Advance Access publication March 2, 2012

Political Analysis (2012) 20:351–368
doi:10.1093/pan/mpr057

Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk

Adam J. Berinsky
Department of Political Science, Massachusetts Institute of Technology, Cambridge, MA 02139
e-mail: berinsky@mit.edu (corresponding author)

Gregory A. Huber
Institution for Social and Policy Studies, Yale University, New Haven, CT 06511
e-mail: gregory.huber@yale.edu

Gabriel S. Lenz
Department of Political Science, University of California, Berkeley, Berkeley, CA 94720
e-mail: glenz@berkeley.edu

Edited by R. Michael Alvarez

We examine the trade-offs associated with using Amazon.com's Mechanical Turk (MTurk) interface for subject recruitment. We first describe MTurk and its promise as a vehicle for performing low-cost and easy-to-field experiments. We then assess the internal and external validity of experiments performed using MTurk, employing a framework that can be used to evaluate other subject pools. We first investigate the characteristics of samples drawn from the MTurk population. We show that respondents recruited in this manner are often more representative of the U.S. population than in-person convenience samples—the modal sample in published experimental political science—but less representative than subjects in Internet-based panels or national probability samples. Finally, we replicate important published experimental work using MTurk samples.

1 Introduction

Interest in experimental research has increased substantially in political science.[1] But experiments can be time consuming and costly to implement, particularly when they involve nonstudent adult subjects. Amazon.com's Mechanical Turk (MTurk) has the potential to facilitate low-cost experiments in political science with a diverse subject pool.

MTurk is an online Web-based platform for recruiting and paying subjects to perform tasks. Relative to other experimental pools, MTurk is extremely inexpensive both in terms of the cost of subjects and the time required to implement studies. Not surprisingly, scholars across the social sciences have begun using MTurk to recruit research subjects.[2] However, despite this burgeoning line of research, the benefits and potential limitations of using MTurk for subject recruitment in political science research remain relatively unexplored. (For related evaluations in psychology and in economics, see Buhrmester, Kwang, and Gosling 2011 and Horton, Rand, and Zeckhauser 2010.)[3] This article addresses a simple but important question: Is MTurk a valid tool for conducting experimental research in political science?[4]

Authors' note: Supplementary data for this article are available on the *Political Analysis* Web site.

[1] Replication code and data are available at *Political Analysis* Dataverse (Berinsky, Huber, and Lenz 2011).

[2] As of October, 2011, Google Scholar lists 769 social sciences articles with the phrase "Mechanical Turk." Relevant studies by economists include, for example, Chandler and Kapelner (2010), Chen and Horton (2010), Horton and Chilton (2010), and Paolacci et al. (2010). Computer scientists have also tested MTurk's suitability as a source of data for training machine learning algorithms (e.g., Sheng et al. 2008; Sorokin and Forsyth 2008). For example, Snow et al. (2008) assessed the quality of MTurkers' responses to several classic human language problems, finding that the quality was no worse than the expert data that most researchers use.

[3] Analyses have generally found that experiments on Internet samples yield results similar to traditional samples. Based on a comprehensive analysis, for example, Gosling et al. (2004) conclude that Internet samples tend to be diverse, are not adversely affected by nonserious or habitual responders, and produce findings consistent with traditional methods.

[4] The MTurk platform is of course limited to conducting research that does not require physical interactions between the subject and either the researcher or other subjects (e.g., to gather DNA samples, administer physical interventions, or observe face-to-face interactions among subjects).

We present a framework for evaluating subject pools in general and then apply the framework to the MTurk subject pool. Although the particular object of study here is the contemporary MTurk community, the types of analysis we undertake can be used to evaluate the strengths and limitations of other subject recruitment procedures. We first identify threats to the internal and external validity of research conducted using MTurk. Even accounting for these threats, we argue that MTurk is a valuable subject recruitment tool. First, the demographic characteristics of domestic MTurk users are more representative and diverse than the corresponding student and convenience samples typically used in experimental political science studies. Second, we replicate experimental studies previously conducted using convenience and nationally representative samples, finding that the estimates of average treatment effects are similar in the MTurk and original samples. Third, we find that potential limitations to using MTurk to recruit subjects and conduct research—in particular, concerns about heterogeneous treatment effects, subject attentiveness, and the prevalence of habitual survey takers—are not large problems in practice.

The remainder of the article proceeds as follows. We begin by providing an overview of the subject recruitment and data-gathering choices involved in using MTurk. We then lay out a framework to evaluate the threats to internal and external validity that arise from the particular characteristics of a given subject pool. The bulk of our article concerns an evaluation of the MTurk pool according to these standards. To do this, we first describe results from a series of surveys measuring MTurk subjects' demographic and political characteristics. Next, we compare our MTurk sample to: (1) those samples used in experiments published in leading political science journals, (2) a high-quality, Internet panel sample, and (3) probability samples used in the Current Population Survey (CPS) and the American National Election Studies (ANES). We then demonstrate that the effects of experimental manipulations observed in the MTurk population comport well with those conducted in other samples. We conclude by addressing two concerns raised about online samples: whether the MTurk population is dominated by subjects who participate in numerous experiments (or participate more than once in a given experiment) and whether MTurk subjects are effectively engaged with the survey stimuli. These concerns, we conclude, are relatively modest in the context of the MTurk platform.

2 Recruiting Experimental Subjects Using MTurk

A core problem for experimental researchers in political science is the difficulty and prohibitive cost of recruiting subjects. In recent years, important innovations, such as the NSF-funded Time-Sharing Experiments in the Social Sciences project, have enabled broader access to nationally representative samples, but access to these resources remains limited.

Amazon.com's MTurk is a promising alternative vehicle for experimental subject recruitment. Amazon.com markets MTurk as a means to recruit individuals to undertake tasks. In practice, these tasks involve a wide array of jobs requiring human intelligence, such as classifying pictures or transcribing handwriting, but tasks can also include taking surveys with embedded experimental manipulations.

To initiate a survey using MTurk, a researcher (a ''Requester'' in Amazon's vernacular) establishes an account (http://www.mturk.com), places funds into her account, and then posts a ''job listing'' using the MTurk Web interface that describes the Human Intelligence Task (HIT) to be completed and the compensation to be paid (Amazon assesses Requesters a 10% surcharge on all payments). Each HIT has a designated number of tasks and the requester can specify how many times an individual MTurk ''Worker'' can undertake the task. Researchers can also set requirements for subjects, including country of residence and prior ''approval rate,'' which is the percent of prior HITs submitted by the respondent that were subsequently accepted by Requesters. When MTurk Workers who meet these eligibility requirements log onto their account, they can review the list of HITs available to them and choose to undertake any task for which they are eligible.[5]

The MTurk interface gives the researcher a great deal of flexibility to conduct a study. In addition to using MTurk's embedded workspace to set up simple tasks, the researcher can also refer subjects to an external Web site. For instance, subjects might be redirected to a Web page to take a survey with an embedded experimental manipulation.[6] Additionally, outside Web sites make it easy to obtain informed consent, implement additional screening procedures, debrief after an experiment, and collect detailed information about the survey process

[5]We present screen shots of a sample HIT from a Worker's view in the Supplementary data.

[6]We have successfully used commercial Web sites like SurveyGizmo and Qualtrics for this process, and any Web survey service that can produce a unique worker code should be suitable. Providing subjects with a unique code and having them enter it in the MTurk Web site ensures that they have completed the task.

(including response times for items and respondents' location when taking the survey as determined on the basis of the respondents' Internet Protocol [IP] address). The final stage for the researcher is compensating subjects. The researcher can easily authorize payment for the task through the MTurk Web interface.[7]

2.1 *Cost per Subject and Ease of Recruitment*

Each HIT advertisement shows the amount a worker will be paid to complete it. Additionally, the description can list how long the task will take and/or other features (e.g., ''fun'' or ''easy'') that may make the task more attractive to MTurk workers. In practice, we have found that relative to other experimental pools, recruiting subjects using MTurk is extremely inexpensive.

