Online Volunteer Laboratories for Social Science Research[1]

Ryan D. Enos[2] Mark Hill[3] Austin M. Strange[4]

This draft: April 21, 2017

[1] Funding for this research and the Harvard Digital Laboratory for the Social Sciences (DLABSS) is provided by the Foundations of Human Behavior Initiative at Harvard University. DLABSS is a scientific program at the Harvard Institute for Quantitative Social Science. Thanks to Yodahe Heramo, Jess Li, Brandon Wang, and Bryant Yang for their contributions to the Lab. For helpful feedback on an earlier draft, we thank Mark Brandt, 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; renos@gov.harvard.edu

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

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

Abstract

We argue for volunteer subjects as a valid source of research subjects in the social sciences and describe the ability to create permanent panels of these subjects, with desirable properties, using online crowdsourcing platforms. We build on Amazon's Mechanical Turk and other online crowdsourcing survey tools by extending the innovation of these platforms to volunteer subjects. Volunteer social science laboratories impose little or no financial costs on researchers and, by relying on intrinsically motivated volunteers, may even avoid some of the challenges associated with paid crowd-sourcing. We discuss the merits and limitations of volunteer subject pools in the context of our experience with one of the first of these labs, the Harvard Digital Lab for the Social Sciences (DLABSS). To test the validity of online volunteer subject pools, we replicate several classic and recent experimental studies and compare the results to replications on other subject pools. Our results suggest that volunteer subject pools can provide high quality data for a diverse range of social science research. We encourage other researchers and institutions to coordinate in creating multiple, integrated volunteer laboratories that will make the gains from these labs widely accessible to social scientists.

1 Introduction

Social Science human subjects research has undergone a major transformation in the last decade. Enabled by the Internet, large web-based samples of paid subjects, including nationally-representative surveys, now dominate human subjects research. And, of course, as social science experiments have grown in popularity and influence (Davis and Holt 1993, Druckman et al. 2006; 2011, Falk and Heckman 2009), low-cost convenience samples, such as Amazon's Mechanical Turk (MTurk), have proliferated across psychology, economics, political science, and other social science disciplines (Buhrmester, Kwang and Gosling 2011, Horton, Rand and Zeckhauser 2011, Berinsky, Huber and Lenz 2012, Paolacci, Chandler and Ipeirotis 2010, Mason and Suri 2012, Santoso, Stein and Stevenson 2016).

These platforms have dramatically lowered the costs, both material and opportunity, of obtaining subjects, allowing for large samples to be recruited rapidly and inexpensively (Mullinix et al. 2015). Subjects drawn from these platforms have also been shown to behave largely similarly to nationally representative samples and convenience samples drawn offline, such as students (Berinsky, Huber and Lenz 2012, Clifford, Jewell and Waggoner 2015). Furthermore, crowdsourced subjects are generally more demographically diverse and representative than traditional convenience samples, including those from the Internet and U.S. college students (Buhrmester, Kwang and Gosling 2011, Huff and Tingley 2015). In addition, they are often comparable to national probability samples (Berinsky, Huber and Lenz 2012). Any observer of human subjects research in social science likely recognizes the dramatic shift in research that these samples have enabled. Yet, despite this shift, a crucial aspect of research remains unchanged from the traditional in-person laboratory or face-to-face survey: subjects are extrinsically rewarded for participation, usually by money, gift, or course credit.

But is material compensation the only way to attract high-quality subjects? In this manuscript we present the case for an underutilized type of research subject, also enabled by the ability to collect inexpensive, large, and diverse subject populations online: the volunteer. We introduce the concept of an online volunteer laboratory and show that a researcher using volunteer subjects can obtain similar experimental results to those obtained from paid representative and convenience samples. Moreover, we argue that volunteer subjects may have properties that make them superior to paid subjects, such as greater attention and less incentive to misrepresent themselves. We also argue that the ability to set up permanent online pools for volunteers means that a large number of subjects can be obtained at little cost to the researcher. Furthermore, because "digital lab" researchers recruit and maintain their own sample, these labs provide control and flexibility not found on other platforms and the labs can become a public good, supporting a broad research agenda.[1]

We illustrate these points by introducing a volunteer lab designed to meet the needs of a broad community of researchers, in a similar manner to a traditional offline shared laboratory: the Harvard Digital Lab for the Social Sciences (DLABSS).[2] In introducing our lab, we address several important questions: How do online volunteer labs recruit subjects? How much and what type of research can these labs support? What are the associated costs of creating and operating a volunteer lab? How do lab subjects compare to samples from other commonly used subject pools? Are volunteer digital labs able to replicate research findings from more traditional platforms, and do they outperform other platforms on certain metrics?

The remainder of this paper is organized as follows. First, we discuss the potential limitations and advantages of volunteer labs for social science. Second, we introduce DLABSS and, in doing so, outline several specific potential advantages of volunteer labs. Third, we test the validity of DLABSS as a social science research tool. We do this by comparing the subject properties of DLABSS with well known online and offline survey pools. We also use volunteer respondents to replicate classic and recent social science experiments on DLABSS. Finally, we summarize our findings and offer concluding thoughts on how institutions and researchers can benefit from digital volunteer social science labs. We ultimately argue that volunteer labs are a highly valuable, underexploited resource for experimental social science. Volunteer labs can build on the innovation of MTurk and other online crowdsourcing survey tools by further lowering the barriers to high quality research subjects.

2 Advantages and Disadvantages of Volunteer Research Subjects

We distinguish between volunteer and non-volunteer research subjects. Volunteer subjects are motivated by intrinsic rewards for participating in research, including psychological benefits from helping others, a sense of duty, or a way to pass the time. Non-volunteer subjects are motivated by extrinsic rewards, such as financial or other non-token material

[1] We interchangeably use terms such as "digital lab," "volunteer digital social science laboratory," "volunteer digital lab," and "volunteer online lab."

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

compensation, or class credit.[3] Some compensation is so small, that the subjects are veritably volunteers,[4] and, for our purposes, we consider such subjects to be volunteers. If subjects do not know they are participating in research, as is the case in some field experiments and other forms of research, these subjects are not included in these definitions and are outside the scope of this article.

Although volunteer subjects were once fairly common in quantitative social science,[5] outside of some small scale surveys, volunteer subjects now only rarely appear in published quantitative studies.[6] While the use of volunteers in past published research shows that people will willingly take part in scientific research simply for the intrinsic rewards, paying subjects allows researchers to collect samples quickly and to maintain stable subject pools, such as those found in campus laboratories or on standing online panels. The number of people willing volunteer for research or the frequency with which the average person will volunteer may be limited, thus making it difficult to use volunteers to do the large N studies that characterize modern survey and much experimental research. While scholars have been able to obtain large samples by providing non-material rewards to their subjects, such as the opportunity to learn something about their own psychology,[7] improving or testing their cognitive function,[8] or being entertained by games, we are among the first to show that a large number of subjects can be obtained, easily and consistently, purely by the intrinsic rewards of volunteering for science.

