Survey Research Methods (2019)
Vol. 13, No. 2, pp. 215-227
doi:10.18148/srm/2019.v13i2.7262

© European Survey Research Association
ISSN 1864-3361
http://www.surveymethods.org

Can nonprobability samples be used for social science research? A cautionary tale

Elizabeth S. Zack
Indiana University
Bloomington, U.S.A.

John M. Kennedy
Indiana University
Bloomington, U.S.A.

J. Scott Long
Indiana University
Bloomington, U.S.A.

Survey researchers and social scientists are trying to understand the appropriate use of nonprobability samples as substitutes for probability samples in social science research. While cognizant of the challenges presented by nonprobability samples, scholars increasingly rely on these samples due to their low cost and speed of data collection. This paper contributes to the growing literature on the appropriate use of nonprobability samples by comparing two online non-probability samples, Amazon's Mechanical Turk (MTurk) and a Qualtrics Panel, with a gold standard nationally representative probability sample, the General Social Survey (GSS). Most research in this area focuses on determining the best techniques to improve point estimates from nonprobability samples, often using gold standard surveys or census data to determine the accuracy of the point estimates. This paper differs from that line of research in that we examine how probability and nonprobability samples differ when used in multivariate analysis, the research technique used by many social scientists. Additionally, we examine whether restricting each sample to a population well-represented in MTurk (Americans age 45 and under) improves MTurk's estimates. We find that, while Qualtrics and MTurk differ somewhat from the GSS, Qualtrics outperforms MTurk in both univariate and multivariate analysis. Further, restricting the samples substantially improves MTurk's estimates, but not enough to close the gap with Qualtrics. With both Qualtrics and MTurk, we find a risk of false positives. Our findings suggest that these online nonprobability samples may sometimes be "fit for purpose," but should be used with caution.

Keywords: Nonprobability Samples; Online Panels; MTurk; GSS

1 Introduction

Over the past ten years, social science researchers have begun testing and using non-traditional nonprobability samples as substitutes for random samples. Response rates for general population surveys have declined over the past twenty years to where single digit response rates are not unusual for telephone surveys (Keeter, Hatley, Kennedy, & Lau, 2017). The low response rate of surveys increases the possibility that nonresponse bias may affect survey results and has also resulted in a substantial increase in survey costs. These changing conditions have led survey researchers and other social scientists to examine the value of low-cost alternatives to the more expensive surveys with randomly selected samples.

Contact information: John M. Kennedy, 1700 East 10th St., Eigenmann Hall, Bloomington IN 47406 (email: kennedyj@indiana.edu).

The current research on the appropriate use of nonprobability samples indicates that the question is no longer if they can be used but rather how they can be used for social science research. This paper contributes to this research and in particular the research that compares nonprobability samples with gold standard probability samples. Specifically, we compare responses to questions asked in the 2014 General Social Survey (GSS) with two commonly used online nonprobability samples. Our primary goals are to determine the similarities and differences in univariate distributions and multivariable models between the low-cost nonprobability samples and the GSS, and to assess a potential method of improving nonprobability sample estimates.

Historically, random sampling for social surveys has been an important factor in the evaluation of research data quality, primarily because the inferential statistics used most often by social scientists are based on the assumption that participants are randomly selected from the target population. Standard statistical measures, such as confidence intervals, allow

readers to carefully evaluate a study's statistical conclusions. However, over the past 20 years, survey researchers have faced increasing challenges to the assumption that their sample surveys can be considered random selections from the target populations. Most current sample surveys have low response rates and while many survey researchers assume that the non-respondents are missing at random or missing completely at random, appropriate measures are not available to carefully evaluate those claims. The costs of random population surveys along with the non-response challenges have made nonprobability samples more attractive.

Social science researchers have a number of sources for low cost nonprobability samples. The most commonly used sampling source is Amazon's Mechanical Turk (MTurk). MTurk is an Amazon web service that is often used by social scientists due to its low cost and the ease of recruiting participants. MTurk is an opt-in online platform where "requesters" can post small tasks, such as translation work or taking surveys, and workers complete those tasks for small amounts of money. While researchers have found consistent evidence that MTurk samples provide high-quality experimental data (e.g. Behrend, Sharek, Meade, & Wiebe, 2011; Berinsky, Huber, & Lenz, 2012; Buhrmester, Kwang, & Gosling, 2011; Coppock, 2018; Gamblin, Winslow, Lindsay, Newsom, & Kehn, 2017; Mullinix, Leeper, Druckman, & Freese, 2015; Paolacci, Chandler, & Ipeirotis, 2010), it is less clear whether they produce useful data for non-experimental survey research.

Other sources of low-cost nonprobability samples include organizations such as Qualtrics and Survey Sampling International (SSI), which offer nonprobability samples that use a variety of opt-in methods to populate panels. These samples differ from MTurk because they are panel-based and selected using a range of proprietary sampling methods. Additionally, researchers using Qualtrics Panels can request quota sampling to ensure that the sample represents the demographic characteristics of a population. While MTurk allows users to include participants based on certain demographic criteria for an additional fee, this method relies on the researcher to match the sample to national demographics themselves, and the universe of potential participants may not closely reflect the U.S. population. For example, it may be difficult to recruit enough older participants to match national demographics (Heen, Lieberman, & Miethe, 2014), and older individuals who do choose to use MTurk may differ meaningfully from the population on variables of interest to many social scientists, including party identification, political ideology, and voting behavior (Huff & Tingley, 2015). As such, there is reason to believe that, while more expensive, Qualtrics Panels may be more suitable to non-experimental social science research compared to MTurk.

To reiterate, while both MTurk and opt-in panels offer relatively low-cost, easily accessible samples, neither are random samples from the targeted populations. Therefore, the samples are not technically appropriate for inferential statistics. While the same statement might be made about low response rate probability samples, the potential for biases with the opt-in samples is much greater. At the same time, these biases may be substantively minimal, and overall results may be similar when using probability and nonprobability samples.