In particular, a listing of different studies we have undertaken, including advertised length, payment, and number of completions per day, appears in Table 1. As that table makes clear, for short surveys advertised as taking between 2 and 4 min, we have been able to obtain more than 200 subjects per day when paying as little as $.25 per survey. When payments were lower (e.g., $.15 per HIT), recruitment was somewhat slower. At higher pay rates—between $.50 and $.75 per completion—we have been able to recruit over 300 subjects per day.[8]

To put these costs in perspective, even the highest pay rate we have used on MTurk of $.50 for a 5-min survey (an effective hourly rate of $6.00) is still associated with a per-respondent cost of $.55 (including Amazon.com's 10% surcharge) or $.11 per survey minute. By contrast, per subject costs for typical undergraduate samples are about $5–10, for nonstudent campus samples about $30 (Kam, Wilking, and Zechmeister 2007), and for temporary agency subjects between $15 and $20. Outside the campus

Table 1 Task title, compensation, and speed of completion for selected MTurk studies

Task title	*Date launched*	*Number of subjects*	*Pay per subject*	*Mean minutes per subject*	*Completions per day*								
					1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Answer a survey about current affairs and your beliefs	January 5, 2010	490	$0.15	7	116	64	41	40	27	36	15	11	15
2- to 3-Min survey for political science research	March 16, 2010	500	$0.25	2	210	68	37	55	53	64	18		
4-Min survey for political science research	April 26, 2010	500	$0.40	4	298	105	79	18					
3-Min survey for political science research	April 29, 2010	200	$0.25	1	200								
3- to 4-Min survey for political science research	May 17, 2010	150	$0.45	2	150								
7- to 9-Min survey	June 24, 2010	400	$0.75	6	400								
5- to 7-Min survey	June 28, 2010	400	$0.75	5	321	79							
5- to 7-Min survey	July 3, 2010	400	$0.50	3	256	115	29						
2- to 3-Min survey	July 16, 2010	200	$0.25	3	200								

Note. The remaining subjects for the January 5, 2010, study were recruited as follows: Day 10 (11), Day 11 (10), Day 12 (17), Day 13 (22), Day 14 (29), and Day 15 (36).

[7]If the researcher has arranged for the external Web site to produce a unique identifier, she can then use these identifiers to reject poor quality work on the MTurk Web site. For example, if the experiment included mandatory filter questions or questions designed to verify the subject was reading instructions, the worker's compensation can be made contingent on responses. Finally, a unique identifier also allows the researcher to pay subjects a bonus based on their performance using either the MTurk Web interface or Amazon.com's Application Programming Interface (API). A Python script we have developed and tested to automate the process of paying individual bonuses appears in the Supplementary data.

[8]We are unaware of research using the MTurk interface to recruit large numbers of subjects for longer surveys, although Buhrmester et al. (2011) report being able to recruit about five subjects per hour for a survey advertised as taking 30 min for a $.02 payment. Other scholars have reported that higher pay increases the speed at which subjects are recruited but does not affect accuracy (Buhrmester, Kwang, and Gosling 2011; Mason and Watts 2009; but see Downs et al. 2010 and Kittur, Chi, and Suh 2008 on potentially unmotivated subjects, a topic addressed in greater detail below).

setting, private survey firms we have worked with charge at least $10 per subject for a 5-min survey when respondents are drawn from an Internet panel. MTurk is, in short, extremely inexpensive relative to nearly every alternative other than uncompensated students.[9]

3 Assessing the Validity of Research Conducted Using MTurk

MTurk provides researchers with access to inexpensive samples. But are these samples of sufficient quality for political science research? To answer this question, we need to consider the validity of experiments conducted through MTurk using the standards that should be applied to evaluate any subject pool.

Threats to the validity of experimental research are generally divided into questions of internal validity and external validity. External validity is an assessment of whether the causal estimates deduced from experimental research would persist in other settings and with other samples. Internal validity pertains to the question of whether causal estimates appropriately reflect the effects of the experimental manipulation among the participants in the original setting.

Concerns about the external validity of research conducted using student samples have been debated extensively (Sears 1986; Druckman and Kam 2011). For experimental research conducted using MTurk, two concerns raised about student samples are pertinent: (1) Whether estimated (average) treatment effects are *accurate* assessments of treatment effects for other samples and (2) whether these estimates are *reliable* assessments of treatment effects for the same sample outside the MTurk setting. The former concern is most likely to be a threat if treatment effects are heterogeneous and the composition of the MTurk sample is unrepresentative of the target population (see Druckman and Kam 2011). For example, if treatment effects are smaller for younger individuals than older ones, a sample dominated by younger individuals will yield estimated treatment effects smaller than what one would observe with a representative sample. The latter concern arises if people behave differently in the MTurk setting than they do outside of that setting.

Given our particular interest in assessing the validity of MTurk samples relative to other convenience samples, we undertake three types of analysis to address potential threats to validity. To address the concerns about *external validity*, we first compare the characteristics of MTurk samples to other samples used in political science research. Second, we use MTurk to replicate prior experiments performed using other samples to compare estimated treatment effects. The replication exercise also allows us to, in part, address the other concern about generalizability—whether MTurk subjects behave differently than similar subjects in other research settings. Third, we consider whether MTurk samples are dominated by habitual participants whose behavior might be unrepresentative of similar populations not exposed to frequent political surveys.[10]

[9] Another promise of MTurk is as an inexpensive tool for conducting panel studies. Panel studies offer several potential advantages. For example, recent research in political science on the rate at which treatment effects decay (Chong and Druckman 2010; Gerber, Gimpel, Green, and Shaw 2011) has led to concerns that survey experiments may overstate the effects of manipulations relative to what one would observe over longer periods of time. For this reason, scholars are interested in mechanisms for exposing respondents to experimental manipulations and then measuring treatment effects over the long term. Panels also allow researchers to conduct pretreatment surveys and then administer a treatment distant from that initial measurement (allowing time to serve as a substitute for a distracter task). Another potential use of a panel study is to screen a large population and then to select from that initial pool of respondents a subset who better match desired sample characteristics. The MTurk interface provides a mechanism for performing these sorts of panel studies. To conduct a panel survey, the researcher first fields a task as described above. Next, the researcher posts a new task on the MTurk workspace. We recommend that this task be clearly labeled as open only to prior research participants. Finally, the researcher notifies those workers she wishes to perform the new task of its availability. We have written and tested a customizable Perl script that does just this (see the Supplementary data). In particular, after it is edited to work with the researcher's MTurk account and to describe the new task, it interacts with the Amazon.com API to send messages through the MTurk interface to each invited worker. As with any other task, workers can be directed to an external Web site and asked to submit a code to receive payment. Our initial experiences with using MTurk to perform panel studies are positive. In one study, respondents were offered 25 cents for a 3-min follow-up survey conducted 8 days after a first-wave survey. Two reminders were sent. Within 5 days, 68% of the original respondents took the follow-up. In a second study, respondents were offered 50 cents for a 3-min follow-up survey conducted 1–3 months after a first-wave interview. Within 8 days, almost 60% of the original respondents took the follow-up. Consistent with our findings, Buhrmester, Kwang, and Gosling (2011) report a two-wave panel study, conducted 3 weeks apart, also achieving a 60% response rate. They paid respondents 50 cents for the first wave and 50 cents for the second. Analysis of our two studies suggests that the demographic profile does not change significantly in the follow-up survey. Based on these results, we see no obstacle to oversampling demographic or other groups in follow-up surveys, which could allow researchers to study specific groups or improve the representativeness of samples.

[10] It should be noted that other convenience samples, such as student or local intercept samples, may also have significant numbers of habitual experimental participants. However, it is important to determine whether this is especially a problem in the MTurk sample, where subjects can easily participate in experiments from their home or work computers.

Turning to the *internal validity* of estimates derived from MTurk samples, two concerns are especially pertinent. The first is the possibility that subjects violate treatment assignment by participating in a given task more than once. The second is subject inattentiveness, in which case some subsets of the sample do not attend to the experimental stimuli and are effectively not treated. To explore these concerns, we directly assess subject attentiveness and the apparent prevalence of subjects participating in the same MTurk HIT multiple times from different user accounts.