Furthermore, while any sample of subjects may be atypical or non-representative in consequential ways and the properties of research populations from low-wage markets, like Mturk, are occasionally criticized (e.g., Stewart et al. (2015)), volunteer subjects may be acutely unrepresentative in ways that raise serious concerns about external validity. They may be more motivated by academic interest or intellectual stimulation than the general public (Rosenthal and Rosnow 1975, Kagel, Battalio and Walker 1979), leading to unknown issues in responses. More generally, in the same way that college students are argued to have a properties that are atypical of the general population, leading to a biased view of human nature, "good subject" affects, and other issues of non-representativeness (Sears 1986, Jones

[3] Outside of social science, for example in medical research, subjects are often referred to as "volunteers," but, of course, such subjects are are often motivated by improving their own medical condition, which does not make them volunteer by our definition.

[4] For example, Klar and Krupnikov (2016) rewarded subjects with stickers.

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

[6] In ethnographic and other qualitative research, subjects are often not compensated.

[7] For example, online labs such as Project Implicit (`https://implicit.harvard.edu/implicit/`) provide tailored feedback to their respondents.

[8] Test My Brain `https://testmybrain.org/`.

2010), we might worry that people who will voluntarily give up their time (and repeatedly, as we show below) are so unusual that results based on volunteers should be viewed with extreme skepticism.

On the other hand, their is reason to believe there are advantages to using volunteer subjects. Of course, volunteer subjects lower the cost to researchers, but, moreover, when the primary motivation of subjects is financial and the primary motivation of researchers is high quality data, the incentives of researchers and subjects may not be aligned. In some cases, for example when subjects are working for an hourly rate as a major source of income (see Williamson (2016)) and, therefore, have an incentive to rush through studies, this mismatch between incentives may be large. Indeed, Mason and Watts (2010) found a negative relationship between the level of financial compensation and performance on MTurk: while larger financial incentives increase the quantity of respondents, quality decreases.

Because they rely on the intrinsic, rather than extrinsic motivation of subjects, volunteer samples may be less vulnerable to these problems. Social scientists have long recognized that extrinsic and intrinsic motivations can be at odds. The Motivation Crowding Effect posits that the introduction of non-zero financial awards for doing something crowds out intrinsic motivation and can actually decrease efficiency. Perhaps the most well known example of this is the argument that paying for blood reduces peoples' incentive to donate it (Titmuss 1970) and more recent surveys of social science literature find strong evidence, across a number of fields, for the tradeoff between financial compensation and intrinsic motivation (Frey and Jegen 2001). As such, we might worry that paying subjects for participating in research actually reduces the quality of responses we would receive from the same subjects if they were volunteers. For example, in one 1978 study testing for differences between paid and volunteer subjects, volunteer respondents committed fewer errors on a selection attention task (Rush, Phillips and Panek 1978).

In summary, there are at least two primary concerns with research using volunteer subjects: the ability to adequately and consistently collect large samples and the potentially unrepresentative nature of the subjects. Conversely, there are at least two potential advantages over paid subjects: the ability collect data at much lower costs than paid labs, and the possibility of obtaining data that is more high quality on certain dimensions. In the remainder of this manuscript, we put these competing propositions to the test. As we explain below, the power of the Internet has largely rendered concerns about the ability to gather adequately large samples moot and has also brought other advantages, including the ability to quickly gather diverse samples of volunteers and to maintain panels of these subjects. Second, by replicating both canonical and recent scholarship in social science, we demonstrate that, not only does the behavior of volunteer subjects largely mirror that of paid subjects, but the demographics of these subjects can be largely representative of the general population on observable characteristics. Finally, we show that volunteer subject responses are arguably of higher quality than paid subjects for certain tests.

3 DLABSS: An Online Volunteer Laboratory

When volunteers subjects are recruited online and brought into a permanent subject pool, the very properties that make paid online samples attractive—speed and cost effectiveness—are also available with volunteer subjects. In addition, some of the pitfalls of research with paid low-wage markets can potentially be avoided or mitigated.

A volunteer digital lab is a platform to host studies that can be completed on the World Wide Web. Subjects are unpaid volunteers who are recruited using a variety of online sources. If contact information, such as an email address, is collected, subjects can enter a standing subject pool and be called on to participate whenever a study is hosted on the lab. This feature can allow for repeated testing of the same subjects in a panel study. To be a laboratory, rather than a specific data collection tool, this platform should be modular and adaptable to a range of research projects, in the same way that a traditional offline psychology or economics laboratory can be used for diverse research agendas.

In July 2014 we launched DLABSS with the objective of attracting volunteer subjects to take part in online social science research at Harvard. DLABSS was created with the principle that volunteer online subjects should be a public good to researchers with diverse substantive interests.[9] Online laboratories for social science research are not new. Over the years many researchers have built websites that rely on volunteer subjects, such as Project Implicit (see Banaji and Greenwald (2013)). These volunteer-based projects have been successful in recruiting large numbers (in some cases, millions) of subjects through a robust web presence and well-known research agendas. However, these platforms are often confined to a certain research programs and therefore are not usable for a broad set of researchers. Additionally, these sites have taken considerable time and resources to build their web presence—recreating this presence for each individual research agenda would be extremely inefficient. DLABSS aims to make the use of volunteer subjects virtually costless

[9] 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), Pilny et al. (2016).

to individual researchers. Harvard faculty, graduate and undergraduate students conducting social science research have equal access to the lab. While DLABSS is currently limited to Harvard researchers, there is no reason that other digital laboratories could not be used for a community of researchers across institutions.

DLABSS is staffed part-time by several 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. The web design is simple. It has a clean, attractive web interface, which includes information on currently active experiments, researcher profiles, the DLABSS blog, and frequently asked question (FAQ)'s. Using Qualtrics, the lab collects basic demographic and attitudinal data on all subjects in the lab.

This simple, modular structure of the lab allows researchers to collect response data for virtually any type of social science research question that can be studied online. Since its inception, DLABSS has hosted over 70 social science studies, 41 of which were still actively collecting subjects as of April 2017: 35% of these were faculty projects, 60% were graduate students, and 5% were undergraduate students. Prior to beginning operation, DLABSS secured approval from our Institutional Review 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, making the lab usable by a wide variety of disciplines and, as such, studies have come from business, economics, sociology, political science, public policy, psychology, and other disciplines. Studies have included a range of target audiences, from undergraduate thesis projects to work intended for publication in scholarly outlets. As we explain below, many studies on DLABSS have subsequently been replicated on MTurk or other populations, and the results from the different platforms published together as complements or robustness checks.

The length of studies on DLABSS is largely similar to those found on MTurk, with most taking under 10 minutes, but the lab has successfully hosted studies of almost 20 minutes in length (see Figure 2)—a length that also highlights the advantages of volunteer subject pools because for many researchers, a study of that length would be prohibitively expensive survey on MTurk. Furthermore, as we describe below, studies on DLABSS have successfully included complex interactions with the subjects, such as follow-up phone call surveys.