While previous research found that online nonprobability samples were effective tools for survey experiments, our goal in this paper is to determine 1) whether these samples are useful in conducting non-experimental research, which is more common in the social sciences, 2) whether Qualtrics Panels, due to their sampling technique and demographic characteristics, perform better than MTurk samples, and 3) whether restricting nonprobability samples to the supports of the demographic data can improve estimates; i.e., while an MTurk sample may not be reflective of the U.S. population as a whole, it may be somewhat more representative of, for example, younger Americans, who are more likely to opt-in to the sample.

2 Similar Studies

This study builds on three recent studies that compare low cost nonprobability samples to gold standard samples to evaluate their utility for social science research. First, Mullinix et al. (2015) compare the results of twenty survey experiments conducted with MTurk workers to the results of the same experiments administered via Timesharing Experiments for the Social Sciences (TESS). TESS is conducted using the GFK Knowledge Panel, an online probability-based nationally representative panel. Mullinix et al. found that 80.6% of the treatment effects in the TESS experiments were replicated in the MTurk samples. While this study found that online nonprobability samples were suitable alternatives to nationally representative probability samples for conducting experiments, it is less clear whether nonprobability samples are suitable for attitudinal social science research.

The second study, conducted by Simmons and Bobo. (2015), compared the results from the 2009 Race Cues, Attitudes, and Punitiveness Survey (RCAPS) with the 2008 General Social Survey (GSS) and the 2008 American National Election Study (ANES). The RCAPS sample was developed using a sample matching process. In the first step, a stratified random sample from the 2006 American Community Survey was drawn. Strata based on relevant demographic characteristics drawn from the probability sample were recreated in an opt-in online panel. In that way, the online sample matched the population sample on the most relevant characteristics for the research. The authors compared the demographic distributions of the three samples with the 2006 Current Population Survey (Simmons and Bobo. (2015), Table 1) and found that each samples' differences from the CPS

averaged less than 4%. While some differences were substantial and statistically significant, for a number of political attitudes questions differences between the ANES, GSS, and RCAPS were statistically different but not necessarily substantively different. Simmons and Bobo. (2015) also compared multivariable models across the datasets. They found that when the dependent variable is more concrete, the models differ less across the samples than when the dependent variable is more abstract. In general, the model differences were not substantially different.

A third paper that exemplifies the type of testing that is being done to determine the appropriate use of nonprobability samples compared the 2012 ANES with an MTurk sample administered in early 2013 (Levay, Freese, & Druckman, 2016). The differences in the distributions of the samples were consistent with previous research. For example, MTurk workers are younger, more educated, and more liberal than samples drawn from a general population sample. The most appropriate finding relevant for this paper is that the MTurk sample is statistically different from the ANES sample for most variables but the substantive impact of the differences is relatively small.

3 Nonprobability Samples

Many researchers have documented that the demographic composition of MTurk workers is very different from a random sample drawn from the U.S. adult population (e.g. Berinsky et al., 2012; Heen et al., 2014; Levay et al., 2016; Paolacci et al., 2010; Shapiro, Chandler, & Mueller, 2013). However, researchers have also noted that randomization to experimental conditions allows MTurk samples to be less affected by their demographic composition. Numerous studies conclude that for social science experiments, the internal validity of nonprobability samples is the equivalent or better than traditional experimental pools of undergraduate students.

In one of the most cited papers on the use of MTurk samples for experimental research, Berinsky et al. (2012) replicate seminal political science experiments using MTurk. They note that the estimates of average treatment effects are similar in the MTurk and original samples. They also found that the potential limitations to using MTurk to recruit subjects and conduct research are tempered by potential benefits. For example, while MTurk subjects are 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. Similarly, Buhrmester et al. (2011) conclude that MTurk participants are more demographically diverse than both undergraduate student samples and other online convenience samples and that the data obtained are at least as reliable for experimental research as those collected via traditional methods. Overall, researchers find consistent evidence that MTurk can be used to obtain high-quality experimental data inexpensively and rapidly.

Survey researchers are aware of the current challenges related to the use of online nonprobability panels and other online opt-in samples. For example, the American Association for Public Opinion Research task force (Baker et al., 2013) analyzed the challenges encountered when using online nonprobability samples for high quality survey research. And in 2014, a book on the use of online panels in survey research (Callegaro et al., 2014) assessed both the challenges and the appropriate use of online panels. Some research has found that, while online nonprobability samples may not be appropriate for obtaining point estimates (Kennedy et al., 2016), they may be appropriate for experimental research and for modeling relationships between variables (Groves, 2004). Nonprobability samples are likely to continue to be used, and more research is needed to determine whether they provide adequate data for the non-experimental survey research that social scientists more typically conduct. Further, more research is needed to determine how online nonprobability samples can be used appropriately in non-experimental research; for example, while these samples may not be representative of the U.S. population as a whole, they may be more representative of certain segments of the U.S. population.

4 Research Focus

This paper contributes to the research on the appropriate use of online panels in three ways: 1) we compare the data quality of two different types of nonprobability online samples; 2) we assess outcome variables from a wide range of subfields of interest to social scientists; and 3) we assess conditions under which nonprobability samples may perform best; specifically we test and find promise in a method to improve MTurk's estimates by restricting the sample to better fit a specific, well-represented population. By benchmarking online nonprobability samples to an established gold-standard survey, we can help define their appropriate use and identify potential pitfalls.

5 Data and Measures

5.1 Data

We compare three datasets in our analyses: the 2014 General Social Survey, an online nonprobability samples of MTurk workers collected in 2015, and a nonprobability sample from a Qualtrics Panel sample also collected in 2015. Our main goal is to determine which of the online nonprobability samples best approximates the nationally representative GSS. If we assume that the GSS, considered the gold-standard survey, accurately measures the variables in the surveys, then differences between the GSS and the nonprobability samples indicate that researchers should be more cautious when using the nonprobability samples. Ideally, all

data would have been collected during the same time period. Since political and social attitudes tend to change slowly over time (Page & Shapiro, 2010), the one to two-year gap between the collection of the probability samples and the nonprobability samples should not affect our analyses. Further, we chose questions for analysis that reflect more enduring attitudes and opinions that should not be strongly related to current events.