4 Assessing Threats to External Validity

We begin our investigation of external validity by assessing the nature of the MTurk subject pool. Specifically, we compare measured characteristics of MTurk survey participants to characteristics of participants in three distinct types of research samples: convenience samples used in experiments published in leading political science journals, a sample generated by a high-quality Internet panel, and probability samples of U.S. residents.

We surveyed 551 MTurk workers in February and March of 2010. We advertised the survey as taking about 10 min and paid respondents 50 cents each.[11] Because we wish to benchmark MTurk against samples of adult U.S. citizens, we restricted the survey to individuals MTurk classified as 18 or older and living in the United States.[12] We also excluded individuals with approval rates below 95% on previous MTurk tasks. As an additional check on U.S. residency, we verified that respondents took the survey from U.S. IP addresses and excluded the 32 individuals (5.8%) who did not.[13]

4.1 *Comparison of Respondent Characteristics: Local Convenience Samples*

Local convenience samples are the modal means of subject recruitment among recent published survey and laboratory experimental research in political science. We examined all issues of the *American Political Science Review*, the *American Journal of Political Science*, and the *Journal of Politics* from January 2005 to June 2010. Of the 961 articles in these issues, 51 used experimental data. Forty-four of these articles used U.S. subjects exclusively (the complete list of these articles, as well as a summary of subject recruitment methods, appears in the Supplementary data). Of these 44 articles, more than half used convenience samples for subjects (including student samples, local intercept samples, or temporary agencies).

Table 2 compares our MTurk sample to several convenience samples for a series of measures reported in prior work. After presenting selected demographics and partisanship of our MTurk sample, it displays the average characteristics of the student and adult samples collected by Kam, Wilking, and Zechmeister (2007). Next, the table lists characteristics of two adult convenience samples used in Berinsky and Kinder (2006)—one of the handful of articles that describes the characteristics of its convenience samples. One of these samples is from around Princeton, NJ, and the other is from around Ann Arbor, MI.

On demographic representativeness, the MTurk sample fares well in comparison with these convenience samples.[14] Not surprisingly, relative to an average student sample, the MTurk population is substantially older, but it is younger than any of the three nonstudent adult samples. The MTurk and student samples are similar in terms of gender distribution, but the adult sample reported by Kam et al. and the Ann Arbor samples from Berinsky and Kinder are substantially more female. The MTurk sample has a similar (high) education level as compared to the two samples by Berinsky and Kinder, whereas the adult sample

[11]The HIT was described as follows: Title: Survey of Public Affairs and National Conditions. Description: Complete a survey to gauge your opinion of national conditions and current events (USA only). Should be no more than 10 mins. Keywords: survey, current affairs, research, opinion, politics, fun. Detailed Posting: Complete this research survey. Usually takes no more than 10 minutes. You can find the survey here: [URL removed]. At the end of the survey, you'll find a code. To get paid, please enter the code below.

[12]MTurk classifies individuals as 18 or older based on self-reports. MTurk does not reveal how it classifies individuals as living in a particular country but may rely on mailing addresses and credit card billing addresses.

[13]These individuals may reside in the United States but be traveling or studying abroad. Additionally, although IP address locators seem reliable, we are unaware of research benchmarking their accuracy. Still, so as to provide as conservative a picture of our sample as is possible, we excluded these questionable respondents. Our results did not change when we included them.

[14]Other researchers have surveyed MTurk respondents and found a similar demographic profile (e.g., Ross et al. 2010).

Table 2 Comparing MTurk sample with other convenience samples

		Convenience Samples			
				Adult samples (Berinsky and Kinder 2006)	
Demographics	*MTurk*	*Student samples (Kam et al. 2007)*	*Adult sample (Kam et al. 2007)*	*Experiment 1: Ann Arbor, MI*	*Experiment 2: Princeton, NJ*
Female	60.1% (2.1)	56.7% (1.3)	75.7% (4.1)	66.0%	57.1%
Age (mean years)	32.3 (0.5)	20.3 (8.2)	45.5 (.916)	42.5	45.3
Education (mean years)	14.9 (0.1)	—	5.48 (1.29)	15.1	14.9
White	83.5 (1.6)	42.5	82.2 (3.7)	81.4	72.4
Black	4.4 (0.9)			12.9	22.7
Party identification					
Democrat	40.8 (2.1)			46.1	46.5
Independent	34.1 (2.0)			20.6	17.6
Republican	16.9 (1.6)			16.3	25.8
None/other	8.2 (1.2)			17.0	10.1
N	484–551	277–1428	109	141	163

Note. Percentages except for age and education with SEs in parentheses. Adult sample from Kam et al. (2007) is for campus employee participants from their Table 1, Column 1. MTurk survey is from February/March 2010.

reported by Kam et al. is much less educated (even compared to national probability samples; see Table 3). Finally, in terms of racial composition, the MTurk sample is much less white than the student sample by Kam et al. (which is again very different from national samples; see Table 3) but similar to the adult sample by Kam et al. and the Ann Arbor sample by Berinsky and Kinder.[15]

More importantly for the purposes of political science experiments, the Democratic party identification skew in the MTurk sample is better, relative to the ANES (see Table 3), than in either adult sample from Berinsky and Kinder. Of course, our point is not to single out Berinsky and Kinder—the distribution of relevant demographic and political variables in their study was, in fact, more representative than that found in several other studies.[16] Instead, we simply wish to emphasize that, when compared to the practical alternatives, the MTurk respondent pool has attractive characteristics—even apart from issues of cost.

4.2 *Comparison of Respondent Characteristics: Internet Samples and High-Quality, Face-To-Face, Probability Samples*

Besides local convenience samples, the other dominant form of sample recruitment in published experimental studies is Internet-based surveys. MTurk also fares reasonably well when compared to these samples. In this section, we compare MTurk to a high-quality Internet survey, the American National Election 2008–09 Panel Study (ANESP). The firm Knowledge Networks conducted the ANESP by recruiting respondents through a random-digit-dial method for a 21-wave Internet-based panel survey (10 waves of the survey concerned political matters; the other 11 waves did not).[17] Since we are treating the ANESP not as a best estimate of true population parameters, but rather as an example of a high-quality Internet sample, we present unweighted results from this survey. Comparing our MTurk survey with the ANESP has an additional advantage. Since both are Internet surveys, we hold the ''mode'' of survey constant. Additionally, in designing our MTurk survey, we followed the ANESP as closely as possible, using identical question wordings and branching formats.

[15]The MTurk sample does have fewer blacks than either of the Berinsky and Kinder adult samples.
[16]Moreover, as the material in the Supplementary data makes clear, many other studies do not report any information about sample characteristics.
[17]Prospective respondents were offered $10 per month to complete surveys on the Internet for 30 min each month.