As of April 2017, the active DLABSS volunteer pool includes over 8,800 subjects. This means that since its inception, DLABSS has attracted about 9.0 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.[10]

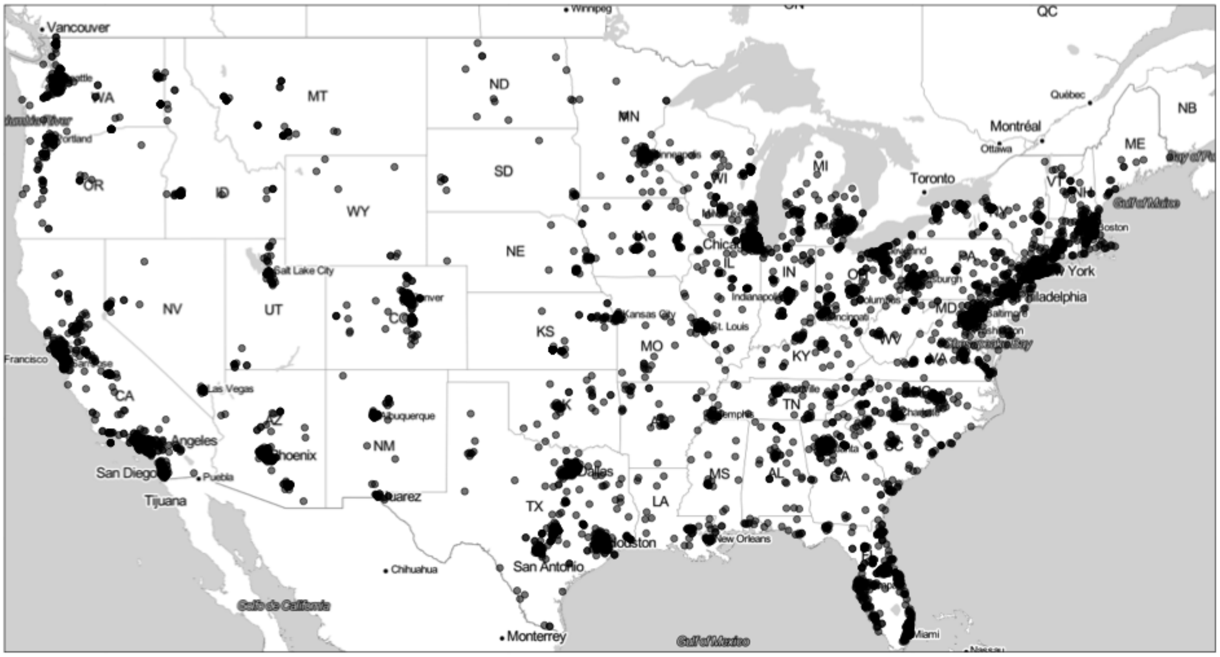
DLABSS staff email a fraction of the active subject pool each week to invite them to participate in the latest studies, and specific studies are often used to recruit new subjects to the lab's pool. As such, subjects for a given study usually consist of two types: 1) new subjects that enter an experiment after being recruited from the web and 2) subjects from our existing subject pool who enter an experiment after being directly solicited by email. Subjects are emailed biweekly on average. These email solicitations can allow a researcher to shape her subject pool, if desired, by targeting certain populations based on known demographics. In this manner, DLABSS shares this targeting ability with other opt-in internet panels, but maintains the low-cost structure of crowd-sourced research. Once a study becomes active on DLABSS, it can usually obtain over 600 responses within 2 weeks, while sometimes exceeding 1,000 responses. As the lab's subject pool continues to grow, so does the typical number of responses.

DLABSS subjects can come from anywhere in the world with Internet access, although 88.9% are in the United States. 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 most successful ways to recruit new subjects. Recruitment usually appeals to the subjects' interest in the substance of the study or their desire to help scientific research. The ability to target certain groups has allowed DLABSS to maintain a balanced subject pool. For example, when we noted that our subject pool was younger than the general population, we placed an advertisement on a bulletin for retired people and, in order to achieve ideological balance, we targeted conservative organizations. Using Craigslist we place advertisements in a wide range of communities to seek demographic, geographic, and ideological balance.

Figures 1 illustrate how the ability to reach subjects via the internet and to target certain populations can lead to a geographically diverse subject pool. For example, within the United States, the DLABSS community has participants from all 50 states and the District

[10] A figure illustrating growth and other information on DLABSS can be found at the DLABSS website: http://dlabss.harvard.edu/.

Figure 1: Geographic Distribution of DLABSS Volunteer Subjects in US



A total of 7,760 points are plotted on this map, each one representing a single DLABSS participant.

of Columbia, and a total of 1,929 cities and towns across the country are represented (we display the world-wide distribution in Figure A1 in the Appendix).

4 Potential Benefits of Digital Volunteer Labs

While other low-cost platforms have obviously been a tremendous boon to human subjects research, volunteer digital labs may have advantages that make them attractive complements or substitutes, not only for MTurk, but also for other subject pools.

Financial: Other than the cost of hosting a website and recruiting subjects, which can be minimal and shared, volunteer digital labs impose no monetary costs on researchers. Because volunteers enter a subject pool and often take part in many studies over time, the costs associated with a single subject are only incurred once. The savings from this model can be substantial. As a benchmark, if a researcher uses MTurk to recruit 1,000 subjects at the Federal minimum wage of $7.25 per hour for a 10 minute study, the resulting cost would be over $1,200 for subjects alone, in addition to at least a 40 percent surcharge by

Amazon.com.[11] The resulting total cost is non-trivial for many social scientists, especially compared to the costs of studies on volunteer labs. For example, for a study we report later un this manuscript, we spent $840 to administer a very simple 10-minute experiment on MTurk with 800 respondents. We conducted an identical study on DLABSS with a marginal cost of $0.

On DLABSS, across all studies, we have collected a total of 41,305 responses as of April 2017. In that time period, our costs associated with subject recruitment have been less than $10,000, resulting in a cost of less than $0.25 per response.[12] The average length of surveys on DLABSS was 7.79 minutes. The recruitment of an equivalent number of subjects for this length of survey on MTurk, assuming minimum wage payments, would exceed $40,000. Notably, nearly all of the costs associated with DLABSS subject recruitment have been in wages paid to student research assistants, so a researcher able to compensate students with course credit or other means could lower costs to virtually nothing. Additionally, overtime the use of our lab has spread around the university, resulting in an increasing number of studies being fielded, so that the cost per response continues to fall.

From an institutional perspective, the moderate cost of the public good of a digital lab can represent substantial savings. Using the costs of an MTurk survey estimated above, if researchers at a single institution conducted 100 studies on MTurk per year (a number that seems like a realistic minimum in an institution with 100 human subjects researchers), then the total institutional cost of MTurk will be over $100,000 per year. And of course, MTurk, to it's credit, has dramatically lowered costs compared to other platforms commonly used for survey research, such as SSI and YouGov, so the savings realized by a volunteer lab can be even more dramatic when compared to other sources of human subjects.