GSS. The GSS is a nationally representative probability sample of English- and Spanish-speaking households in the United States, conducted annually or bi-annually since 1972. Respondents are 18 years of age or older. The 2014 GSS dataset was collected from February through April 2014, through face-to-face interviews. In total, 2,538 interviews were completed with a response rate of 69 percent.

Mechanical Turk. The MTurk sample is an online nonprobability sample of MTurk workers residing in the United States who are age 18 or older. MTurk respondents were recruited through a task posted to the MTurk website, titled "Political and Social Opinions Survey". Respondents were paid $1.50 for their participation in a 15-minute survey. Four-hundred seventy MTurk workers completed the survey. Following survey completion, we re-contacted the original respondents to ask two additional demographic questions. The follow-up survey was completed by 355 of the original respondents. We use the 355 respondents who completed both the original and the follow-up study as our final sample. Sensitivity analyses showed minimal differences in other demographics between respondents who did and did not respond to the follow-up survey.[1] The original MTurk survey was conducted in March 2015. The follow-up survey was conducted in April 2015. MTurk workers were required to have an approval rating of 95% or higher to participate. Data was collected over several days and at varying times of day because the characteristics of MTurk workers may vary by day and time (Arechar, Kraft-Todd, & Rand, 2017; Casey, Chandler, Levine, Proctor, & Strolovitch, 2017).

Qualtrics Panel. The Qualtrics Panel is an online nonprobability panel provided by Qualtrics, a research software company. Using Qualtrics Panels, researchers are able to build panels to their desired specifications. The Qualtrics panel used in this study was created to approximately match national demographics of the adult population on age, race, and gender. This Qualtrics panel was conducted in April 2015, and 547 respondents completed the survey. Qualtrics makes available panels with more sample stratification variables at a higher cost but we decided that a simpler stratification plan would be more comparable to other low-cost samples.

Mode Differences. Data from the two nonprobability samples were collected online while the GSS was collected in-person. Thus, differences between the GSS and the nonprobability samples could be due to mode differences. For example, people might respond differently during face-to-face interviews than they do via online survey for social desirability reasons. To reduce the possibility of mode effects, we selected non-sensitive questions.

Weighting. We did not use weights in the analyses for two reasons. First, some evidence shows that weights do not always reduce the bias of nonprobability samples (e.g. Tourangeau, Conrad, & Couper, 2013, p. 33). A recent Pew report stated among its key findings that even the most effective adjustment procedures were unable to remove most bias (Mercer, Lau, & Kennedy, 2018, p. 3). Second, most social science researchers do not weight the data when using multivariate models. Our goal was to examine how typical researchers would use the three data sources.

5.2 Measures

We examined 22 attitudinal questions that are often used in sociological analysis. These included questions on a range of topics, including health and well-being, science and technology, gender and family, redistributive policy, social mobility, criminology, racial attitudes, religion, and knowledge-based questions. For a full list of the outcome variables, see Table 1. To avoid bias due to question order, we selected questions that were first in a series of similar questions or were stand-alone questions.[2] These variables are the outcome variables in our analyses. Due to the small samples in MTurk and Qualtrics we dichotomized each variable. This allows us to use binary logit in the analyses, perhaps the most commonly used models for analyzing attitudinal questions. The independent variables are standard demographic variables: age, gender, region, ethnicity, race, education, marital status, political ideology, party identification, and household income (see Table 2 for descriptive statistics).

We fielded the demographic and outcome variables on MTurk and Qualtrics using the original GSS wording for each variable, with minor modifications to make the questions appropriate for online data collection. Two exceptions are the marital status and income variables, which have slightly different response options between the nonprobability samples and the GSS. For marital status, the nonprobability samples include a category for "Living with a partner",

[1] The two samples differ significantly on marital status and age. People living with a partner or widowed were more likely to respond to the follow-up survey, while people who were never married were less likely to respond. Older people were more likely to respond to the follow-up survey than were young people. Because MTurk samples tend to be younger than U.S. population averages, the mean age of respondents in our final sample is closer to the mean age of the GSS sample.

[2] The only exception is the GSS variable *abany* which was asked last in a series of more specific questions on abortion. It was selected because it was the broadest question on abortion, asking "whether or not you think it should be possible for a woman to obtain a legal abortion if the woman wants it for any reason."

Table 1
Outcome Variables

Variable Name	Description	Coding
ABANY	Do you think it should be possible for a pregnant woman to obtain a legal abortion if the woman wants it for any reason?	1=Yes; 0=No
ADVFRONT	Even if it brings no immediate benefits, scientific research thatadvances the frontiers of knowledge is necessary and should be supported by the federal government.	1=Strongly agree; 0=Does not strongly agree
BIBLE	Which of these statements comes closest to describing your feelings about the Bible: 1) the Bible is the actual word of God and is to be taken literally, word for word; 2) the Bible is the inspired word of God but not everything in it should be taken literally, word for word; 3) the Bible is an ancient book of fables, legends, history, and moral precepts recorded by men.	1=Literal or inspired word of God; 0=Ancient book of fables
CAPPUN	Do you favor or oppose the death penalty for persons convicted of murder?	1=Favor; 0=Oppose
COURTS	In general, do you think the courts in this area deal too harshly or not harshly enough with criminals?	1=Too harshly; 0=Not harshly enough
ELECTRON	Electrons are smaller than atoms.	1=True; 0=False or Don't know
FECHLD	A working mother can establish just as warm and secure a relationship with her children as a mother who does not work.	1=Strongly agree; 0=Does not strongly agree
GETAHEAD	Some people say that people get ahead by their own hard work; others say that lucky breaks or help from other people are more important. Which do you think is more important?	1=Hard work; 0=Luck or Both equally
HAPPY	Taken all together, how would you say things are these days–would you say that you are very happy, or not too happy?	1=Very happy; 0=Not very happy
HELPFUL	Would you say that most of the time people try to be helpful, or that they are mostly just looking out for themselves?	1=Try to be helpful; 0=Looking out for themselves
HELPNOT	Some people think that the government in Washington is trying to do too many things that should be left to individuals and private businesses. Others disagree and think that the government should do even more to solve our country's problems. Still others have opinions somewhere in between. Where would you place yourself on this scale, or haven't you made up your mind on this?	1=Government should do more or Neutral; 0=Government does too much
HELPPOOR	Some people think that the government in Washington should do everything possible to improve the standard of living of all poor Americans. Other people think it is not the government's responsibility, and that each person should take care of himself. Where would you place yourself on this scale, or haven't you made up your mind on this?	1=Government should do everything possible to help poor; 0=Neutral or People should take care of themselves
HOTCORE	The center of the earth is very hot.	1=True; 0=Don't know or False
LASERS	Lasers work by focusing sound waves.	1=False; 0= Don't know or True
LIFE	In general, do you find life pretty exciting, routine, or dull?	1=Exciting; 0=Not exciting
NEXTGEN	Because of science and technology, there will be more opportunities for the next generation.	1=Strongly agree; 0=Does not strongly agree