Table 3 Comparing MTurk sample demographics to Internet and face-to-face samples

	Internet sample		*Face-to-face samples*	
	MTurk	*ANESP*	*CPS 2008*	*ANES 2008*
Female	60.1% (2.1)	57.6% (0.9)	51.7% (0.2)	55.0% (1.3)
Education (mean years)	14.9 (0.1)	16.2 (0.1)	13.2 (0.0)	13.5 (0.1)
Age (mean years)	32.3 (0.5)	49.7 (0.3)	46.0 (0.1)	46.6 (0.5)
Mean income	$55,332 ($1,659)	$69,043 ($794)	$62,256 ($130)	$62,501 ($1,467)
Median income	$45,000	$67,500	$55,000	$55,000
Race				
White	83.5 (1.6)	83.0 (0.7)	81.2 (0.1)	79.1 (0.9)
Black	4.4 (0.9)	8.9 (0.5)	11.8 (0.1)	12.0 (0.6)
Hispanic	6.7 (1.1)	5.0 (0.4)	13.7 (0.1)	9.1 (0.5)
Marital status				
Married	39.0 (2.1)	56.8 (0.9)	55.7 (0.2)	50.1 (1.3)
Divorced	7.1 (1.1)	12.1 (0.6)	10.2 (0.1)	12.9 (0.8)
Separated	2.5 (0.7)	1.3 (0.2)	2.1 (0.1)	2.9 (0.4)
Never married	50.6 (2.1)	14.2 (0.6)	25.7 (0.2)	26.2 (1.1)
Widowed	0.7 (0.4)	4.9 (0.4)	6.3 (0.1)	7.8 (0.6)
Housing status				
Rent	52.7 (2.3)	14.3(0.1)		32 (1.2)
Own home	47.3 (2.3)	80.8 (0.8)		66.1 (1.2)
Religion				
None	41.8 (2.1)	13.1 (0.8)		26.9 (1.2)
Protestant	20.7 (1.7)	38.7 (1.4)		28.2 (1.2)
Catholic	16.5 (1.6)	22.9 (1.0)		17.5 (1.0)
Jewish	4.4 (0.9)	3.0 (0.4)		1.2 (0.3)
Other	16.5 (1.6)	22.2 (1.0)		26.2 (1.1)
Region of the United States				
Northeast	22.1 (1.8)	16.9 (0.7)	18.4 (0.1)	14.6 (0.9)
Midwest	26.6 (1.9)	28.3 (0.9)	21.9 (0.1)	21.2 (1.1)
South	30.9 (2.0)	31.4 (0.9)	36.5 (0.2)	42.8 (1.2)
West	20.4 (1.7)	23.4 (0.8)	23.1 (0.2)	21.4 (0.9)
N	548–551	2,727–3,003	100,008	2,307–2,312

Note. Percentages except for education, age, and income with SEs in parentheses. CPS 2008 and ANES 2008 are weighted. MTurk survey is from February/March 2010. Tests of statistical significance of differences across samples appear in the Supplementary data.

To put these comparisons in perspective, we benchmark them against nationally representative samples, including the CPS and the 2008 American National Elections Study (ANES 2008). These latter two studies use face-to-face probability samples and are widely considered the ''gold standard'' for survey sampling. In comparing these samples, it should be noted that most differences were statistically significant, even when those differences were substantively trivial (this is due, in part, to the large sample sizes of the ANESP, the CPS, and the ANES). We present these significance tests in the Supplementary data.[18]

In Table 3, we begin by presenting means and SEs for key demographic variables. For continuous and near-continuous measures (Age, Education, and Income), we also plot the distribution for these four samples in Fig. 1. On many demographics, the MTurk sample is very similar to the unweighted ANESP. Starting with gender, MTurk is only slightly more female than the ANESP, 60% versus 58%, and slightly less educated, 14.9 versus 16.2 years. As Fig. 1 shows, both education estimates are somewhat higher than

[18]We therefore only report significance tests in the exceptional cases when they are not statistically significant, relying on Kolmogorov-Smirnov tests of differences in distributions (and proportion tests for categorical variables). As shown in the Supplementary data, about 85% of the tests between MTurk and the other samples are statistically significant at the 0.10 threshold. In comparison, about 60% of the tests between ANES 2008 and CPS 2008 are significant.

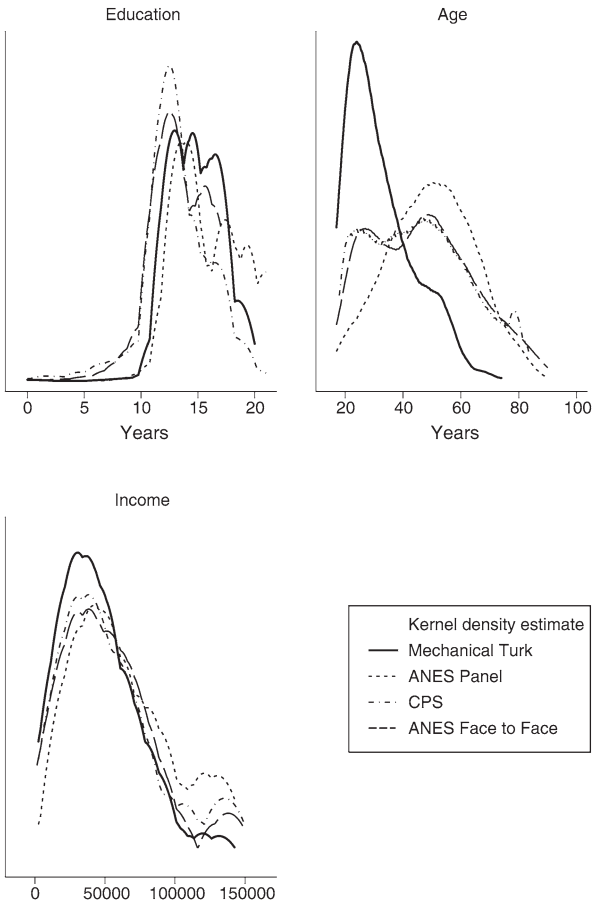

Fig. 1 Comparing MTurk sample on selected demographic measures to face-to-face samples. CPS 2008 and ANES Face to Face 2008 are weighted. Tests of statistical significance of differences across samples appear in the Supplementary data.

in the face-to-face national probability samples, indicating that both MTurk and ANESP somewhat underrepresent low education respondents. On age, the MTurk sample is notably younger. Consistent with this age gap, both mean and median income are lower than in the other samples. On race, MTurk's characteristics are mixed: It is similar to the CPS on percent white, but underrepresents blacks and Hispanics.

Not surprisingly, MTurk fares worse in comparison to both the ANESP and CPS on demographic characteristics related to life cycle events, such as marital status, homeownership, and religious preference. MTurk subjects are more likely to: have never married (51%), rent rather than own their home (53%), and report no religious affiliation (42%).[19] Finally, the MTurk sample is broadly similar to the other samples in region of residence, with perhaps a slightly larger prevalence of those living in the Northeast.

We next compare the samples on key political and psychological measures, including partisanship, ideology, political interest, political knowledge, need for cognition, and need to evaluate.[20] These measures are often used in experimental research and are common on key political science surveys. These distributions are presented in Table 4 and Fig. 2. Beginning with registration and 2008 turnout, the MTurk

[19]On demographics, the only nonsignificant differences between MTurk and the other samples are on gender, marriage separation, Catholic, and region.

[20]The need for cognition and need to evaluate scales are from the 2008 ANES. These items were placed on a separate survey of 699 MTurk respondents conducted in May 2011. This study also contained the Kam and Simas (2010) replication discussed below. The HIT was described as follows: Title: Survey of Public Affairs and Values. Description: Relatively short survey about opinions and values (USA only). 10–12 minutes. Keywords: survey, relatively short. Detailed Posting: Complete this research survey. Usually takes 10–12 minutes. You can find the survey here: [URL removed]. At the end of the survey, you'll find a code. To get paid, please enter the code below.