Subject and Researcher Accessibility: In addition to financial accessibility, volunteer labs also help increase accessibility for citizens as well as social scientists. From the perspective of potential volunteers, there are very few barriers to entry into volunteer social science labs. As such, individuals throughout the world can join a digital social science research community, possibly extending participation and voice in social science to a wider range of

[11] The 40% figure is based on the baseline MTurk fee of 20% plus an additional 20% for tasks with more than 10 workers, see https://requester.mturk.com/pricing. As another benchmark, we administered a survey to 116 Harvard social scientists and found that, on average, researchers recruited roughly 290 respondents per study on MTurk and paid them just over $4 each.

[12] These numbers are imprecise because much of this cost goes toward wages for staff with multiple duties. Notably, as an early adopter in this space, our startup costs, including costs associated with research, will likely be higher than subsequent similar labs.

communities. In the case of DLABSS, a volunteer must be 18 years old and possess a valid email address. For the researcher, in addition to the lowered barriers to entry provided by monetary savings, the modular nature of the laboratory means it is easily accessible to a range of research programs and researchers do not face the sometimes considerable barrier of setting up a platform specific for a single project. As social science has moved toward "larger scale, collaborative, interdisciplinary" research (King 2014), this modularity is increasingly valuable.

Stable Unit Treatment Value Assumption and other subject to subject interference: In prominent theories of causal inference, the Stable Unit Treatment Value Assumption (SUTVA) is a critical assumption for valid inference, stating that an individual's potential outcomes depend only on her own treatment assignment (Rubin 1974). Crowdsourced markets where communication cannot be controlled by the researcher raise the potential for violations of this assumption. For example, MTurk workers can digitally interface on MTurk Forum,[13] Mturkgrind,[14] and MTurk Crowd,[15], and often do so.[16] On these forums, MTurk workers who have completed studies commonly comment on the content of studies, thereby potentially influencing the behavior of other subjects before the subjects have enrolled and been assigned to treatment.[17] These platforms also raise related concerns about subject crosstalk biases (Edlund et al. 2009) caused by inter-subject interactions (Paolacci, Chandler and Ipeirotis 2010).

In contrast, the potential for violations of SUTVA may be reduced in volunteer labs because subject interactions via online discussion forums are less likely, simply because such fora do not exist and, perhaps more importantly, individual volunteers have little incentive to create such platforms given that they select into the lab for non-financial reasons.

Observation Uniqueness and Effective Sample Size: A common concern of online labor markets is that biases will result from subjects participating multiple times or in multiple related experiments (Chandler, Mueller and Paolacci 2014). This concern is related to similar critiques that the universe of social science MTurk research is plagued by surprisingly

[13] http://mturkforum.com/

[14] http://www.mturkgrind.com/

[15] http://www.mturkcrowd.com/

[16] As of December 2016, MTurk forum had over 58,000 members and 11,000 discussions. Hitsworthturking for had over 35,000 members. MTurk Grind had over 12,000 members with nearly 30,000 discussions. MTurk Crowd had nearly 3,000 visitors and over 1,000 discussions.

[17] Indeed, in conducting the MTurk studies for this manuscript, we checked MTurk Forum, MTurk Crowd, and HitsWorthTurkingFor and found postings about our studies.

small effective sample sizes (Stewart et al. 2015). Of course, platforms use terms-of-use agreements, membership fees, or software requirements to deter subjects from creating multiple accounts (Horton, Rand and Zeckhauser 2011). But these solutions, by imposing high costs on the subjects, creates tension with these markets' ability to provide cheap and diverse subjects to researchers and, for some subjects, the incentives for participation are high enough that they may try to overcome such barriers dishonestly. Of course, because there is little incentive to do so, problems of multiple accounts and repeat users should be potentially less likely to occur with volunteer laboratories and as volunteer pools grow, then the problem of a single small effective sample being shared by so many researchers is mitigated.

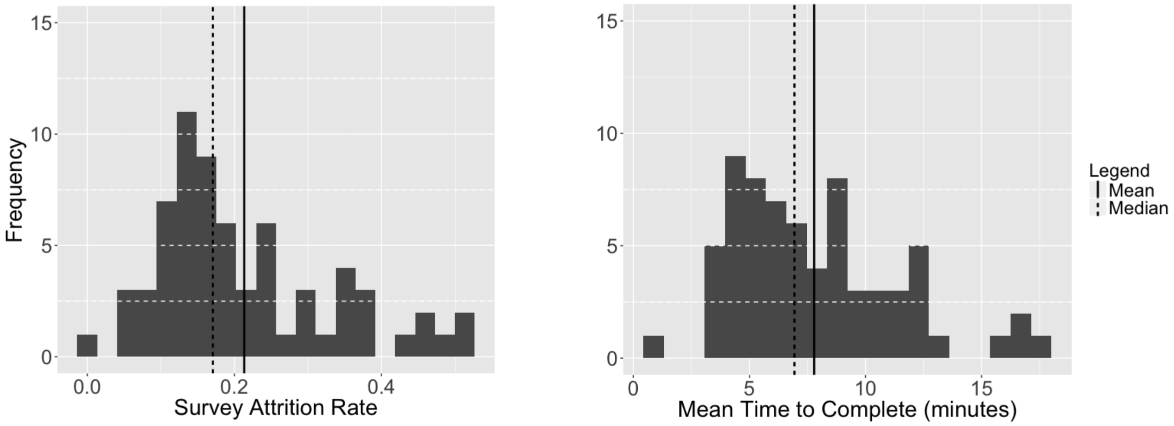
Attrition: In experimental research, attrition rates might be uneven and cause selection bias if the nature of one or more treatment conditions causes subjects to drop out at higher rates than the other. If, for example, treatment conditions unevenly increases survey time, financially-motivated subjects may unevenly leave the survey across conditions. As such, as others have pointed out, selective attrition appears to be an important issue that threatens inferences made from MTurk studies (Zhou et al. 2016).

In contrast, because of the intrinsic motivations of subjects, in volunteer labs selective attrition may be less severe. As illustrated in Figure 2, the mean and median attrition rate for DLABSS studies is 20.9% and 17.3%, respectively (these numbers drop considerably with surveys under 10 minutes in length). This represents an important improvement over attrition in MTurk, where researchers have noted the potential for alarming attrition-induced bias: in the study by Zhou et al. (2016) *every* replication of published studies from MTurk had an attrition rate of greater than 20%.

5 Validity of Volunteer Labs as a Research Tool

This section describes the results of several assessments of the validity of a volunteer lab. First, we compare basic subject properties across DLABSS and other well-known online and offline survey resources: MTurk, the American National Election Studies (ANES), and the U.S. Census Bureau's Current Population Survey (CPS) (King et al. 2010). Our initial tests build off Berinsky, Huber and Lenz (2012), which is often cited to justify the use of MTurk in social science research. We do this because MTurk is an extremely common source of low-cost convenience-samples in the social science and, therefore, the platform that volunteer digital labs are most likely to complement or replace (see also Santoso, Stein and Stevenson (2016)).

Figure 2: 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 all studies hosted by DLABSS.