Continues on next page

Continued from last page

Variable Name	Description	Coding
ODDS1	A doctor tells a couple that their genetic makeup means that they've got one in four chances of having a child with an inherited illness. Does this mean that if their first child has the illness, the next three will not have the illness?	1=Yes; 0=No
PILLOK	Methods of birth control should be available to teenagers between the ages of 14 and 16 if their parents do not approve.	1=Yes; 0=No
RADIOACT	All radioactivity is man-made.	1=False; 0=Don't know or True
SEXEDUC	Would you be for or against sex education in the public schools?	1=For; 0=Against
TOOFAST	Science makes our way of life change too fast.	1=Yes; 0=No
WRKWAYUP	Irish, Italians, Jewish and many other minorities overcame prejudice and worked their way up. Blacks should do the same without special favors.	1=Agree; 0=Disagree

which the GSS variable does not include; to make the variable equivalent across samples, "Living with a partner" was collapsed into the "Never married" category. For income, the 25-category GSS variable was collapsed into the 7-category income variable that was fielded in the Qualtrics and MTurk surveys. While most of the GSS categories fit cleanly into the 7-category version, two of the upper-income categories differed by $10,000.

6 Analytic Strategy

To compare MTurk and Qualtrics to the GSS standard, we began with chi-square tests to assess whether MTurk and Qualtrics respondents have demographics, beliefs, attitudes and knowledge that are similar to the GSS respondents. Our next step was to evaluate whether substantive conclusions from regression models were consistent across the three datasets using the methods described in (Long & Mustillo, forthcoming) for comparing groups using logit models. In our application, the data source is the group variable. Group differences in unobserved heterogeneity invalidate traditional tests for comparing regression coefficients across groups (Allison, 1999). While Allison (1999), Williams (2009) present tests that account for differences in unobserved heterogeneity and Breen, Holm, and Karlson (2014) develop methods based on correlations between the latent outcome and each regressor, we wanted to compare effects that were measured in the metric of the probability of the outcome variable. This was done by fitting a regression model in which the regression coefficient of each independent variable is allowed to differ across groups. From these estimates, the average discrete change for each regressor on each outcome (ADCs) was computed for each group. The ADC estimates the average change in the probability of the outcome for a discrete change in an independent variable. For example, in the GSS, on average being white decreases the probability of being happy by .068. After ADCs were estimated, we tested if the effects were equal in each sample. For example, we tested if the ADC for race on being happy is the same in the GSS sample as the MTurk sample.

Since the ADCs are computed by averaging over the sample (e.g., what is the average effect of race on being happy in the GSS), they reflect the distribution of independent variables. Accordingly, if the ADC for a regressor differs between GSS and one of the nonprobability datasets, it could be due to differences in the distribution of regressors across samples even if the regression coefficients were identical. This possibility is addressed in two ways. First, we compute discrete changes at representative values (DCRs). Instead of an average of values, DCRs compute the change in the probability as one regressor changes by a specified amount, holding other variables at specific values. Often, the mean is used as the representative values. For our comparisons, we made comparisons at *age*=35, *polviews*=Moderate, *partyid*=Democrat, with other variables held at the GSS means. Note that GSS means were used for all three samples.

Given that MTurk samples are notably younger than the U.S. population (Berinsky et al., 2012; Levay et al., 2016; Paolacci et al., 2010) and older MTurk workers may differ meaningfully from the population (Huff & Tingley, 2015), we created restricted samples that included only respondents who were between 18and 45 years old (about 85% of our MTurk sample). This should make the comparisons between the three samples less dependent on the lack of representativeness of the nonprobability samples. ADCs and DCRs were then computed using the restricted sample.

Computations were completed using Stata 15.1 (StataCorp, 2015) using SPost commands (Long & Freese, 2014).

Table 2
Descriptive Statistics and Chi-square Tests of Differences Between Samples for Demographic Variables

	Mean			t/χ^2	
	GSS	MTurk	Qualtrics	GSS/MTurk	GSS/Qualtrics
Female	0.54	0.46	0.51	9.67^{**}	2.31
Non-Hispanic White	0.66	0.78	0.76	19.77^{***}	16.74^{***}
Bachelor's Degree	0.33	0.46	0.36	23.19^{***}	1.77
Age	48.73 (17.02)	34.61 (10.61)	45.90 (17.17)	-15.09^{***}	-3.38^{***}
South	0.37	0.36	0.40	0.09	1.50
Marital Status				186.70^{***}	41.25^{***}
Married	46.05	34.76	43.95	-	-
Widowed	7.59	0.57	3.52	-	-
Divorced	16.80	2.85	12.30	-	-
Separated	3.14	1.42	1.56	-	-
Never Married	26.42	60.40	38.67	-	-
Household Income[1]				59.58^{***}	71.36^{***}
$7,500	14.56	11.68	11.52	-	-
$20,000	11.86	15.67	12.70	-	-
$30,000	11.32	15.67	12.50	-	-
$42,500	11.81	18.23	16.60	-	-
$62,500	18.96	20.80	20.12	-	-
$87,500	7.55	9.97	15.23	-	-
$125,000	14.78	6.27	8.40	-	-
$200,000	9.16	1.71	2.93	-	-
Political Ideology				212.18^{***}	63.80^{***}
Extremely Liberal	3.95	20.80	8.79	-	-
Liberal	12.76	23.36	14.65	-	-
Slightly Liberal	10.87	13.68	11.52	-	-
Moderate	39.89	17.95	28.52	-	-
Slightly Conservative	13.70	10.26	16.80	-	-
Conservative	14.65	9.40	10.55	-	-
Extremely Conservative	4.18	4.56	9.18	-	-
Party Identification				22.92^{***}	34.64^{***}
Democrat	33.78	41.88	38.48	-	-
Republican	22.01	17.38	25.00	-	-
Independent	41.73	35.04	30.47	-	-
Something Else	2.47	5.70	6.05	-	-
Religion				212.51^{***}	154.22^{***}
Protestant	45.24	15.10	26.76	-	-
Catholic	24.03	15.38	24.80	-	-
Other	10.15	17.38	30.08	-	-
None	20.58	52.14	18.36	-	-
N	2226	351	512		