Table 4 Comparing MTurk sample political and psychological measures to Internet and face-to-face samples

	Internet sample		*Face-to-face samples*	
	MTurk	*ANESP*	*CPS 2008*	*ANES 2008*
Registration and turnout				
Registered	78.8% (1.7)	92.0% (0.7)	71.0% (0.2)	78.2% (1.1)
Voter turnout 2008	70.6 (2.0)	89.8 (0.5)	63.6 (0.2)	70.4 (1.1)
Party identification (mean on 7-point scale, 7 = Strong Republican)	3.48 (0.09)	3.90 (0.05)		3.70 (0.05)
Ideology (mean on 7-point scale, 7 = Strong conservative)	3.39 (0.09)	4.30 (0.05)		4.24 (0.04)
Political Interest (mean on 5-point scale, 5 = Extremely interested)	2.43 (0.04)	2.71 (0.02)		2.93 (0.03)
Political knowledge (% correct)				
Presidential succession after Vice President	70.0 (1.3)	65.2 (2.0)		
House vote percentage needed to override a veto	81.3 (1.7)	73.6 (1.3)		
Number of terms to which an individual can be elected president	96.2 (0.8)	92.8 (0.7)		
Length of a U.S. Senate term	45.0 (2.1)	37.5 (1.3)		
Number of Senators per state	85.4 (1.5)	73.2 (1.2)		
Length of a U.S. House term	50.1 (2.1)	38.9 (1.3)		
Average	71.3	63.5		
Need for cognition (mean on 0–1 scale)	.625 (0.012)	.607 (0.006)		.559 (0.009)
Need to evaluate (mean on 0–1 scale)	.628 (0.008)	.579 (0.004)		.558 (0.005)
N	506–699	1,466–2,984	92,360	1,058–2,323

Note. Means with SEs in parentheses. CPS 2008 and ANES 2008 are weighted. Political measures are from the February/March 2010 MTurk survey (N = 551). Need for Cognition and Need to Evaluate are from the May 2011 MTurk survey (N = 699). Tests of statistical significance of differences across samples appear in the Supplementary data.

sample is more similar to the nationally representative samples than is the ANESP.[21] MTurk respondents are slightly more Democratic in their partisan identification than are ANESP respondents and are substantially more liberal in their ideology (a difference especially visible in Fig. 2).

MTurk respondents are also somewhat more interested in politics than ANESP respondents are, and both samples are considerably more interested than are ANES respondents.[22] We also administered a battery of six political knowledge items from the ANESP. This battery includes questions about the line of succession for the presidency, the length of a U.S. Senate term, and the number of federal senators per state. Just before asking these questions, we instructed respondents to provide just their best guess and not to look up answers. For each item, we offered four answer options in a multiple-choice format.[23] Based on their responses, MTurk subjects appear more knowledgeable than ANESP respondents, but the gap is not large (though it is statistically significant in four of six cases). Finally, for both need for cognition and need to evaluate, the MTurk sample scores higher than the ANESP, which is itself higher than the ANES 2008. But Fig. 2 shows that the distributions are nevertheless quite similar.

[21]In fact, differences between MTurk and ANES 2008 on registration and turnout are not statistically significant (see the Supplementary data). There are inconsistencies in the ANESP's measures of turnout and registration (e.g., the survey contains respondents who say they are not registered to vote, but report voting) that suggest caution here.

[22]This result is somewhat odd because workers visit MTurk to make money, not because they are interested in politics. The higher levels of interest may be due to advertising the survey as about ''public affairs.''

[23]To check whether MTurk subjects looked up answers to knowledge questions on the Internet, we asked two additional multiple choice questions of much greater difficulty: who was the first Catholic to be a major party candidate for president and who was Woodrow Wilson's vice president. Without cheating, we expected respondents to do no better than chance. On the question about the first Catholic candidate, MTurk subjects did worse than chance with only 10% answering correctly (Alfred Smith; many chose an obvious but wrong answer, John F. Kennedy). About a quarter did correctly answer the vice presidential question (Thomas Marshall), exactly what one would expect by chance. These results suggest political knowledge is not inflated much by cheating on MTurk.

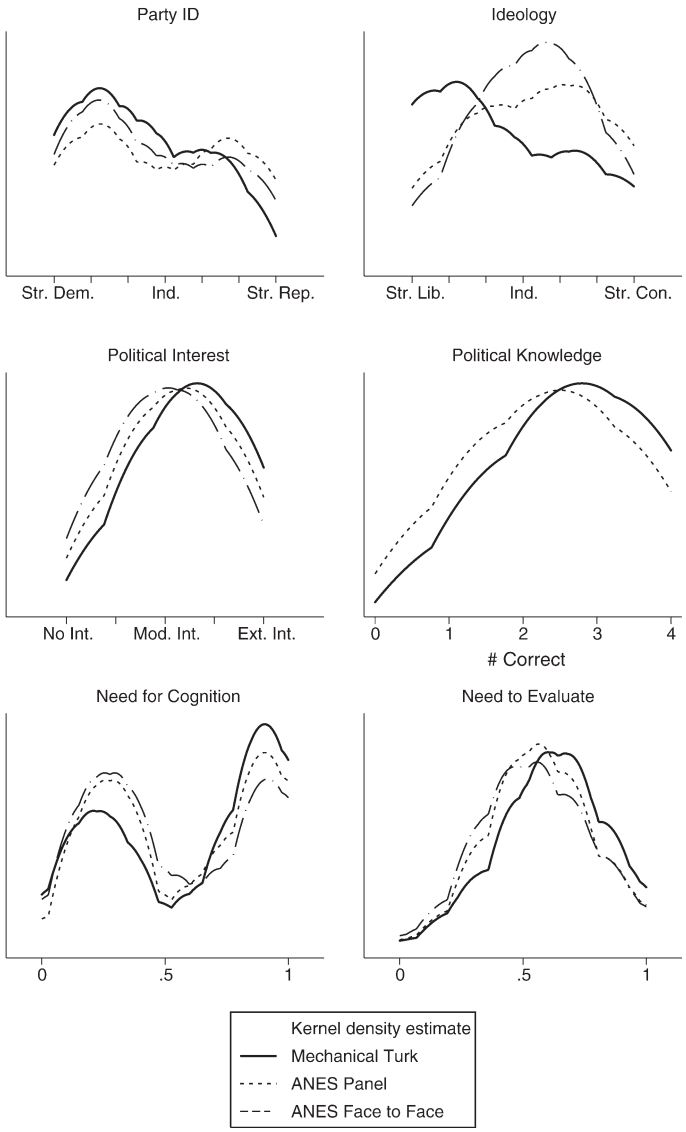

Fig. 2 Comparing MTurk sample selected demographics to face-to-face samples. ANES Face to Face 2008 is weighted. Tests of statistical significance of differences across samples appear in the Supplementary data.

Finally, we asked the MTurk sample several attitudinal questions that mirrored questions on the ANESP and the ANES 2008 (see Table 5). These questions asked about support for the prescription drug benefit for seniors, universal health care, and a citizenship process for illegal immigrants. The MTurk responses match the ANES well on universal health care—about 50% of both samples support it—whereas those in the ANESP are somewhat less supportive at 42%. MTurk also compares reasonably well on the question about a citizenship process for illegal immigrants. Perhaps as a function of the age skew of the sample or a different political environment after the political discussions surrounding the Obama health care initiative, MTurk respondents are less supportive of the prescription drug benefit for seniors compared to the ANES and ANESP—64% of MTurk respondents favor the benefit, compared to 75% of ANESP and 80% of ANES respondents.

Our MTurk survey also included three additional policy questions from the ANESP that were not included on the 2008 ANES. These asked about support for a constitutional amendment banning gay marriage, raising taxes on people making more than $200,000, and raising taxes on people making less than $200,000. Compared to the ANESP, MTurk subjects express somewhat more liberal views on all three items, with only 16% supporting a constitutional amendment banning gay marriage, compared

Table 5 Comparing MTurk sample policy attitudes to Internet and face-to-face samples

	MTurk	*Internet sample ANESP*	*Face-to-face samples ANES 2008*
Favor prescription drug benefit for seniors	63.5% (2.0)	74.8% (1.1)	80.1% (1.5)
Favor universal health care	47.8 (2.1)	41.7 (1.2)	51.0 (1.9)
Favor citizenship process for illegals	38.1 (2.1)	42.7 (1.2)	49.1 (1.9)
Favor a constitutional amendment banning gay marriage	15.6 (1.5)	30.7 (1.2)	
Favor raising taxes on people making more than $200,000	61.2 (2.1)	55.4 (1.2)	
Favor raising tax on people making less than $200,000	6.2 (0.1)	7.1 (0.6)	
N	551	1,614–1,618	1,142–1,156

Note. Percentages supporting each policy with SEs in parentheses. ANES 2008 is weighted. MTurk survey is from February/March 2010. Tests of statistical significance of differences across samples appear in the Supplementary data.

to 31% in the ANESP (Table 5).[24] On both tax increase items, MTurk subjects are only a few percentage points more liberal in their views (and these differences are not statistically significant).