We then use volunteer subjects from DLABSS to replicate several classic and contemporary social science experiments. These include those replicated by Berinsky, Huber and Lenz (2012): the well-known Asian Disease Problem (Tversky and Kahneman 1981), the framing effect on welfare spending (Rasinski 1989), and a recent experiment on framing and risk (Kam and Simas 2010). In addition, we replicate two prominent, recent experimental studies in political science: one on the drivers of immigration attitudes (Hainmueller and Hiscox 2010), and the other on the individual-level foundations of audience costs in international relations (Tomz 2007). We also briefly discuss other studies that researchers have carried out on DLABSS and replicated on other platforms. We then test our hypotheses that volunteer subjects may be more attentive and deliver higher quality responses than paid subjects by replicating part of a recent study on anxiety and immigration attitudes that draws heavily on open-ended survey responses (Gadarian and Albertson 2014). Finally, we discuss examples of "complex studies" successfully undertaken on DLABSS and which we believe might be more difficult to undertake on a platform like MTurk.

5.1 Comparing Subjects on DLABSS and Other Subject Pools

Table 1 compares the demographics, political behavior, and political knowledge of DLABSS respondents with a sample of MTurk participants from December 2016,[18] the online 2008-09 American National Elections Study Panel Study (ANESP), the 2012 American National

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

Elections Study (ANES), and the 2012 Current Population Survey (CPS), the latter two often considered the gold-standard for nationally-representative face to face samples. The latter four samples replicate the comparisons in Berinsky, Huber and Lenz (2012).

Broadly speaking, DLABSS respondents appear quite similar to those from both online and offline survey platforms.[19] On average, the DLABSS sample of respondents is demographically similar to MTurk, but with some important differences. For example, compared to MTurk, DLABSS appears more representative of the general population in terms of age. There are certain demographic characteristics that one might expect to be correlated with the leisure time necessary for volunteering, such as education, income, and race. However, we note that compared to MTurk, DLABSS participants are not substantially more educated, higher income, or likely to be white. And, compared to the nationally representative samples, while DLABSS subjects tend to be more educated, they also tend to have lower income.

Also in Table 1, we compare DLABSS to the other samples on common measures of ideology and political participation, sophistication, and knowledge. Although, like MTurk, the DLABSS sample expresses a higher level of political interest than the nationally representative samples, this does not appear to translate into a sample with skewed political knowledge: DLABSS more closely resembles the nationally representative surveys than does the MTurk sample, demonstrating the added usefulness of DLABSS for political science studies, where such variables are crucial to prominent theories in political behavior (e.g., Zaller (1992).

In Tables A1 and A2 in the Appendix, we complete the replication of Berinsky, Huber and Lenz (2012) by comparing policy support of DLABSS respondents with those found in the other four surveys and by comparing DLABSS demographics with small in-person convenience samples. Results from DLABSS once again approximate these other samples in a way similar to MTurk.

5.2 Using DLABSS to Replicate Existing Findings in the Social Sciences

Having established that volunteer samples can look broadly similar to nationally representative and MTurk samples, we further test the properties of the DLABSS panel by directly replicating a series of experiments and by describing other replications that have

[19] The sample reported here is sometimes a subset of the entire volunteer population of DLABSS because, while we collect demographics on all subjects, to a subset, we administered a questionnaire with the specific demographic and knowledge questions necessary for this comparison.

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	55.6% (0.7)	48.0% (1.9)	57.6% (0.9)	51.9% (0.2)	52.0% (0.1)
Education (mean years)	15.2 (0.0)	14.9 (0.1)	16.2 (0.1)	13.4 (0.0)	13.6 (0.1)
Age (mean years)	44.1 (0.2)	37.8 (0.5)	49.7 (0.3)	46.7 (0.1)	47.3 (0.4)
Mean income	$49,174 ($565)	$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	74.7 (0.5)	78.3 (1.6)	83.0 (0.7)	79.5 (0.1)	74.5 (1.0)
Black	6.4 (0.3)	8.4 (1.0)	8.9 (0.7)	12.2 (0.1)	12.2 (0.7)
Hispanic	7.8 (0.4)	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	21.1 (0.6)	21.5 (1.6)	16.9 (0.7)	18.2 (0.1)	18.2 (0.8)
Midwest	21.9 (0.6)	25.4 (1.7)	28.3 (0.9)	21.6 (0.1)	22.6 (0.9)
South	31.8 (0.7)	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	46.9 (0.6)	44.3 (1.9)			
Independent/Other	29.6 (0.6)	30.1 (1.7)			
Republican	23.5 (0.5)	22.8 (1.6)			
Ideology					
Liberal	59.1 (0.6)	62.6 (1.9)			
Conservative	34.4 (0.6)	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	909-6,280	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.*

been carried out on DLABSS. In many respects, this is the crucial test of volunteer labs because convenience samples, such as MTurk and DLABSS, are overwhelmingly used for experimental tests, where the representativeness of the sample is usually a secondary concern to internal validity. We summarize these replications in Table 2, describe the replications below, and provide detailed results, where available, either in the manuscript or Appendix. Using volunteers subjects, DLABSS has replicated, at least, 15 studies, six of which were replications we conducted to test the properties of DLABSS. The replications had the same approximate magnitude, directionality, and level of statistical significance as in the original study The fifteen studies include replications of studies using a range of other platforms, including MTurk or more expensive samples, including General Social Survey (GSS), Knowledge Networks (KN), Qualtrics, and the Danish firm Epinion. These studies also cover classic and more recent research and a variety of academic disciplines and substantive topics.

5.2.1 Replicating the Studies in Berinsky, Huber and Lenz (2012)

We begin by showing that all the same three experiments replicated by Berinsky, Huber and Lenz (2012) were replicable on DLABSS.

First, Rasinski (1989) using data from the GSS found that framing policy choices dramatically changes stated individual preferences for the policies. Specifically, citizens are much more likely to support redistribution when it is worded as "assistance to the poor" compared to "welfare." Our replication is in line with the original findings and previous MTurk replications. All experiments show a significant difference between the "poor" and "welfare" groups, with the poor group always more likely than the welfare group to suggest that too little is being spent. The DLABSS replication poor group matched the original experiment's poor group exactly, while the DLABSS welfare group was higher than both of the other platforms. It is, of course, possible that shifts in public opinion since the original experiment may partially explain this difference. See Table A3 in the Appendix.