$^{*}\ p \leqslant 0.05$, $^{**}\ p \leqslant 0.01$, $^{***}\ p \leqslant 0.001$
Note: Standard errors in parentheses. χ^2 statistics are displayed for all indicator variables; *t*-statistics are displayed for continuous variables (age).
[1] Set to category midpoints.

7 Results

7.1 Differences in analysis variables between samples

In our analysis of demographic variables, we use all cases except those missing on any of the ten demographic variables. The resulting GSS sample has 2,226 cases, MTurk has 351 cases; and Qualtrics has 512 cases. Demographic variables in MTurk significantly differ from those in the GSS on gender ($p < 0.01$), race/ethnicity, education, age, marital status, income, political ideology, party identification, and religion ($p < 0.001$). In particular, the MTurk sample is less female, more educated, younger, and more liberal than the GSS sample. There is no significant difference in region. Similarly, Qualtrics significantly differs from the GSS sample on race/ethnicity, age, marital status, income, political ideology, party identification, and religion ($p < 0.001$). However, substantively the sizes of these differences are relatively small. There are no significant differences in gender, education, or region. (See Table 2).

In the remainder of our analyses, we reduce our samples to cases that are not missing on the outcome variable or the demographic variables. The size of the resulting samples vary by the outcome variable. GSS sample size ranges from 1,012 to 2,223 (because some outcome variables were not asked of all respondents); MTurk from 329 to 351; and Qualtrics from 472 to 512. We compare outcome variables across the three data sources in Figure 1. For 19 out of 22 outcome variables, the MTurk variables differ significantly from the GSS (including *happy, life, nextgen, toofast, advfront, fechld, pillok, abany, helppoor, helpnot, getahead, cappun, courts, wrkwayup, hotcore, radioact, lasers, electron,* and *bible*). Seventeen differ significantly at the .001 level, one at the .01 level, and two at the .05 level. There are no significant differences for *helpful*, *sexeduc*, or *odds1*. By comparison, Qualtrics distributions significantly differ from GSS distributions on only 8 out of the 22 outcome variables (including *happy, life, advfront, fechld, sexeduc, abany, getahead,* and *electron*), each at the .001 level. Differences were not significant for *helpful, nextgen, toofast, pillok, helppoor, helpnot, cappun, courts, wrkwayup, hotcore, radioact, lasers, odds1,* or *bible*.

7.2 Logit analysis—full sample

Average Discrete Change. To assess whether demographic variables are associated with outcome variables in similar ways across datasets, we estimate ADCs for each demographic variable on each outcome variable. First, we estimate separate models for each sample. For example, we estimate the average discrete change of political ideology on *abany*, the belief that a woman should be able to have an abortion for any reason, controlling for all other demographic variables (see Table 3).

In this example, we find that in the GSS, on average a one unit increase in conservatism is associated with a .110 decrease in the probability of supporting abortion (p<.001), in Qualtrics we estimate a .116 decrease in support (p<.001), while in MTurk, there is a small, positive, non-significant effect. We then tested whether these ADCs differ significantly across samples. In the *abany-polviews* example, MTurk's ADC is significantly larger than the effect in the GSS (p<.001), while the effects in Qualtrics and GSS are not significantly different (see Table 4).

Using the method illustrated above, we estimate and test ADCs for six demographic variables (*age, white, female, education, income,* and *polviews*), controlling for religion, party identification, marital status, and region, on 22 outcome variables, for a total of 132 ADCs for each sample. Of these 132 ADCs, 38 (28.8%) differ significantly between GSS and MTurk, while 14 (10.6%) differ significantly between GSS and Qualtrics (see Table 5). Online Appendix table A1 includes results for each ADC. These findings suggest that Qualtrics models may better approximate the relationships between demographic and outcome variables found in the GSS than do MTurk models.

Discrete Change at Representative Values. Next, we estimate the discrete change at representative values holding *age*=35, *polviews*=Moderate, *partyid*=Democrat, and all other variables at the GSS means. We then test if the effects differ between samples. As explained in the Methods Section, DCRs do not reflect differences in the distribution of regressors across the samples since they are computed at the same values of the regressors. As with the ADCs, we estimated 132 DCRs. Of these, 28 (21.2%) significantly differ between GSS and MTurk, while 13 (9.8%) significantly differ between GSS and Qualtrics (Table 5). Online Appendix table A2 includes results for each DCR. Again, these findings suggest that the Qualtrics sample performs better than the MTurk sample, even when accounting for differences in demographic distributions between samples.

For both ADCs and DCRs, when demographic variables significantly predict outcome variables in the GSS and a nonprobability sample, the direction of the effects is the same in both samples in the overwhelming majority of cases. For example, when the effects were significant in both samples, the direction of the ADC was the same in 17 of 20 cases (85%) for MTurk and in all 29 cases for Qualtrics. However, when the ADC for demographic variables were significant in the nonprobability samples, they were only significant in the GSS 57% of the time for MTurk and 78% of the time for Qualtrics, suggesting that the use of the nonprobability samples creates a risk of false positives.