All told, these comparisons reinforce the conclusion that the MTurk sample does not perfectly match the demographic and attitudinal characteristics of the U.S. population but does not present a wildly distorted view of the U.S. population either. Statistically significant differences exist between the MTurk sample and the benchmark surveys, but these differences are substantively small. MTurk samples will often be more diverse than convenience samples and will always be more diverse than student samples. Thus, if we treat the MTurk as a means for conducting internally valid experiments, instead of a representative sample, the MTurk respondent pool is very attractive. At the same time, if one is interested in estimating treatment effects that may differ due to any of the factors for which the MTurk sample is less representative, then the MTurk sample may yield estimates that are inaccurate for the larger population. For example, if one believes that older or more conservative citizens are particularly responsive to treatments, researchers should be cautious about drawing broader conclusions. Furthermore, given the relative dearth of these sorts of individuals in the MTurk pool, large sample sizes may be necessary to obtain sufficient diversity on these dimensions to estimate differences in treatment effects for these groups, an issue that may guide ex ante targets for sample populations.[25] Tables 3–5 and Figs. 1 and 2, which compare the MTurk to Internet and face-to-face samples, provide clear guidance for which variables these concerns should be most salient.

4.3 *Benchmarking via Replication of Experimental Effects*

To further assess MTurk's usefulness as a vehicle for experimental research, we also replicated the results reported in three experiments. The first is a classic study of the effect of question wording on survey responses, the second is a canonical framing experiment, and the third is a recently published political science experiment on the effects of risk preferences on susceptibility to framing. In all three cases, the experimental results found using the MTurk sample are highly similar to those found in published

[24]As with the drug benefit, this difference may be due to age or to differences in political circumstances.

[25]These discrepancies also suggest the potential utility of using an initial survey to screen large numbers of individuals and then inviting a more representative subset of those respondents to participate in the experiment itself. We discuss the technique for contacting selected respondents for a follow-up survey in footnote 8.

research. (Additional question wording and design details necessary to conduct these replications appear in the Supplementary data.)

4.3.1 Experiment 1: Welfare spending

Rasinski (1989) reports results from a question wording experiment that asked representative samples from the General Social Surveys (GSS) from 1984 to 1986 whether too much or too little was being spent on either ''welfare'' or ''assistance to the poor.'' The GSS is a nationally representative face-to-face interview sample with characteristics similar to the ANES face-to-face surveys. Even though welfare and assistance to the poor are thought by policy experts to refer to the same policy, the study found important differences in levels of support between the two question forms. Although 20%–25% of the respondents in each year said that too little was being spent on welfare, 63%–65% said that too little was being spent on assistance to the poor (Rasinski 1989, 391). The GSS has continued to ask the spending experiment and the gap in support for increasing spending between the question forms remains similar over time, ranging from 28% to 50%, with an average difference of 37% (Green and Kern 2010).

We ran the same between-subjects experiment on MTurk ($N = 329$) in our original MTurk demographic survey described above. Respondents were randomly assigned to either the welfare or assistance to the poor version of the question. We found a statistically significant 38 percentage point gap (p value $< .001$) between the two conditions that is similar in magnitude to that found using the GSS (the comparable gap was 44% in the 2010 GSS). Only 17% of MTurkers said too little was being spent on welfare, whereas 55% said too little was being spent on assistance to the poor.[26] We also explored whether these treatment effects were heterogeneous. In both the 2010 GSS and the MTurk sample, the experimental effect does not differ reliably across gender, education, or racial lines.[27] Furthermore, a test of the difference in the estimated treatment effect by demographic group crossed with differences in experimental administration (MTurk versus GSS) also yields null results.[28] We present this analysis in the Supplementary data.

4.3.2 Experiment 2: Asian disease problem

On a separate MTurk survey ($N = 450$), we also replicated a classic framing experiment—the ''Asian Disease Problem'' reported in Tversky and Kahneman (1981).[29] This experiment was first conducted using a student sample but has been replicated in many other settings (e.g., Takemura 1994; Kuhberger 1995; Jou, Shanteau, and Harris 1996; Bless, Betsch, and Franzen 1998; Druckman 2001). All respondents were initially given the following scenario:

> Imagine that your country is preparing for the outbreak of an unusual disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows:

They were then randomly assigned to one of the two following conditions:

Condition 1, Lives Saved: If Program A is adopted, 200 people will be saved. If Program B is adopted, there is one-third probability that 600 people will be saved, and two-third probability that no people will be saved.

[26] The support for increased spending is, on average, somewhat higher in *both* conditions on the GSS. Specifically, in 2010, the GSS data show that 24% think that too little is being spent on welfare, whereas 68% think that too little is spent on assistance to the poor.

[27] Prior work similarly finds no evidence that gender or race are associated with differences in effect sizes in the GSS in years earlier than 2010 (Green and Kern 2010).

[28] To conduct these tests, we pooled the MTurk and GSS samples. We then ran three ordered probits using the three-category welfare spending response scale as the dependent variable—one for each of the demographic variables (men versus women; college educated versus other; blacks versus all other races). For each of these probits, we included as independent variables a dummy variable for sample (GSS versus MTurk), a dummy variable for the treatment (welfare versus assistance to the poor question form), the demographic variable of interest (education, gender, or race), and interactions between all the variables. The interactions between the treatment and the demographic variables allow us to test whether heterogeneous treatment effects are present, whereas the three-way interaction between the demographic variable, the sample, and the treatment allow us to test whether the treatment effect varies by demographic subgroup across forms. In all cases, these interaction terms were insignificant (the p values on the terms range from .28 to .78). The full ordered probit results are presented in the Supplementary data.

[29] This survey was fielded in January 2010. HIT was described as follows: Title: Answer a survey about current affairs and your beliefs. Description: ''Answer a survey about current affairs and your beliefs. Should take less than 5 minutes.'' Paolacci et al. (2010) also reports an MTurk replication of this experiment.

Condition 2, Lives Lost: If Program A is adopted, 400 people will die. If Program B is adopted there is one-third probability that nobody will die, and two-third probability that 600 people will die.

In each of these conditions, respondents are asked to choose one of two policy options. The first is a program with certain consequences. The second is a program that comes with risks—the outcome is probabilistic. These scenarios are exactly the same in their description of the expected consequences of each program, but differ in framing. In Condition 1, both the certain program and the risky program are described in terms of likelihood of positive outcomes, namely the lives saved by the programs. In Condition 2, by contrast, the two programs are described in terms of likelihood of negative outcomes—the lives lost by the different options. Tversky and Kahneman report that, when the problem was framed in terms of ''lives saved,'' respondents were more likely to pick the certain choice, while when it was framed in terms of ''lives lost,'' respondents were more likely to pick the risky choice. Framing the outcomes in positive terms therefore produced a reversal of participants' preferences for the two programs compared to when it was presented in negative terms.

In the original Tversky and Kahneman experiment, which was run with student samples, 72% of respondents picked the certain choice in the ''lives saved'' condition, as compared to 22% who picked the certain choice in the ''lives lost'' condition. We find a similar pattern among our MTurk sample: 74% pick the certain choice in the lives saved condition and 38% select the certain choice in the lives lost condition (p value of test of statistical significance of difference $< .001$). Thus, although the gap is smaller than in the student sample, we still observe the large preference reversal reported in Tversky and Kahneman (and replicated in subsequent experiments with different samples).