Second, we replicate the well-known Asian Disease Problem popularized by Tversky and Kahneman (1981), who presented two different groups of undergraduate students with the problem of a "rare Asian disease" threatening their country and suggest two possible programs to deal with the disease. They find that respondents primed with a frame that describes policy options in terms of losses (deaths), rather than gains (lives saved) are more likely to choose probabilistic (rather than certain) outcomes. Across all platforms, people in the positively framed group prefer the certain outcome to the probabilistic outcome, and vice versa for the negatively framed group. The original experiment displays the strongest

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 (2017)	survey bias	272	✓	524		
Kaufman and Kim (2017)	Audience costs	360	✓	429		
Kaufman, King and Komisarchik (2017)	district compactness	373	✓	764		

*Each of these studies was successfully replicated on DLABSS in that the replicated result had the same approximate magnitude, directionality, and level of statistical significance as in the original study. Column 1 lists the name of the original study. The first six replications listed were carried out by the authors to test the properties of DLABSS. The next nine were replicated by other authors in the course of research carried out using DLABSS. For citations to the papers in which the replications appeared, see the text. Some replications and the original finding are carried out in the same study, for example Enos and Carney (2015) did the same study on both DLABSS and MTurk. In several cases, the original article contains more than one study and not all were attempted to be replicated. 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. So, for example, Tversky and Kahneman (1981) was first studied using undergraduates, but has been replicated on many samples since then. 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, in all other cases, it was carried out by other researchers.*

difference in groups, while the online laboratories find smaller but equally significant effects. See Table A4 in the Appendix.

Lastly, we look at the more recent experiment by Kam and Simas (2010), who use KN to demonstrate that individuals' risk acceptance affects their preferences for different policies. Specifically, those who are willing to accept higher amounts of risk are more likely to support probabilistic policy outcomes. The coefficient magnitudes and levels of significance are very similar across platforms. See Table A5 in the Appendix.

5.2.2 Replicating Other Recent Prominent Studies

We further explore the ability of volunteer digital labs to be a useful research tool by revisiting two recent political science experiments that have generated broad scholarly attention. These studies are far removed from basic psychological research and, rather, explore complex sociopolitical attitudes.

First, we replicate Tomz (2007) well-known experiment on the microfoundations of audience costs in international relations. Tomz uses a survey panel from KN to demonstrate that audience costs can be found in an experimental setting: citizens are more likely to disapprove of their president when he or she makes an empty threat versus not making a threat at all. We highlight this experiment to show the broad usefulness of volunteer subjects: first, the experiment comes from the field of international relations, a field that, in our assessment, has received less attention than others when discussing crowd-sourced samples; and, second, the experiment is arguably more qualitatively complex than other the other experiments we have presented. In particular, though we only present the core findings in line with Tomz (2007), the number of varying factors in the experiement—such as the political nature of the adversary, the military strength of the adversary, and the origins of the dispute—is greater than in some of the other experiments we replicated. Table 3 present the core results of our replication and shows that results from the DLABSS volunteer panel essentially mirror the original results from Tomz (2007).

Second, we replicate Hainmueller and Hiscox (2010)'s experiment that challenges traditional political economy theories about preferences for and against immigration. Their results, drawn from an experiment embedded in a KN sample, suggest that high-skilled immigrants who will benefit the national economy as a whole are universally preferred by citizens regardless of income status, education level and other covariates, lending support to the idea of sociotropic preferences for immigration. For robustness, we replicated this experiment on both DLABSS and a new MTurk sample. In Figure A2 in the Appendix we

Table 3: 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.

show that we replicate the key finding in Hainmueller and Hiscox (2010) on both DLABSS and MTurk. As in the original study, respondents in DLABSS were far more supportive of highly-skilled immigration than low-skilled immigration: over 60% of respondents supported allowing more highly-skilled immigrants into the U.S., while only about 40% were supportive of allowing in more low-skilled immigrants.

5.2.3 Other Replications on DLABSS

In addition to the above replications, which we performed to explicitly test the validity of volunteer digital labs, other researchers have used DLABSS to replicate many other findings. We briefly mention these replications to provide a broader sense of what type of research volunteer labs have been used to reproduce. We include studies based on our own communications with DLABSS researchers. This may be an incomplete list.

Several studies hosted on DLABSS have explored racial politics. One study (Enos 2017), using both DLABSS an 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 variant 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 2017). 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).

5.3 Testing Open-Ended Survey Response Quality between MTurk and DLABSS

We now turn to demonstrating that volunteer samples not only can replicate findings from non-volunteer samples, but that volunteer samples, due to the intrinsic motivation of subjects, may sometimes have increased response quality relative to other experimental platforms. To do this, we turn to open-ended survey responses and response time.

First, in the top panel of Table 4, we examine the response times of subjects on DLABSS and MTurk to the political knowledge questions listed in Table 1. Across every question, DLABSS respondents take more time before initially answering and before advancing to the next question. All else equal, this may be evidence that DLABSS respondents are more carefully considering their responses. This is some, very preliminary, evidence that data collected from volunteer subjects, perhaps because they are not under pressure of maximize their hourly wage, is less prone to certain types of measurement error. We note too that this careful consideration by subjects does not appear to make volunteer subjects so out of the ordinary that studies of short-term, heuristic, mental processing (see Kahneman (2003)), which is central to many research agendas in psychology and other fields cannot be undertaken: DLABSS successfully replicated the Tversky and Kahneman (1981) framing experiment, the paradigmatic example of heuristic processing, as well as other experiments relying on similar mental processing.

Another simple metric of response quality is the amount of effort expended answering open-ended questions where subjects are not paid an hourly rate and therefore not incentivized to move quickly through a survey. On both MTurk and DLABSS, we partially replicated a recent political science finding that individuals primed with anxiety will seek and assess information about immigration in different ways (Gadarian and Albertson 2014).[20] MTurk respondents were paid $.75 to complete the study. We examined whether DLABSS respondents wrote more or less than their paid MTurk counterparts. As the bottom panel in Table 4 illustrates, across three of the four open-ended response questions included in Gadarian and Albertson (2014), DLABSS participants on average wrote 15% more than

[20] The replication was partial in that some of the dependent variables in Gadarian and Albertson (2014) involve capturing the proportion of stories read by respondents, which we did not measure.

Table 4: Comparing Response Length in DLABSS and MTurk

Mean number of seconds spent answering political knowledge questions

Question	DLABSS	MTurk	Difference
Presidential succession after Vice President	19.54 (.36)	15.40 (.42)	4.14***
House vote % needed to override a veto	14.41 (.35)	8.65 (.30)	5.76***
Maximum number of presidential terms	15.14 (4.79)	8.77 (.24)	6.37
Length of a U.S. Senate term	13.52 (.35)	9.80 (.34)	3.72***
Number of Senators per state	8.77 (.27)	8.30 (.34)	0.47
Length of a U.S. House term	12.52 (.33)	10.85 (.37)	1.67***

Mean number of words written per open-ended question in Gadarian and Albertson (2014)

Question	DLABSS	MTurk	Difference
Treatment	22.88 (2.03)	17.90 (0.98)	4.97*
Control	24.22 (4.77)	20.01 (1.05)	4.21
Memory of Stories	22.78 (1.31)	25.24 (0.90)	-2.47
Opinion of Stories	32.17 (1.80)	23.69 (0.73)	8.48***

*Standard errors in parentheses. * indicates a t-test with a p-value below .05. ** below .01. *** below .001.*

respondents on MTurk.

5.4 Complex Studies on DLABSS

Finally, we highlight the potential advantage of volunteer subjects in that subjects driven by intrinsic motivation may be especially willing to participate in unconventional surveys that require more work by subjects than typical studies. We do not formally test whether this is more true of volunteer subjects than others, however, here we describe two complex experiments that DLABSS has supported. Our experience makes us believe such tests would be difficult to execute on some paid samples, such as MTurk.