7.3 Logit analysis—age-restricted sample

Since the greater number of significant differences in MTurk, compared to Qualtrics, may simply reflect the truncated distributions of demographic variables in the MTurk sample, particularly age, we repeat the previous steps on the

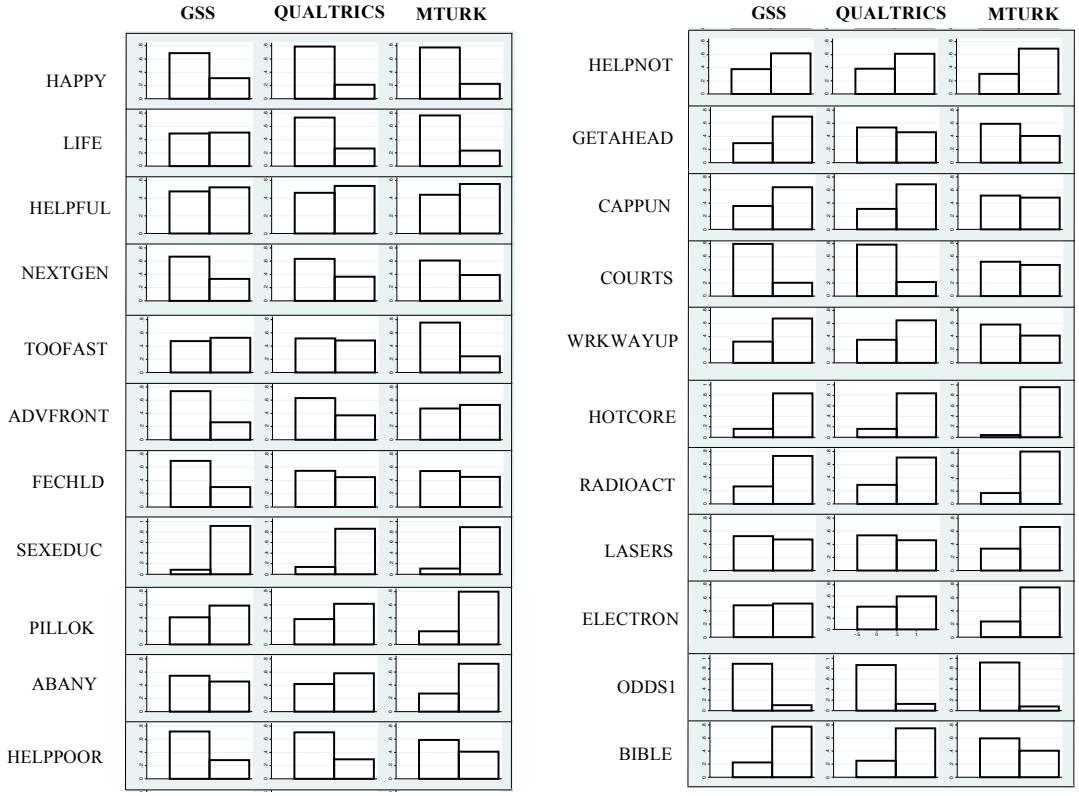

Figure 1. A Comparison of Outcome Variable Distributions, by Sample

Table 3
ADCs of Political Ideology on Support for Abortion for by Sample

Outcome *Abany*	GSS	MTurk	Qualtrics
ADC(*polviews*)	-0.110^{***}	0.032	-0.116^{***}
Std. Err.	(0.013)	(0.028)	(0.022)

$^{*}\ p \leqslant 0.05, \quad ^{**}\ p \leqslant 0.01, \quad ^{***}\ p \leqslant 0.001$

Table 4
Differences in ADCs of Political Ideology on Support for Abortion Between Samples

Outcome *Abany*	GSS-MTurk	GSS-Qualtrics	MTurk-Qualtrics
ADC(*polviews*)	-0.142^{***}	0.006	0.148^{***}
Std. Err.	(0.031)	(0.025)	(0.036)

$^{*}\ p \leqslant 0.05, \quad ^{**}\ p \leqslant 0.01, \quad ^{***}\ p \leqslant 0.001$

Table 5
Differences in ADCs and DCRs between Samples – Full Sample

	ADCs		DCRs	
	MTurk (%)	Qualtrics (%)	MTurk (%)	Qualtrics (%)
% of effects that differ at .05 level from GSS	28.8	10.6	21.2	9.8

age-restricted samples. About 85% of our MTurk sample is between the ages of 18–45 (comparable to typical MTurk samples, which have significantly younger respondents than nationally representative samples). As such, we repeated our analyses using only respondents ages 18 to 45 for each sample. This left us with a sample size of 478 to 1,017 for GSS, 276 to 295 for MTurk, and 232 to 253 for Qualtrics. Restricting the age of the sample leads to estimates in the MTurk sample that are closer to those for the GSS (Table 6). As the Qualtrics sample was selected to reflect the national age distribution, we did not expect the age restriction to improve Qualtrics estimates, which it did not.

Using the age restricted sample reduced the number of significantly different ADCs between GSS and MTurk from 38 to 22 (28.8% to 16.7%), representing a 42% decrease in significant differences. Similarly, significant differences in DCRs between GSS and MTurk decreased from 28 to 15 (21.2% to 11.4%), representing a 46% decrease. As expected, the percentage of Qualtrics ADCs and DCRs that significantly differed from the GSS did not substantively change between the full and age-restricted samples. As such, the age-restricted sample narrowed the apparent advantage of using Qualtrics over MTurk from about an 18-percentage point advantage to a single point advantage for ADCs, and from about an 11-percentage point advantage to a 4-percentage point advantage for DCRs, almost eliminating the difference with Qualtrics by over half.

8 Discussion

8.1 Variable distributions

In line with previous research, we found that MTurk demographics differed significantly from GSS demographics; specifically, the MTurk sample was younger, less female, more educated, and more liberal. Further, despite sample stratification in the Qualtrics Panel to approximate national distributions on age, race, and gender, the Qualtrics Panel also differed significantly from the GSS on several demographics. However, Qualtrics differed less often than did MTurk, and substantively these differences were relatively small.