4.3.3 Experiment 3: Framing and risk

The first two studies we replicated were both straightforward experiments with large effect sizes. To demonstrate the utility of MTurk as a tool for subject recruitment for more complex political science experiments, we also replicated a third study conducted by Kam and Simas (2010) that uses a Knowledge Networks probability sample and was published in the *Journal of Politics*.[30]

Kam and Simas used a modified version of the Tversky and Kahneman framing experiment and an original scale measuring individual differences in risk orientation to explore how both the frame of a problem and variation in risk proclivities affected choice under uncertainty. They first use a between-subject design to test framing effects. As in the original Tversky and Kahneman framing study, the frame (lives lost versus lives saved) influenced the choices respondents made. Moreover, as predicted, individuals who scored high on the risk orientation scale were more likely to choose the risky policy option, regardless of the frame.[31]

For this replication, the experimental materials were taken verbatim from Kam and Simas (2010). In Table 6, we present the results for our MTurk sample of the Kam and Simas between-subjects experiment alongside the original Kam and Simas results from Table 2 of their article (the analysis from their Tables 3 and 4 are presented in the Supplementary data). The dependent variable in this analysis is a binary variable measuring preference for the risky policy option over the certain policy option. The similarities between the experimental results are striking. The lives lost mortality frame increases support for the risky policy option—the probabilistic outcome—in both experiments and across all three model specifications. Moreover, the coefficients on the effect of the frame are very similar between the original study and our replication (1.07 in the original article compared to 1.18 in the MTurk sample). Additionally, higher levels of risk acceptance are associated with greater support for the risky policy option in both experiments and the coefficients are also similar across all three specifications.[32] Finally, in the third specification for each sample, like Kam and Simas, we also find a statistically insignificant coefficient on the interaction between risk

[30]We chose to replicate the study by Kam and Simas because it is an excellent example of the way in which contemporary political scientists use experimentation to understand key political dynamics. That study examines the importance both of framing and of the relationship between framing and underlying preferences for risk aversion (i.e., heterogeneity in treatment effects).

[31]Kam and Simas also employed a within-subjects design to show that high levels on the risk acceptance scale reduce susceptibility to framing effects across successive framing scenarios. We replicated these results as well (see Supplementary data).

[32]We also find a somewhat different pattern of signs for the coefficients on the control variables—notably education and income. However, the coefficients on these variables—both in our analysis and in the original article by Kam and Simas—fall short of statistical significance by a wide margin.

Table 6 Replication of Table 2 by Kam and Simas (2010)—risk acceptance and preference for the probabilistic outcome

	Kam and Simas (2010)			*MTurk replication*		
	(H1a) Mortality frame and risk acceptance	*(H1b) Adding controls*	*(H2) Frame × Risk acceptance*	*(H1a) Mortality frame and risk acceptance*	*(H1b) Adding controls*	*(H2) Frame × Risk acceptance*
Mortality frame in Trial 1	1.068 (0.10)	1.082 (0.10)	1.058 (0.29)	1.180 (0.10)	1.180 (0.10)	1.410 (0.31)
Risk acceptance	0.521 (0.31)	0.628 (0.32)	0.507 (0.48)	0.760 (0.29)	0.780 (0.31)	0.990 (0.42)
Female		0.105 (0.10)			−0.018 (0.11)	
Age		0.262 (0.22)			0.110 (0.31)	
Education		−0.214 (0.20)			0.025 (0.23)	
Income		0.205 (0.23)			−0.024 (0.23)	
Partisan ideology		0.038 (0.19)			0.006 (0.15)	
Risk acceptance × Mortality frame			0.023 (0.62)			−0.450 (0.58)
Intercept	−0.706 (0.155)	−0.933 (0.259)	−0.700 (0.227)	−1.060 (−0.170)	−1.100 (−0.290)	−1.190 (−0.230)
lnL	−453.185	−450.481	−453.184	−409.740	−409.662	−409.439
$p > \chi^2$	0.000	0.000	0.000	0.000	0.000	0.000
N	752	750	752	699	699	699

Note. Entries are probit coefficients with SEs in parentheses. Dependent variable is Preference for the Probabilistic Outcome (0 = deterministic outcome; 1 = probabilistic outcome). All independent variables are scaled to range from 0 to 1. MTurk survey is from May 2010. None of the differences between coefficients across studies are statistically significant (see the Supplementary data).

assessment and exposure to the mortality frame (though the coefficient is in both cases statistically insignificant by a wide margin, we find a negative sign on the interaction between risk assessment and mortality frame, whereas Kam and Simas find a positive sign). Overall, although our estimate of the predictive power of risk assessment is modestly larger than in the original article, the basic pattern of effects is the same.

4.4 *Habitual Participants Analysis*

A final source of concern about external validity for any self-selected sample is the potential for ''habitual customers.'' If the same subjects take several surveys, there is the potential for cross-experiment stimuli contamination (Transue, Lee, and Aldrich 2009). To assess the severity of this problem, we asked our respondents how many political surveys they had taken in the last month on MTurk. The mean was 1.7; 39% of the respondents took no other survey, whereas 78% took two or fewer surveys.

We also assessed the prevalence of habitual participants by examining a broad range of experiments run on MTurk. We gathered the unique MTurk ID number for all workers who participated in each of seven studies we conducted from January 2010 to April 2010. The compensation for these studies ranged from 10 to 50 cents, and the *N* ranged from 200 to 587. Across the seven experiments, there were a total of 1,574 unique subjects. Of these subjects, 70% participated in only one experiment; another 18% participated in two experiments. Only 2% of the subjects participated in five or more experiments.

Although this set of experiments represents only a small proportion of those conducted on MTurk, our findings may illuminate broader trends. Although there are certainly a handful of respondents who participate habitually in experiments, the majority of MTurkers are not chronic study participants. Furthermore, the presence of these habitual responders does not seem to pose a threat to our inferences. In the experiments presented above, we found that the effects did not differ—in either a statistical or a substantive sense—when we examined the habitual respondents and the nonhabitual respondents separately.[33] All told, our results, combined with other replications of well-known experiments in other fields by other scholars (Horton, Rand, and Zeckhauser 2010), provide further support for the external validity of MTurk as an experimental platform.[34]

5 Assessing Threats to Internal Validity

In the previous section, we noted that estimates of experimental treatment effects are similar for habitual and one-time participants. This result addresses both external and internal validity concerns. To further explore the internal validity of experiments conducted using MTurk samples, we also examined whether MTurk users appear to violate treatment assignment and their engagement with experimental stimuli.

5.1 *Do MTurk Workers Violate Assignment by Participating in Experiments Multiple Times?*

We sought to assess whether a given respondent took our survey more than once. By default, each HIT (survey) can only be completed by a single worker. However, an individual could potentially subvert this process by opening multiple MTurk accounts (though this behavior would violate the terms of the MTurk user agreement). They could then take the survey once from each account, which might expose them to more than one treatment condition. Given the relatively low pay rate of our studies and the availability of other paid work, we do not believe our work is likely to encourage such behavior. Nevertheless, we did check to see if multiple responses came from a single IP address. We found that a total of seven IP addresses produced two responses each to our demographic survey (i.e., 14 of 551 responses or 2.4% of the total responses). This pattern is not necessarily evidence of repeat survey taking. It could, for example, be the case that these IP addresses were assigned dynamically to different users at different points in time or that multiple people took the survey from the same large company, home, or even coffee shop.

[33] We conducted the habitual responder analysis for the welfare and Asian flu experiments (see Supplementary data). We do not perform this analysis for the study by Kam and Simas because it was conducted a year after our data on frequent participants was collected.

[34] Lawson et al. (2010) successfully replicate the ratings of 2006 Senate candidate faces on MTurk by Ballew and Todorov (2007). Horton, Rand, and Zeckhauser (2010) replicate several experimental findings in economics. Gabriele Paolacci's Experimental Turk blog (http://experimentalturk.wordpress.com/) has collected reports of successful replications of several canonical experiments from a diverse group of researchers, including the Asian Disease Problem discussed in this section and other examples from psychology and behavioral economics.

But even if these are cases of repeat survey taking, only a handful of responses would be contaminated, suggesting that repeat survey taking is not a large problem in the MTurk subject pool.[35]

5.2 *Attention, Demand, and Subject Motivation*

Given their incentives, MTurk respondents may generally pay greater attention to experimental instruments and survey questions than do other subjects. Since Requesters often specify at least a 95% prior ''approval rate''—that is, previous Requesters accepted 95% or more of the HITs submitted by an individual—respondents have an incentive to read instructions carefully and consider their responses.