One experiment required participants to begin a conventional online survey, then eventually asked them to call a phone number and answer questions via an automated phone message. The attrition rate was abnormally high for this study (above 50%), but the researcher was able to collect hundreds of responses. Another study included two parts in which respondents completed a series of questions then, after a certain amount of time, were subsequently emailed to complete more questions. The treatment and control conditions were differentiated by the amount of time to elapse before receiving an email invitation to complete the second step. This study could also prove difficult in an online paid lab setting because recontacting subjects through email is prohibited by terms of service not allowing the collection of personally identifiable information.[21]

6 Conclusion

Human subjects research in social science has become dramatically less expensive and more accessible following the emergence of online crowdsourcing tools like MTurk. In this paper we have introduced the concept of volunteer digital labs as a social science research tool and argued that these labs can be a high-quality tool. Building on the innovation of online crowdsourcing, we argue that volunteer labs extend the accessibility and quality frontiers of online experiments even further and that volunteer labs have enormous potential as vehicles to lower the costs and improve the quality of social science inquiry for researchers and the institutions that support them. We have demonstrated this by replicating a series of diverse studies using a volunteer laboratory and by highlighting potential advantages of volunteer subjects over non-volunteer subjects.

Of course, even if volunteer labs are cheaper and prove more effective at avoiding sampling and other biases, we believe that volunteer labs will most likely emerge as complements, rather than replacements, to other existing survey platforms. In the same way that paid crowdsourcing, such as MTurk, has not meant the demise of other subject pools, we believe that volunteer labs will have advantages and disadvantages that will allow them occupy a particular niche in human subjects data collection.

While we have started our laboratory to aid researchers at a single university, coordination between researchers and across institutions would significantly increase the efficiency of volunteer labs, in much the same way that economies of scale in survey research, found in designs like the Cooperative Congressional Election Study (CCES) and Cooperative Campaign Analysis Project (CCAP) have greatly improved the efficiency of survey research. We urge researchers to undertake such collaborations.

[21] See https://requester.mturk.com/help/

References

Banaji, Mahzarin and Anthony Greenwald. 2013. *Blindspot: Hidden Biases in Good People.* New York: Delacorte Press.

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." Working Paper, Harvard University.

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.

Chandler, Jesse, Pam Mueller and Gabriele Paolacci. 2014. "Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers." *Behavior research methods* 46(1):112–130.

Clifford, Scott, Ryan M Jewell and Philip D Waggoner. 2015. "Are samples drawn from Mechanical Turk valid for research on political ideology?" *Research & Politics* 2(4):2053168015622072.

Davis, Douglas D and Charles A Holt. 1993. *Experimental economics.* Princeton university press.

Druckman, James N, Donald P Green, James H Kuklinski and Arthur Lupia. 2006. "The growth and development of experimental research in political science." *American Political Science Review* 100(04):627–635.

Druckman, James N, Donald P Green, James H Kuklinski and Arthur Lupia. 2011. *Cambridge handbook of experimental political science.* Cambridge University Press.

Edlund, John E, Brad J Sagarin, John J Skowronski, Sara J Johnson and Joseph Kutter. 2009. "Whatever happens in the laboratory stays in the laboratory: The prevalence and prevention of participant crosstalk." *Personality and Social Psychology Bulletin* .

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." Midwest Political Science Association, annual meeting, Chicago.

Falk, Armin and James J Heckman. 2009. "Lab experiments are a major source of knowledge in the social sciences." *science* 326(5952):535–538.

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.

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." Working Paper, Harvard University.

Horton, John J, David G Rand and Richard J Zeckhauser. 2011. "The online laboratory: Conducting experiments in a real labor market." *Experimental Economics* 14(3):399–425.

Huff, Connor and Dustin Tingley. 2015. "Who are these people? Evaluating the demographic characteristics and political preferences of MTurk survey respondents." *Research & Politics* 2(3):2053168015604648.

Jones, Dan. 2010. "A WEIRD View of Human Nature Skews Psychologists' Studies." *Science* 328(5986):1627–1627.

Kagel, John H, Raymond C Battalio and James M Walker. 1979. Volunteer artifacts in experiments in economics: Specification of the problem and some initial data from a small-scale field experiment. Technical report The Field Experiments Website.

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

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." Working Paper, Harvard University.

Kaufman, Aaron, Gary King and Mayya Komisarchik. 2017. "How to Measure Legislative District Compactness If You Only Know it When You See It." Working Paper, Harvard University.

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

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

King, Miriam, Steven Ruggles, J Trent Alexander, Sarah Flood, Katie Genadek, Matthew B Schroeder, Brandon Trampe and Rebecca Vick. 2010. "Integrated public use microdata series, current population survey: Version 3.0.[machine-readable database]." *Minneapolis: University of Minnesota* 20.

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 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?" Working Paper, University of Copenhagnen.

Mason, Winter and Duncan J Watts. 2010. "Financial incentives and the performance of crowds." *ACM SigKDD Explorations Newsletter* 11(2):100–108.

Mason, Winter and Siddharth Suri. 2012. "Conducting behavioral research on Amazons Mechanical Turk." *Behavior research methods* 44(1):1–23.

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.

Pilny, Andy, Brian Keegan, Brooke Wells, Chris Riedl, David Lazer, Jason Radford, Katya Ognyanova, Leslie DeChurch, Michael Macy, Noshir Contractor et al. 2016. Designing online experiments: Citizen science approaches to research. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*. ACM pp. 498–502.

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.”.

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

Rubin, Donald B. 1974. “Estimating causal effects of treatments in randomized and non-randomized studies.” *Journal of educational Psychology* 66(5):688.

Rush, Michael C, James S Phillips and Paul E Panek. 1978. “Subject recruitment bias: The paid volunteer subject.” *Perceptual and Motor Skills* 47(2):443–449.

Santoso, Lie Philip, Robert Stein and Randy Stevenson. 2016. “Survey Experiments with Google Consumer Surveys: Promise and Pitfalls for Academic Research in Social Science.” *Political Analysis* 24(3):356–373.

Sears, David O. 1986. “College Sophomores in the Laboratory: Influences of a Narrow Data Base on Social Psychology’s View of Human Nature.” *Journal of Personality and Social Psychology* 51(3):515–530.

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.

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.