Similarly, we found that the distributions of outcome variables in the Qualtrics sample better approximated the GSS samples than did the MTurk distributions. Qualtrics distributions differed significantly 33% of the time, less than half as often as the MTurk distributions did. Even when Qualtrics distributions significantly differed from the GSS, a cursory visual assessment (Figure 1) illustrates that the Qualtrics sample is a substantively closer match to the GSS than is the MTurk sample. This finding is likely to be explained in part by the difference in sampling methods between MTurk and Qualtrics (a convenience sample versus the creation of a panel matching national distributions for several demographic characteristics). As expected due to the liberal lean of the MTurk sample, when MTurk distributions differed from the GSS, it was universally in a more liberal direction.

Our results suggest that even without the weighting methods that scholars are developing to adjust nonprobability sample point estimates to those of gold standard probability samples (often with mixed results; e.g. Kennedy et al., 2016), distributions are somewhat similar between Qualtrics and GSS. Many sociologists and other social scientists do not use complicated weighting techniques, and our research suggests that, for univariate analysis, certain nonprobability samples might be used with appropriate cautions.

8.2 ADCs and DCRs

To use a nonprobability online sample with all of its obvious limitations, researchers need to assume that the missing data from nonrespondents, those who do not join panels, those without internet access, etc. are missing at random (see Heitjan & Basu, 1996). That is, the differences between the online panel participants and others can be adjusted by using some statistical method. Based on the assumption that the nonprobability sample responses are missing at random and that demographic characteristics can be used to adjust the sample for nonresponse and other errors, we tested whether the multivariate models from probability and nonprobability samples that used demographic variables as predictors would be roughly similar.

Estimating ADCs and DCRs, we determined how often effects significantly differed between the GSS and the two nonprobability samples, and found that Qualtrics was statistically indistinguishable from the GSS about 90% of the time, while MTurk was indistinguishable from the GSS about 70% of the time for ADCs and about 80% of the time for DCRs. Qualtrics' better performance may simply reflect the fact that MTurk samples tend to lack demographic data in certain parts of the curve (e.g., respondents over 45 years old), and the possibility that, while MTurk may be somewhat representative of younger Americans, older adults who choose to use MTurk may differ meaningfully from those who do not (Huff & Tingley, 2015). In other words, while MTurk may not be representative of the United States population as a whole, it may successfully approximate certain groups, such as, 18-45 year olds, within the U.S. population. Thus, we re-examined each sample, restricting age to the supports of the MTurk data, to make apples-to-apples comparisons across samples. Restricting the samples improved MTurk's ADCs from differing from GSS 29% of the time to 17% and MTurk's DCRs from 21% to 11%, and eliminated the advantage of using Qualtrics over MTurk by about half. Nonetheless, even with the restricted samples, Qualtrics continued to outperform MTurk.

Future research may want to examine other ways in which

Table 6
Differences in ADCs and DCRs between Samples – Age Restricted Sample

	ADCs		DCRs	
	MTurk (%)	Qualtrics (%)	MTurk (%)	Qualtrics (%)
% of effects that differ at .05 level from GSS	16.7	11.4	11.4	10.6

MTurk samples may be more representative of a population. For example, researchers may want to restrict the sample on political ideology. In our analysis, we found that *polviews* was a particularly problematic predictor in MTurk, accounting for about 25% of the significant differences in ADCs and DCRs between MTurk and the GSS. As MTurk samples are notably more liberal than the U.S. population, limiting samples not only by age but also to "moderate" to "extremely liberal" respondents (about 75% of our sample) may better reflect a segment of the U.S. population. In Qualtrics, education stood out as a particularly problematic predictor. Although Qualtrics did not significantly differ from the GSS in the distribution of education, education accounted for over 20% of significant differences in ADCs and DCRs between Qualtrics and the GSS, suggesting that education may function differently in Qualtrics than it does in gold-standard nationally-representative datasets. Future research should examine whether this is a consistent finding or an anomaly in our sample.

While we attempted to explore a broad range of outcome variables, we did not cover everything. Future research may want to explore additional topics, as well as specific topics in more depth (e.g., Simmons and Bobo's recent work on the use of nonprobability samples for assessing racial attitudes). Further, this paper explored only binary outcome variables; researchers may want to examine whether findings are similar with different types of outcome variables, such as continuous and categorical variables. An additional limitation of this study was the conditional selection of the MTurk sample based on MTurk workers' completion of both the initial survey and a two-question follow-up survey. While 76 percent of the initial MTurk sample completed the follow-up survey, the MTurk workers who did so may be particularly engaged with the MTurk platform, potentially introducing bias.

8.3 Conclusion

Our findings suggest that Qualtrics performs better than MTurk for both univariate and multivariate social science research. However, researchers may continue to be interested in using MTurk data due to its substantially lower costs. Our research suggests that making small adjustments to MTurk samples can improve Murk data; it appears that restricting MTurk samples on key demographics such as age may make the sample more representative of a segment of the U.S. population. Restricting the sample in this way requires conducting a short initial survey to exclude certain respondents, paying an extra fee for Amazon's "Premium Qualifications" to exclude people outside of the desired categories, or collecting a full sample and then dropping unwanted cases from the analysis; thus, it will cost a bit more than fielding a traditional MTurk survey. However, our results suggest that the dataset produced will be more informative about the corresponding segment of U.S. population. Even with these increased costs, MTurk is likely less expensive than a Qualtrics Panel, and may be preferred for researchers fielding a survey on a tight budget. While our research suggests a means of reducing the advantage of Qualtrics over MTurk, and future research may make further improvements, as of now these findings nonetheless point to using Qualtrics rather than MTurk, if the researcher's budget allows. Overall, these findings are exploratory, and more work is needed to narrow the gap between online nonprobability samples and gold standard probability surveys. Although Qualtrics effects only differed from the GSS about 10% of the time (not much greater than the 5% one would expect to find due to chance), and the direction of effects were overwhelmingly the same when examining effects that were significant in both samples, our analyses also showed the possibility for a substantial number of false positives in both MTurk and Qualtrics. While there is a consensus that online nonprobability samples are promising tools for experimental research, researchers should use care when conducting attitudinal research. Both MTurk and Qualtrics may be useful for specific purposes, but should be used with care. The intention of this paper is to provide a note of caution rather than a definitive answer regarding the utility of nonprobability samples in social science research. As other researchers conduct similar tests, we hope that a greater understanding of what social scientists can and cannot learn from nonprobability samples will be gained.

