiguous faces as Black." *Journal of Experimental Social Psychology* 49(6): 1196–1203.

Mahler, Daniel. 2016. "Do Altruistic Preferences Matter for Voting Outcomes?".

Rasinski, Kenneth A. 1989. "The effect of question wording on public support for government spending." *Public Opinion Quarterly* 53(3): 388–394.

Sears, David O., and Donald R. Kinder. 1971. "Racial tensions and voting in Los Angeles." In *Los Angeles: viability and prospects for metropolitan leadership*, ed. Werner Z Hirsch. New York: Praeger.

Tomz, Michael. 2007. "Domestic audience costs in international relations: An experimental approach." *International Organization* 61(04): 821–840.

Tversky, Amos, and Daniel Kahneman. 1981. "The framing of decisions and the psychology of choice." *Science* 211(4481): 453–458.