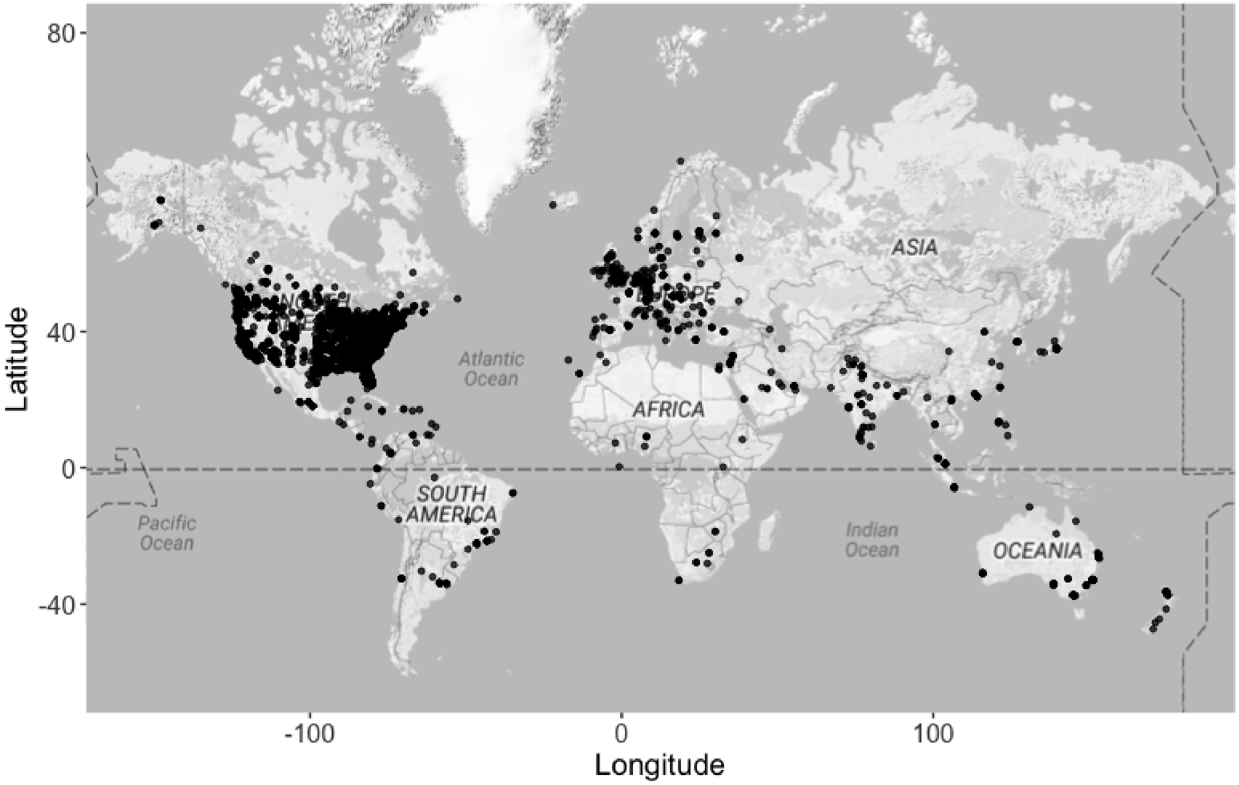
Tversky, Amos and Daniel Kahneman. 1981. “The framing of decisions and the psychology of choice.” *Science* 211(4481):453–458.

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* .

Figure A1: World-wide Distribution of DLABSS Volunteer Subjects



A total of 8,736 points are plotted on this map, each one representing a single DLABSS participant.

Appendix

This appendix includes various supplementary figures and tables referenced in the main text of the manuscript.

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).

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.*

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

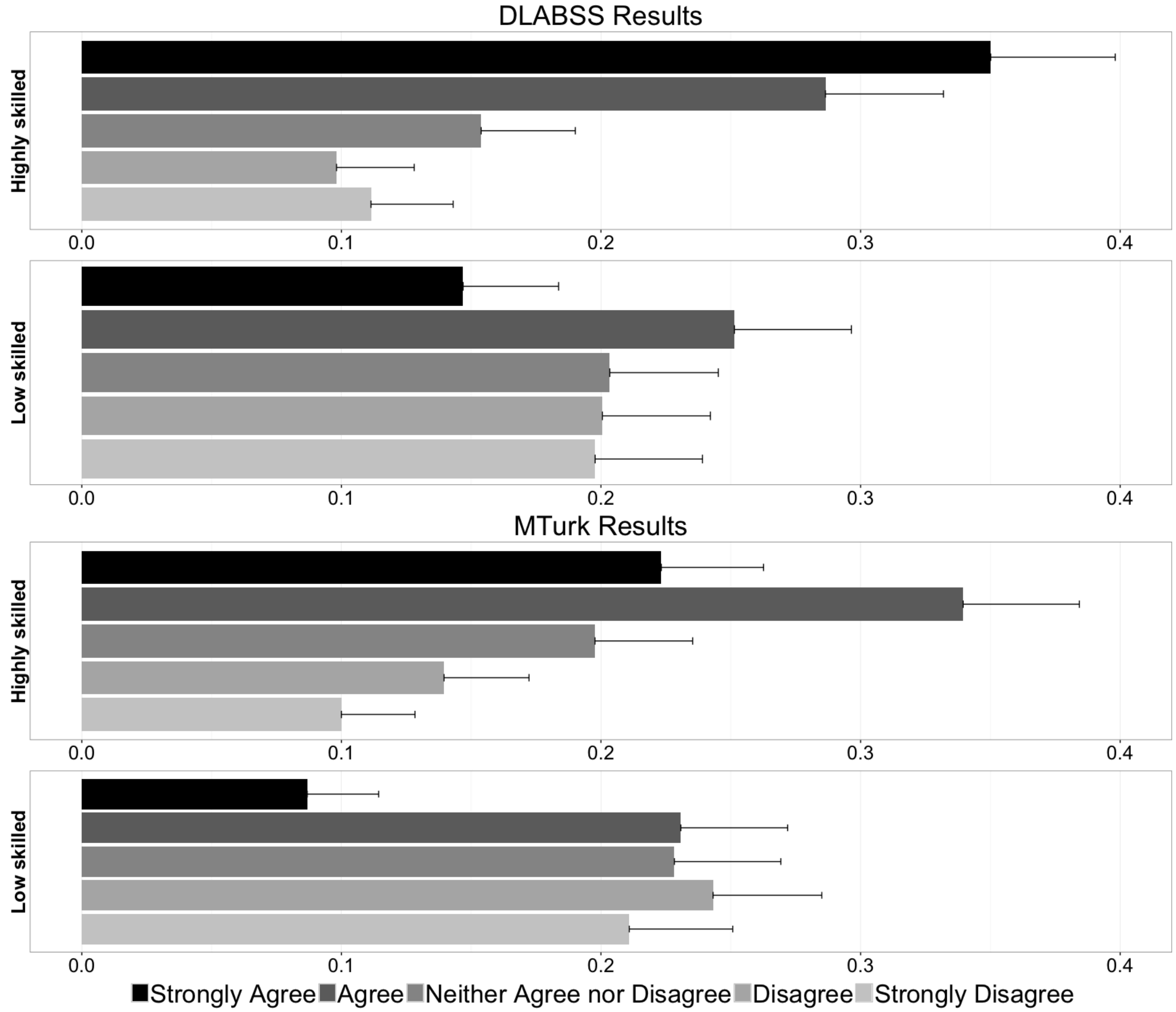
Cells are percent of respondents choosing non-probabilistic (certain) outcome with each frame.

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.

Figure A2: 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)?"