Our research falls in line with the similar studies on nonprobability samples discussed in our literature review. That is, the nonprobability samples are not so far from the gold standard samples that they never can be used. Just as a researcher rarely knows if a random sample is not representative of a population, the accuracy of a nonprobability sample similarly cannot be assessed. While they may be functionally equivalent at times, our analysis indicates that a researcher might come to different conclusions when using probability and nonprobability samples. While probability samples are very costly and have low response rates, they likely still gen-

erate more accurate results than nonprobability samples.

We recognize that not all survey samples need to be perfect and that "fitness for use" (Biemer, 2010) is also an important criterion to evaluate surveys. For example, survey experiments do not necessarily need truly random samples to detect differences. Additionally, Qualtrics Panels and other large nonprobability panels offer researchers an inexpensive method of surveying subpopulations (e.g., the Muslim population in the US), whose data would be very difficult and expensive to obtain using random samples. Low-cost nonprobability samples can be used in many ways to improve our understanding of the social world, and they should not be dismissed as a legitimate tool for social scientists.

References

Allison, P. D. (1999). Comparing logit and probit coefficients across groups. *Sociological Methods & Research*, *28*(2), 186–208.

Arechar, A. A., Kraft-Todd, G. T., & Rand, D. G. (2017). Turking overtime: How participant characteristics and behavior vary over time and day on Amazon Mechanical Turk. *Journal of the Economic Science Association*, *3*(1), 1–11.

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., … Tourangeau, R. (2013). Summary report of the AAPOR task force on nonprobability sampling. *Journal of Survey Statistics and Methodology*, *1*(2), 90–143.

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods*, *43*(3), 800–813.

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

Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, *74*(5), 817–848.

Breen, R., Holm, A., & Karlson, K. B. (2014). Correlations and nonlinear probability models. *Sociological Methods & Research*, *43*(4), 571–605.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5.

Callegaro, M., Baker, R. P., Bethlehem, J., Göritz, A. S., Krosnick, J. A., & Lavrakas, P. J. (Eds.). (2014). *Online panel research: A data quality perspective*. New York: John Wiley & Sons.

Casey, L. S., Chandler, J., Levine, A. S., Proctor, A., & Strolovitch, D. Z. (2017). Intertemporal differences among MTurk workers: Time-based sample variations and implications for online data collection. *SAGE Open*, *7*(2), 1–15.

Coppock, A. (2018). Generalizing from survey experiments conducted on Mechanical Turk: A replication approach. *Political Science Research and Methods*, 1–16.

Gamblin, B. W., Winslow, M. P., Lindsay, B., Newsom, A. W., & Kehn, A. (2017). Comparing in-person, Sona, and Mechanical Turk measurements of three prejudice-relevant constructs. *Current Psychology*, *36*(2), 217–224.

Groves, R. M. (2004). *Survey errors and survey costs*. John Wiley & Sons.

Heen, M. S., Lieberman, J. D., & Miethe, T. D. (2014). *A comparison of different online sampling approaches for generating national samples*. Center for Crime and Justice Policy.

Heitjan, D. F. & Basu, S. (1996). Distinguishing 'missing at random' and 'missing completely at random'. *The American Statistician*, *50*(3), 207–213.

Huff, C. & Tingley, D. (2015). "Who are these people?" Evaluating the demographic characteristics and political preferences of MTurk survey respondents. *Research & Politics*, *2*(3), 1–12.

Keeter, S., Hatley, N., Kennedy, C., & Lau, A. (2017). *What low response rates mean for telephone surveys*. Pew Research Center.

Kennedy, C., Mercer, A., Keeter, S., Hatley, N., McGeeney, K., & Gimenez, A. (2016). *Evaluating online nonprobability surveys*. Pew Research Center.

Levay, K. E., Freese, J., & Druckman, J. N. (2016). The demographic and political composition of Mechanical Turk samples. *SAGE Open*, *6*(1), 1–17.

Long, J. S. & Freese, J. (2014). *Regression models for categorical dependent variables using Stata* (3rd ed.). College Station, Texas: Stata Press.

Long, J. S. & Mustillo, S. A. (forthcoming). Using predictions and marginal effects to compare groups in regression models for binary outcomes. *Sociological Methods and Research*.

Mercer, A., Lau, A., & Kennedy, C. (2018). *For Weighting Online Opt-In Samples, What Matters Most?* Pew Research Center.

Mullinix, K. J., Leeper, T. J., Druckman, J. N., & Freese, J. (2015). The generalizability of survey experiments. *Journal of Experimental Political Science*, *2*(2), 109–138.

Page, B. I. & Shapiro, R. Y. (Eds.). (2010). *The rational public: Fifty years of trends in americans' policy preferences*. Chicago: University of Chicago Press.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*(5), 411–419.

Shapiro, D. N., Chandler, J., & Mueller, P. A. (2013). Using Mechanical Turk to study clinical populations. *Clinical Psychological Science*, *1*(2), 213–220.

Simmons, A. D. & Bobo., L. D. (2015). Can non-full-probability internet surveys yield useful data? a comparison with full-probability face-to-face surveys in the domain of race and social inequality attitudes. *Sociological Methodology*, *45*(1), 357–387.

StataCorp. (2015). *Stata 15*. College Station, TX: Stata Press.

Tourangeau, R., Conrad, F. G., & Couper, M. P. (Eds.). (2013). *The science of web surveys*. Oxford: Oxford University Press.

Williams, R. (2009). Using heterogeneous choice models to compare logit and probit coefficients across groups. *Sociological Methods and Research*, *37*(4), 531–559.