Our experiences are consistent with this expectation. In a study conducted by one of the authors, subjects were asked to identify the political office held by a person mentioned in a story they had just read. The format of this question was a multiple-choice item with five possible responses. On the MTurk study, 60% of the respondents answered the question correctly. An identical question concerning the same article was also included on experiments run through Polimetrix/YouGov (another high-quality Internet panel) and with a sample collected by Survey Sampling International (SSI). The correct answer rates on these platforms were markedly lower than in the MTurk sample—49% on Polimetrix/YouGov and 46% on SSI.

Although a concern for pleasing the researcher has benefits, it may also have costs. MTurk respondents may pay close attention to experimental stimuli, but they may also exhibit experimental demand characteristics to a greater degree than do respondents in other subject pools, divining the experimenter's intent and behaving accordingly (Orne 1962; Sears 1986). To avoid this problem and the resulting internal validity concerns, it may be desirable to avoid signaling to subjects ahead of time the particular aims of the experiment.[36] Demand concerns are relevant to any experimental research, but future work needs to be done to explore if these concerns are especially serious with respect to the MTurk respondent pool and how they are affected by recruitment and consent text.

6 Conclusion

This article describes the potential advantages and limitations of using Amazon.com's MTurk platform as a subject recruitment device for experimental research. We demonstrate that relative to other convenience samples often used in experimental research in political science, MTurk subjects are often more representative of the general population and substantially less expensive to recruit. MTurk subjects appear to respond to experimental stimuli in a manner consistent with prior research. They are apparently also not currently an excessively overused pool, and habitual responding appears to be a minor concern. Put simply, despite possible self-selection concerns, the MTurk subject pool is no worse than convenience samples used by other researchers in political science. The analysis we undertake for the MTurk pool also provides a template for evaluating the desirability of other means of subject recruitment.

Despite these advantages, several aspects of MTurk should engender caution. In particular, MTurk subjects are notably younger and more ideologically liberal than the public, which may limit their suitability for some research topics. They also appear to pay more attention to tasks than do other respondents. Finally, as use increases, habitual responding may pose more of an external validity problem.

Interest in experimental research has risen substantially in political science, but experiments can be difficult and costly to implement. MTurk potentially provides an important way to overcome the barrier to conducting research raised by subject recruitment costs and difficulties by providing easy and inexpensive access to nonstudent adult subjects. Our results provide researchers with a clearer understanding of the potential advantages of the MTurk tool for conducting experiments as well as areas where caution may be in order.

References

Ballew, Charles C. II, and Alexander Todorov. 2007. Predicting political elections from rapid and unreflective face judgments. *Proceedings of the National Academy of Sciences of the United States of America* 104:17948–53.

[35]Researchers can reject and block future work by suspected retakers or simply exclude duplicate work from their analysis by selecting only the first observation from a given IP address.

[36]In the case of experiments involving deception, it is also feasible to debrief at the conclusion of the experiment.

Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz. 2011. *Replication data for: Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk*. IQSS Dataverse Network [Distributor] V1 [Version]. http://hdl.handle.net/1902.1/17220 (accessed January 19, 2012).

Berinsky, Adam J., and Donald R. Kinder. 2006. Making sense of issues through media frames: Understanding the Kosovo crisis. *Journal of Politics* 68:640–56.

Bless, Herbert, Tilmann Betsch, and Axel Franzen. 1998. Framing the framing effect: The impact of context cues on solutions to the 'Asian disease' problem. *European Journal of Social Psychology* 28:287–91.

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

Chandler, Dana, and Adam Kapelner. 2010. *Breaking monotony with meaning: Motivation in crowdsourcing markets*. University of Chicago Mimeo.

Chen, Daniel L., and John J. Horton. 2010. *The wages of pay cuts: Evidence from a field experiment*. Harvard University Mimeo.

Chong, Dennis, and James N. Druckman. 2010. Dynamic public opinion: Communication effects over time. *American Political Science Review* 104:663–80.

Downs, Julie S., Mandy B. Holbrook, Steve Sheng, and Lorrie Faith Cranor. 2010. Are your participants gaming the system? Screening Mechanical Turk workers. Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010, 2399–402. New York: ACM Press.

Druckman, James N. 2001. Evaluating framing effects. *Journal of Economic Psychology* 22:91–101.

Druckman, James N., and Cindy D. Kam. 2011. Students as experimental participants: A defense of the 'narrow data base'. In *Handbook of experimental political science*, eds. James N. Druckman, Donald P. Green, James H. Kuklinski, Arthur Lupia, 41–57. New York: Cambridge University Press.

Gerber, Alan S., James G. Gimpel, Donald P. Green, and Daron R. Shaw. 2011. How large and long-lasting are the persuasive effects of televised campaign ads? Results from a randomized field experiment. *American Political Science Review* 105:135–50.

Green, Donald P., and Holger L. Kern. 2010. Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. Yale University Mimeo.

Gosling, Samuel D., Simine Vazire, Sanjay Srivastava, and Oliver P. John. 2004. Should we trust web-based studies? *American Psychologist* 59:93–104

Horton, John J., David G. Rand, and Richard J. Zeckhauser. 2010. *The online laboratory: Conducting experiments in a real labor market*. Available at SSRN: http://ssrn.com/abstract=1591202 (accessed January 19, 2012).

Horton, J., and L. Chilton. 2010. The labor economics of paid crowdsourcing. Proceedings of the 11th ACM Conference on Electronic Commerce, Cambridge, MA.

Jou, Jerwen, James Shanteau, and Richard Harris. 1996. An information processing view of framing effects: The role of causal schemas in decision making. *Memory & Cognition* 24(1):1–15.

Kam, Cindy D., and Elizabeth N. Simas. 2010. Risk orientations and policy frames. *Journal of Politics* 72:381–96.

Kam, Cindy D., Jennifer R. Wilking, and Elizabeth J. Zechmeister. 2007. Beyond the 'narrow data base': Another convenience sample for experimental research. *Political Behavior* 29:415–40.

Kittur, Aniket, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. Proceedings of the 26th Annual CHI Conference on Human Factors in Computing Systems, CHI 2009, 453–6. New York: ACM Press.

Kuhberger, Anton. 1995. The framing of decisions: A new look at old problems. *Organizational Behavior & Human Decision Processes* 62:230–40.

Lawson, C., Gabriel S. Lenz, Mike Myers, and Andy Baker. 2010. Looking like a winner: Candidate appearance and electoral success in new democracies. *World Politics* 62:561–93.

Mason, Winter, and Duncan J. Watts. 2009. Financial incentives and the performance of crowds. Proceedings of the ACM SIGKDD Workshop on Human Computation, 77–85. New York: ACM Press.

Orne, M. T. 1962. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist* 17:776–83.

Paolacci, Gabriele, Jesse Chandler, and Panagiotis G. Ipeirotis. 2010. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5:411–19.

Rasinski, Kenneth A. 1989. The effect of question wording on public support for government spending. *Public Opinion Quarterly* 53:388–94.

Ross, Joel, Lily Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who are the crowdworkers? Shifting demographics in Amazon Mechanical Turk. In *CHI EA 2010*, 2863–72. New York: ACM Press.

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* 51:515–30.

Sheng, Victor S., Foster Provost, and Panagiotis G. Ipeirotis. 2008. Get another label? Improving data quality and data mining using multiple, noisy labelers. Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 614–22. New York: ACM Press.

Snow, Rion, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—But is it good? Evaluating non-expert annotations for natural language tasks. EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 254–63. Morristown NJ: Association for Computational Linguistics.

Sorokin, Alexander, and David Forsyth. 2008. *Utility data annotation with Amazon Mechanical Turk. Computer Vision and Pattern Recognition Workshops '08* 51:1–8.

Takemura, Kazuhisa. 1994. Influence of elaboration on the framing of decision. *Journal of Psychology* 128:33–9.

Transue, John E., Daniel J. Lee, and John H. Aldrich. 2009. Treatment spillover effects across survey experiments. *Political Analysis* 17:143–61.

Tversky, Amos, and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *Science* 211:453–8.